République de Côte d'Ivoire

Union -Discipline -Travail

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Institut National Polytechnique
Félix HOUPHOUËT-BOIGNY

EDP
École Doctorale Polytechnique

UMRI 78

N° d'ordre : 068/2020

THESE

Pour l'obtention du grade de

Docteur de l'Institut National Polytechnique Félix Houphouët Boigny

Mention : Informatique

Spécialité : Fouille de données (*Data mining*)

Thème :

APPROCHE D'UNE METHODE DE FOUILLE DE DONNEES POUR L'ANALYSE DES COMPORTEMENTS À RISQUES LIÉS À L'ACTIVITÉ DOUANIÈRE

Présentée et soutenue publiquement le mercredi 30 septembre 2020 par

M. ZEHERO Bi Bolou Ernest

Devant le jury d'examen composé de :

M. ZOUEU Jérémie	**Professeur Titulaire** Institut National Polytechnique - Houphouët Boigny Yamoussoukro - Côte d'Ivoire	**Président**
M. ASSEU Olivier	**Professeur Titulaire** Institut National Polytechnique - Houphouët Boigny Yamoussoukro - Côte d'Ivoire	**Directeur de thèse**
M. BOURGET Daniel	**Maître de Conférences** Institut Mines-Télécom Atlantique Brest- France	**Co-Encadreur**
M. KERMARREC Yvon	**Professeur HDR** Institut Mines-Télécom Atlantique Brest- France	**Rapporteur**
M. DIABATE Nabongo	**Maître de Conférences** Université Alassane Ouattara Bouaké - Côte d'Ivoire	**Rapporteur**
M. MONSAN Vincent	**Maître de Conférences** Université Félix Houphouët Boigny Abidjan - Côte d'Ivoire	**Examinateur**

INSTITUT NATIONAL POLYTECHNIQUE -FELIX HOUPHOUET BOIGNY

ECOLE DOCTORALE POLYTECHNIQUE

Approche d'une méthode de fouille de données pour l'analyse des comportements à risques liés à l'activité douanière.

Présenté par : **ZEHERO BI BOLOU ERNEST**

en vue de l'obtention du diplôme de : Docteur en Informatique

dans la spécialité de Fouille de Données (Datamining)

Septembre, 2020.

DEDICACE

Ces quelques lignes sont dédiées

à mon épouse Georgette et mes enfants Tadiale, Méliane et Emmanuel,

ils ont toujours été là dans les moments difficiles et que ce travail soit pour eux le

témoignage de ma reconnaissance pour leur soutien inconditionnel.

REMERCIEMENTS

Je souhaite remercier ici tous ceux qui ont contribué à l'aboutissement de ce travail. Qu'ils trouvent ici toute ma reconnaissance.

Je commencerai avec une mention spéciale à mon directeur de thèse, Monsieur ASSEU Olivier, Directeur de la recherche et de l'Innovation Technologique (ESATIC), Professeur Titulaire à l'Institut National Polytechnique Félix Houphouët Boigny, et mon Co-encadreur Monsieur BOURGET Daniel, Maître de Conférences à l'Institut Mines Télécom Atlantique à Brest en France pour sa disponibilité.

Chers maîtres, sans vous ce travail n'aurait certainement pas vu le jour. Toute ma reconnaissance pour la confiance, les conseils, les échanges. Je ne saurai énumérer toutes les choses pour lesquelles je devrais vous remercier, Merci pour tout.

Nous remercions également :

- Professeur YAO Benjamin, Directeur de l'Ecole Doctorale Polytechnique (INPHB) ;
- Professeur SORO Doudjo, Directeur des études de l'Ecole Doctorale Polytechnique (INPHB) ;
- Professeur ZOUEU Jérémie, Directeur de l'UMRI Electronique et Electricité Appliquée de l'Ecole Doctorale Polytechnique (INPHB).

Qui contribuent par leurs actions à la promotion de la Recherche Fondamentale et Appliquée en Côte d'Ivoire

Je remercie particulièrement Docteur BROU Aguié Pacôme, Assistant à l'Ecole Supérieure Africaine des TIC (ESATIC), Direction de la Recherche et de l'Innovation Technologique pour ses conseils, son encadrement rapproché, assistance et aide technique, tout au long de ce projet de thèse.

En ma qualité de fonctionnaire au sein de l'Administration des douanes ivoiriennes, je voudrais exprimer ma profonde gratitude à l'endroit des autorités des douanes ivoiriennes :

- Général Pierre Alphonse DA, Directeur Général des douanes de la République de Côte d'Ivoire,
- Colonel Major Louis Albert KADIO, Directeur Général Adjoint des douanes de la République de Côte d'Ivoire
- Colonel Major Issa OUATTARA, Directeur Général Adjoint des douanes de la République de Côte d'Ivoire

Pour leurs bienveillantes interventions auprès des différents services des douanes en vue de la facilitation de la collecte des informations utiles au projet de thèse.

Je remercie également les différents responsables des services des douanes ivoiriennes pour leur immense contribution.

J'adresse également mes remerciements aux membres du jury pour l'honneur qu'ils me font en participant à l'évaluation de ce travail :

M. ZOUEU Jérémie	Professeur Titulaire	Président
M. KERMARREC Yvon	Professeur, HDR	Rapporteur
M. DIABATE Nabongo	Maître de Conférences	Rapporteur
M. MONSAN Vincent	Maître de Conférences	Examinateur

RESUME

Les travaux de cette thèse s'inscrivent dans le cadre de l'extraction de connaissances et de la fouille de données appliquées à un entrepôt de données d'infractions douanières, afin d'extraire des résumés linguistiques sur la base de motifs fréquents, exprimant des corrélations entre des attributs de la forme, Si « *Antécédent* » alors « *Conséquence* ». Notre objectif est de découvrir des connaissances pour ensuite faire une analyse prédictive du comportement à risque dans le contrôle douanier pour anticiper les risques de fraudes liés à l'activité douanière.

L'administration douanière est une institution fiscale dont la mission principale est de veiller à l'application de la réglementation douanière et à la perception des droits et taxes exigibles sur les marchandises importées et exportées.

Dans ce contexte, analyser, étudier et identifier les risques de fraude dans le processus de dédouanement des biens et services est essentiel pour remplir sa mission régalienne de l'État. L'accroissement des échanges commerciaux couplé aux progrès technologiques ces dernières années ont fortement contribué à la digitalisation des administrations douanières par la mise en place de systèmes de stockage relatif aux opérations de dédouanement.

L'analyse à posteriori de ces différentes données liées au dédouanement et de risques de fraude présente des perspectives intéressantes pour la compréhension et l'aide à la modélisation des comportements à risques dans le domaine douanier. En effet, dans le cadre de notre projet de recherche, nous proposons une méthodologie basée sur la fouille de données et adaptée au contexte douanier pour extraire automatiquement des connaissances en proposant des règles d'association permettant de découvrir des connaissances dans une base

de données transactionnelles entre une opération de dédouanement et la nature de l'infraction constatée dans un premier temps, puis en explorant la structure symbolique des différentes données liées à ces transactions, nous pouvons extraire de nouvelles règles d'association au niveau des exportateurs et importateurs pour étudier leurs comportements dans le processus de dédouanement. Pour ce faire, nous exploitons, une masse de données d'infractions douanières tirées du Procès-Verbal Simplifié (PVS) du Système Automatisé de dédouanement des Douanes de la République de Côte d'Ivoire (SYDAM World), de la période allant de Mai 2015 et Mai 2018. Notre conception exploite principalement la force des techniques de fouille de données, à partir de l'algorithme Apriori (règles d'association), pour chercher l'information pertinente dans la masse de données en question. Nos travaux montrent que l'approche Apriori et son extension à partir des données symboliques, permettent d'obtenir une corrélation potentielle entre une opération de dédouanement et une nature de fraude, puis faire une analyse prédictive du comportement à risque des opérateurs.

Mots clés : Fouille de données ; extraction de données ; règles d'association ; algorithme Apriori, Données symboliques, Activité douanière.

ABSTRACT

Work of this thesis falls within the framework of knowledge extraction and data mining applied to a customs offence data warehouse, in order to extract linguistic summaries on the basis of frequent patterns, expressing correlations between attributes of the form, If "Antecedent" then "Consequence". Our objective is to discover knowledge and then make a predictive analysis of risk behavior in customs control in order to anticipate fraud risks related to customs activity.

Customs administration is a fiscal institution whose main mission is to ensure the application of customs regulations and the collection of duties and taxes due on imported and exported goods.

In this context, analyzing, studying and identifying the risks of fraud in the process of customs clearance of goods and services is essential to fulfill the State's regalian mission. The increase in trade coupled with technological advances in recent years have greatly contributed to the digitalization of customs administrations through the implementation of storage systems for customs clearance operations.

Posteriori analysis of these various data related to customs clearance and fraud risks presents interesting perspectives for understanding and helping to model risk behaviors in the customs field. Indeed, within the framework of our research project, we propose a methodology based on data mining and adapted to the customs context to automatically extract knowledge by proposing association rules allowing to discover knowledge in a transactional database between a customs clearance operation and the nature of the infraction observed at first, then by exploring the symbolic structure of the different data related to these transactions, we can extract new association rules at the level of exporters

and importers to study their behaviors in the customs clearance process. In order to do so, we exploit a mass of customs infringement data from the Simplified Procès-Verbal Simplifié (PVS) of the Automated Customs Clearance System of the Republic of Côte d'Ivoire (SYDAM World), for the period between May 2015 and May 2018. Our design mainly exploits the strength of data mining techniques, based on the Apriori algorithm (association rules), to search for relevant information in the mass of data in question. Our work shows that the Apriori approach and its extension based on symbolic data, allows us to obtain a potential correlation between a customs clearance operation and a type of fraud, and then to make a predictive analysis of the risk behavior of operators.

Keywords: Data mining; data extraction; association rules; Apriori algorithm, Symbolic data, Customs activity

TABLE DES MATIERES

DEDICACE 3

REMERCIEMENTS 4

RESUME 6

ABSTRACT 8

TABLE DES MATIERES 10

LISTE DES TABLEAUX 16

LISTE DES FIGURES 18

LISTE DES SIGLES ET ABREVIATIONS 19

INTRODUCTION GENERALE 20

Contexte et étude 20

Notion de fouille de données 21

Problématique et question de recherche 23

Objectifs recherchés 24

Méthodologie de l'approche 25

Contributions 26

Méthodologie de travail et structuration du mémoire 28

CHAPITRE 1 : CONCEPTS GENERAUX : LA FOUILLE DES MOTIFS 34

1.1 Introduction 35

1.2 Data Mining : Concepts et principes 36

1.2.1 Composantes du Data Mining 36

1.2.2 Extraction de Connaissances versus Fouille de Données (ECBD vs FD) 37

1.3 Fondements mathématiques : Notions de base....38

1.3.1 Contexte formel....39

1.3.2 Correspondance de Galois....40

1.3.3 Concept formel....41

1.3.4 Théorie des treillis : Notion de base....41

1.4 Règles d'association et implications....45

1.4.1 Règles d'association à partir de données binaires....47

1.4.2 Règles d'associations quantitatives....53

1.4.3 Règles d'associations floues....56

1.4.4 Autre extension des règles d'associations : *Cas des motifs séquentiels*....58

1.5 Techniques de fouille de données....59

1.5.1 Techniques de fouille de données d'apprentissage supervisé....60

1.5.2 Technique de fouille de données d'apprentissage non-supervisé....62

1.6 Conclusion....63

CHAPITRE 2 : ANALYSE BIBLIOGRAPHIQUE : PROBLEMATIQUE, APPLICATIONS ET OUTILS.....64

2.1 Introduction....65

2.2 Fouille de Données : Problématique sur l'étude....65

2.3 Travaux majeurs....66

2.3.1 Fouille des données en météorologie et en astrologie....68

2.3.2 Fouille des données en Bio-informatique....69

2.3.3 Fouille des données pour la détection des profils communautaires 69

2.3.4 Fouille des données pour le suivi de trajectoires d'objets mobiles 70

2.3.5 Applications du datamining dans les autres domaines d'activités 70

2.3.6 Fouille des données et activité douanière 71

2.4 Méthodes de Fouille de Données Structurées 73

2.4.1 La fouille des graphes 73

2.4.2 Programmation logique inductive 74

2.4.3 La fouille de données Multi-tables 74

2.5 Algorithmes d'extraction de connaissances 74

2.5.1 L'approche par segmentation 74

2.5.2 L'approche par classification 80

2.5.3 L'Approche Naïve 89

2.6 Application de la fouille des données 92

2.6.1 Chaîne d'extraction des connaissances sur les données 96

2.6.2 Techniques de génération de motifs fréquents 98

2.7 Environnement libre des fouilles de motifs 99

2.7.1 WEKA 100

2.7.2 SPMF 101

2.7.3 KNIME 101

2.7.4 Rattle et R 102

2.7.5 Tanagra 102

2.7.6 Mahout 103

2.7.7 Orange 103

2.7.8 ELKI 104

2.8 Analyse des travaux présentés de la littérature 106

2.9 Question de recherches et positionnement de nos travaux 108

2.10 Conclusion 109

CHAPITRE 3 : REGLES D'ASSOCIATIONS SUR LA BASE DE MOTIFS FREQUENTS 110

3.1 Introduction 111

3.2 Motivation et problématique 111

3.3 Problème d'apprentissage : Les règles d'associations 114

3.4 Observations sur les travaux de la littérature : *Limites et pistes d'amélioration* 115

3.5 Approche méthodologique : Définition du problème 118

3.5.1 Définition du problème 122

3.5.2 Algorithme Apriori 126

3.5.3 Règles valides 135

3.6 Modélisation de la cartographie du risque douanier 137

3.6.1 Modèle mathématique de la cartographie des risques 138

3.7 Espace des données à explorer 144

3.7.1 Base de données des infractions douanières issues du Procès-Verbal Simplifié du Système de dédouanement de la république de Côte d'Ivoire....144

3.7.2 Acquisition de données : Nettoyage....146

3.7.3 Sélection des données....146

3.8 Expérimentation : *Mise en contexte et résultats*....147

3.8.1 Résultats et Analyses....150

3.9 Comparaison entre le modèle économétrique et l'approche de la Fouille de Données....156

3.10 Conclusion....157

CHAPITRE 4 : ANALYSE PREDICTIVE DES COMPORTEMENTS A RISQUES. 159

4.1 Introduction....160

4.2 Positionnement du problème....160

4.3 Notion de la fouille de données symbolique : Concept de base....163

4.3.1 Description de la notion d'objet symbolique....164

4.3.2 Définition de la notion de données symboliques....164

4.3.3 Présentation d'une matrice symbolique....165

4.4 Méthode de l'algorithme Apriori Étendu....166

4.4.1 Principe de la méthode....166

4.4.2 Précision du découpage h....167

4.4.3 Définition des indicateurs : *Support, Confiance, Confiance Diagramme*....168

4.5 Étapes algorithme de l'approche Apriori Étendu 169

4.6 Application et Résultats 170

4.6.1 Résultats et Analyses 170

4.7 Comparaison de l'algorithme Apriori et de l'algorithme Apriori étendu 173

4.8 Conclusion 174

CONCLUSION GENERALE ET RECOMMANDATIONS 175

1. Bilan des contributions 175

2. Intérêts des travaux pour l'administration douanière 176

3. Limites des travaux effectués 177

4. Travaux futurs : Perspectives 177

REFERENCES BIBLIOGRAPHIQUES 179

ANNEXE A : PROCES-VERBAL SIMPLIFIE (PVS) 201

ANNEXE B: DICTIONNAIRE DES DONNEES ET TABLE DE LA BASE DE DONNEES DU PVS (SYDAM WORLD) 202

ANNEXE C: PUBLICATION SCIENTIFIQUE #1 211

ANNEXE D: PUBLICATION SCIENTIFIQUE #2 212

LISTE DES TABLEAUX

Tableau 1.1: Contexte d'extraction B 39

Tableau 1.2: Exemple d'une Base de Données binaires 48

Tableau 1.3: Principale mesures de qualité d'une règle d'association $X \rightarrow Y$ 52

Tableau 1.4 : Tableau comparatives de trois règles d'associations quantitatives....55

Tableau 2.1: Classification des méthodes de la segmentation 75

Tableau 3.1: Modèle de base de données au format attribut /valeur 123

Tableau 3.2: Modèle de base de données au format transactionnel 124

Tableau 3.3 : *Algorithme 1* - Génération de motifs fréquents 129

Tableau 3.4 : *Algorithme 2* - Génération des ($k+1$)- motifs candidats 130

Tableau 3.5 : *Algorithme 3* - Génération des règles d'association....133

Tableau 3.6 : *Algorithme 4* - Génération des règles d'associations de plus d'un item....134

Tableau 3.7: Règles d'association valides 136

Tableau 3.8 : Exemple représentatif d'une Base de données binaires....149

Tableau 3.9:Résultats des implémentations 151

Tableau 3.10: Tableau de comparaison entre le modèle économétrique et l'approche de fouille de données 156

Tableau 4.1 : Exemple d'une matrice avec 10 cas de transactions classiques.....163

Tableau 4.2 : Matrice des données symboliques composée d'une valeur diagramme....166

Tableau 4.3 : Matrice pour des opérations classiques de dédouanement....171

Tableau 4.4 : Matrice de données symboliques composées d'une seule variable.....171

Tableau 4.5 : Règles d'associations symboliques 172

Tableau 4.6: Tableau de comparaison entre les algorithmes Apriori et la méthode par diagramme.....173

LISTE DES FIGURES

Figure 1.1 : Les étapes d'extraction de règles d'association 23

Figure 1.2 : Démarche méthodologique d'avancement du projet de thèse. 28

Figure 1.3 : Organigramme du manuscrit de thèse. 32

Figure 1.5: Architecture type d'un système d'ECBD [Han, 2000]. 37

Figure 1.6 : Processus d'extraction de connaissances [Fayyad, 1996]. 38

Figure 1.7 : Treillis des parties associé à A = {A,B,C,D,E} 45

Figure 2.1: Processus de classification 81

Figure 2.2: Schéma de construction d'un modèle de *k=5* plus proche voisins 83

Figure 2.3: Modèle d'un neurone artificiel 85

Figure 2.4 : Architecture d'un réseau de neurones artificiel 86

Figure 2.5 : Schéma de la classification bayésienne 88

Figure 2.6 : Les différents domaines d'applications de la fouille de données 95

Figure 3.1: Modèle économétrique pour la gestion du risque en douane 116

Figure 3.2 : Processus d'analyse des données 117

Figure 3.3: Les différentes étapes de l'Extraction de Connaissances des Données 120

Figure 3.4 : Treillis des motifs potentiellement fréquents quand C n'est pas fréquent 128

Figure 3.5 : Modèle conceptuel de la base de données du Procès-Verbal Simplifié (PVS) 145

LISTE DES SIGLES ET ABREVIATIONS

ACF	Analyse de Concepts Formels
AFC	Analyse Formelle de Concepts
AN	Approche Naïve
DARRV	Direction de l'Analyse du Risque, du Renseignement et de la Valeur
DHP	Direct Hashing and Pruning
DIC	Dynamic Itemset Counting
ECBD	Extraction de Connaissances à partir de Bases de Données
ELKI	Environment for DeveLoping KDD-Applications supported by Index structures
FD	Fouille de Données
FP	Frequent-Pattern
IJCAI	International Joint Conference On Artificial Intelligence
KNIME	Konstanz Information Miner
PVS	Procès-Verbal Simplifié
RA	Règle d'Association
RATTLE	The R Analytical Tool To Learn Easily
RNA	Réseaux de Neurones Artificiels
SPMF	Sequential Pattern Mining Framework
WEKA	Waikato Environment for Knowledge Analysis

INTRODUCTION GENERALE

Contexte et étude

Les progrès réalisés dans les technologies d'acquisition, de stockage et de distribution des données ont entraîné une croissance rapide de la taille des bases de données disponibles. En effet, la quantité de données estimée est doublée presque chaque année, d'où l'intérêt croissant de valoriser ces données. Par conséquent, un des plus grands défis auxquels sont confrontées les administrations et organisations est de savoir comment transformer leur volume de données qui ne cesse d'augmenter en des connaissances utiles et exploitables. Les travaux qui ont abordé cette problématique relèvent de domaines très variés tels que les statistiques, l'intelligence artificielle, l'apprentissage automatique, les bases de données, etc. Ces travaux ont donné lieu à un nouveau domaine de recherches connu sous le terme de Fouille de Données.

Dans le cadre du processus de dédouanement, les administrations douanières sont confrontées de nos jours à un volume croissant de marchandises en liaison avec le développement du commerce international. Une gestion automatisée des risques liés à l'activité douanière à partir des méthodes de fouilles de données ; est ainsi requise afin de limiter le contrôle intrusif et optimiser le contrôle douanier.

Le risque douanier peut être défini comme la probabilité que survienne un élément ayant une incidence sur les objectifs de dédouanement. Dans cette optique, la gestion des risques en douane consiste à contrôler et à gérer le risque afin d'améliorer la qualité des contrôles réalisés ; nécessaire à la facilitation des échanges et à l'amélioration des recettes fiscales. Le traitement douanier réservé à une marchandise lors d'opérations d'importation et d'exportation requiert la connaissance et la maitrise de trois notions essentielles : *l'espèce*

tarifaire, l'origine et la valeur en douane. Ces informations sont à transmettre aux autorités douanières lors des passages en douane, via la déclaration douanière. C'est donc en anticipant et en maîtrisant ces données que l'on peut supprimer le risque douanier, et ainsi organiser et gérer les flux à l'international au mieux.

Notion de fouille de données

La croissance exponentielle de la masse des données stockées dans les entrepôts de données, dans des plusieurs secteurs d'activités tels que : la météorologie, données médicales issues des capteurs, la génétique, etc., exige à ne plus opérer de manière traditionnelle pour la transformation de ces données en connaissances [Fayyad et al., 1996]. De nos jours des bases de données contenant un nombre d'enregistrements de l'ordre de 10^9 n'a plus rien d'exceptionnel [Han et Kamber, 2012]. L'expression fouille de données tire son fondement de la définition du terme « Exploratory data analysis » donnée par John Tukey en 1977 [Tukey, 1977]; en effet, c'est la tâche d'extraire des informations inconnues et potentiellement importantes de grandes bases de données [Piateski et Frawley, 1991] (*Voir la figure 1.1 pour les différentes étapes de la fouille de données*). Il s'agit d'un domaine de recherches récent et important dans les bases de données, l'intelligence artificielle, de la statistique, des interfaces homme/machine et de la visualisation. [Piatetsky-Shapiro 1991]. À partir de données recensées, il s'agit de proposer des connaissances nouvelles qui enrichissent les interprétations du champ d'application, tout en fournissant des méthodes automatiques qui exploitent cette information. Son processus est décliné en quatre procédures différentes :

1.Collecte de données : Il s'agit de recueillir des données de différentes sources comme le Web, les entrepôts de données ou les bases de données. Ces données sont accompagnées de plusieurs types de bruit (redondance de données, non-conformité de données ; données

manquantes). Nettoyer et optimiser la base de données est une étape importante avant l'utilisation de ces données.

2. Prétraitement des données collectées : Implique plusieurs tâches différentes, comme la sélection des caractéristiques, le classement des caractéristiques, la sélection d'échantillons, etc. Son but est de sélectionner les caractéristiques ou les échantillons les plus importants de l' entrepôt de données nécessaires pour le processus de datamining. Les fonctionnalités sélectionnées sont parfois accompagnées d'un format inapproprié. Leur transformation en fonction de l'approche de la tâche de datamining fait également partie de la procédure de prétraitement.

3. Exploration de données : La phase principale de la fouille de données au cours de laquelle les données traitées, génèrent différents types de résultats en fonction des approches utilisées. Le regroupement, la classification, la régression, la génération de règles d'association, la sélection de modèles, la correspondance de modèles sont quelques-unes des nombreuses tâches du forage de données [Han 2011]. Le forage de données est le processus de recherche et d'analyse qui permet de trouver, généralement dans des bases de données, des corrélations cachées ou des informations nouvelles, ou encore, de dégager certaines tendances.

4. Recherche d'informations : Les résultats sont interprétés par l'expert de l'application pour obtenir des informations cachées réelles à partir des données collectées lors de la première étape. Il est applicable à nombreux domaines d'application : *De la biologie cellulaire aux applications spatiales, etc.* Ce qui le rend populaire pour extraire des connaissances cachées à partir des données stockées.

L'étape de la fouille de données correspond à l'ensemble des techniques et méthodes mises en œuvre afin d'extraire, à partir de données préparées, des informations exploitables non accessibles par les méthodes classiques [Abbas, 2012].

Ces informations sont des règles d'associations, des motifs fréquents ou rares, des regroupements, des anomalies et des modèles.

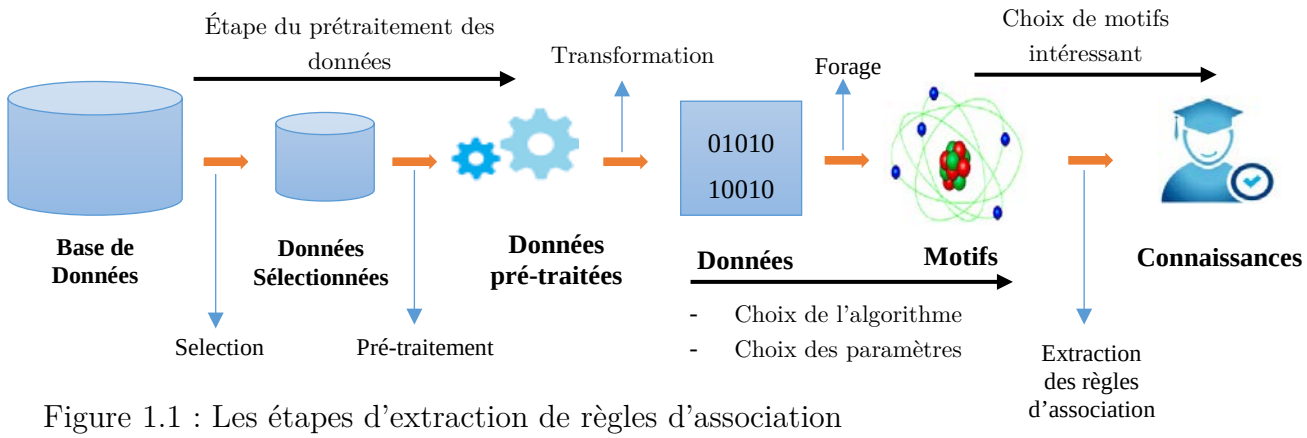

Figure 1.1 : Les étapes d'extraction de règles d'association

Problématique et question de recherche

Dans un monde qui s'ouvre de plus en plus facilement aux nouvelles technologies, il devient très facile de collecter et de stocker des informations. Il en est de même pour les administrations douanières qui sont confrontées de nos jours à un volume important de marchandises en lien avec le développement du commerce international. Le volume croissant de marchandise au cordon douanier induit des contrôles intrusifs, autrement dit plus le volume de marchandises augmentent, plus le contrôle douanier s'intensifie en raison des risques de fraude. Pour limiter les contrôles intrusifs, les administrations douanières doivent recourir à l'analyse du risque pour faciliter les operation douanières. Aussi l'analyse du risque contribue-t-elle à la mobilisation des recettes qui constitue une priorité tant au regard de l'équilibre budgétaire qu'en matière de réduction de la pauvreté. Par ailleurs, la plupart des administrations douanières modernes sont dotées de technologies d'acquisition de stockage et de distribution de données massives. D'où l'intérêt de la fouille de données en douane pour optimiser le contrôle douanier dans le cadre de la facilitation du commerce [OMD, 2002].

Quelle démarche globale structurée et systématique mettre en place pour la gestion des risques liés à l'activité douanière afin de :

- Etablir une cartographie des risques de fraude
- Analyser et prédire les comportements à risques des opérateurs

Objectifs recherchés

Dans le cadre de nos travaux de thèse, nous nous intéressons à l'extraction de motifs intéressants : non triviaux, implicites, inconnus auparavant, potentiellement utiles, à partir d'une masse de données d'infractions issues de l'activité douanière. Tan et al. Mettent en évidence trois grandes catégories de fouille [Tan et al., 2006] :

- La classification ;
- Le regroupement automatique ;
- Et la découverte d'associations.

Au regard de la problématique, nous visons dans ce projet de thèse comme objectifs principaux, de :

- Construire un modèle descriptif pour l'analyse temps-réel des infractions douanières
- Déterminer un modèle prédictif pour l'anticipation des comportements des opérateurs à risque lié à l'activité douanière.

Ces deux objectifs principaux se déclinent en quatre sous points :

- Formuler un modèle mathématique de la cartographie des risques liés aux activités douanières ;
- Identifier une corrélation entre une opération de dédouanement et la nature de l'infraction survenue à partir d'une méthode de fouille de données non supervisée ;

- Extraire des connaissances d'un entrepôt de données des infractions douanières en vue de prédire les comportements à risque des opérateurs ;
- Modéliser des concepts explorant la structure symbolique des données de fraudes douanières ;

Notre démarche dans ce mémoire nous orientera vers la découverte de règles d'association, sur la base de motifs fréquents. Puis à partir de ces règles, nous extrayons des connaissances relatives aux comportements à risque, utiles d'aide à la décision.

Méthodologie de l'approche

La méthodologie adoptée dans notre projet de thèse se présente en plusieurs phases, et regroupée en deux grandes parties :

- Le contexte d'étude et l'état de l'art
- Les contributions apportées

La première partie vise à situer le contexte de l'étude en exposant sur les travaux de la littérature qui mettent en avant la notion de gestion du risque douanier et les questions de l'application de la fouille de données dans différents domaines d'application dont celle de l'activité douanière en matière de contrôle, dans l'optique d'extraire des connaissances fondamentales pour une anticipation des infractions et analyser le comportement à risque des opérateurs économiques.

Dans cette section, il est question des différents domaines et cas d'utilisation de la fouille de données. Cette étape a permis de faire ressortir des pistes de solution pour la suite du travail. Une étude des grandes familles d'algorithmes est réalisée, soit la segmentation, soit la classification, ainsi que les algorithmes et la théorie s'y rattachant [Devroye et al., 1996], [Vladimir N,1982 ; 1998], et [Hastie et al., 2001]. Plusieurs algorithmes ont été analysés lors de cette phase. Nous définissons également les indicateurs sur lesquels vont s'appuyer les

différentes étapes de génération des Règles d'Association (RA). Par la suite, il est question de la préparation : des données et l'adaptation formelle des algorithmes choisis ; en effet, entrepôt de données contenant des données bruitées qui ne reflètent pas toujours la réalité. Ainsi, il faut opérer un nettoyage de ces informations pour les rendre plus accessibles aux algorithmes afin d'obtenir des résultats interprétables. Cette phase est d'une grande importance car les différents résultats sont le reflet de la qualité des données.

La seconde partie du mémoire est relative aux contributions et à l'utilisation des modèles théoriques définis à l'étape précédente. Pour ce faire, un algorithme est développé pour lier le module de datamining et la base de données. Cette contribution a été réalisée à partir d'expérimentation. Les objectifs expérimentaux étaient de tester différents algorithmes de classification sur des données réelles provenant d'administration douanière. Dans cette optique, il a fallu avoir recours à une base de données contenant des caractéristiques d'infractions douanières entre 2015 et le premier semestre de 2018. Le programme va chercher l'information ciblée grâce aux motifs fréquents de fraude et les transmet aux algorithmes de datamining. Les résultats sont ensuite transférés aux modules d'évaluation des données. Le résultat est conservé pour fins d'analyse.

Contributions

La contribution du mémoire prend la forme de l'adaptation, de l'implémentation et de l'expérimentation d'algorithmes de Datamining pour l'analyse du comportement à risque dans l'activité douanière. Ainsi, le mémoire fournit des pistes de solutions pour l'interprétation des règles d'association entre un type et une nature de fraude lors d'opération de dédouanement d'une part et l'analyse du comportement d'un opérateur économique *alpha* ou *oméga* en lien avec un type de fraude d'autre part.

La première contribution fait cas d'un algorithme classique appliqué sur la base de motifs fréquents permettant d'extraire des connaissances dans une base de données d'infractions issues des transactions douanières. En effet, trois grandes règles d'association issues de données binaires : les règles de prévisions, les règles de ciblage et les règles neutres ont permis d'établir une association entre les types et la nature des fraudes.

La seconde contribution est relative à une modélisation des concepts par exploration de la structure symbolique des données avec **une extension de l'algorithme Apriori (Algorithme Apriori Etendu)**. En effet, nous étudions le comportement des opérateurs en lien direct avec les fraudes constatées en analysant la structure symbolique des données, avec une extension de l'algorithme Apriori. Ce qui a permis de mettre en évidence de nouvelles règles d'associations et l'extraction de nouvelles connaissances sur les comportements à risques des opérateurs économique en considérant un nouvel indicateur : « *Confiance Diagramme* (ConfD) ».

La plus-value de cette seconde contribution est la découverte automatique de comportements à risques des opérateurs et leur lien direct avec la nature des infractions découvertes dans la première contribution.

L'efficacité de ces méthodes algorithmes a été expérimentée sur des données recueillies dans entrepôt de données des informations relatives aux infractions tirées du Procès-Verbal Simplifié et issues exclusivement des opérations douanières de 2015 à 2018 en Côte d'ivoire (Direction Générale des Douanes de Côte d'Ivoire). Ces informations concernent 6854 infractions issues soit :

(1) les opérations de dédouanement des marchandises à l'importation et à l'exportation ; et des contrôles de change

(2) Du contrôle de l'application de la réglementation douanière et de l'ensemble des procédures ;

Enfin, ce travail démontre de la possibilité d'utiliser les approches de fouille de données dans le secteur de la douane pour conjuguer les efforts de réduction du risque notamment en matière de fraude.

Méthodologie de travail et structuration du mémoire

Cette section est subdivisée en deux sous sections :

- (1) Méthodologie de travail

Nous présentons à la figure 1.2 les étapes générales représentant la méthode de travail et d'avancement suivies dans notre projet de thèse. Les lectures bibliographiques effectuées ont été organisées en plusieurs thèmes à savoir, la gestion des risques et l'analyse de comportements dans l'administration douanière, la fouille de données ainsi que les méthodes d'extraction des règles d'association.

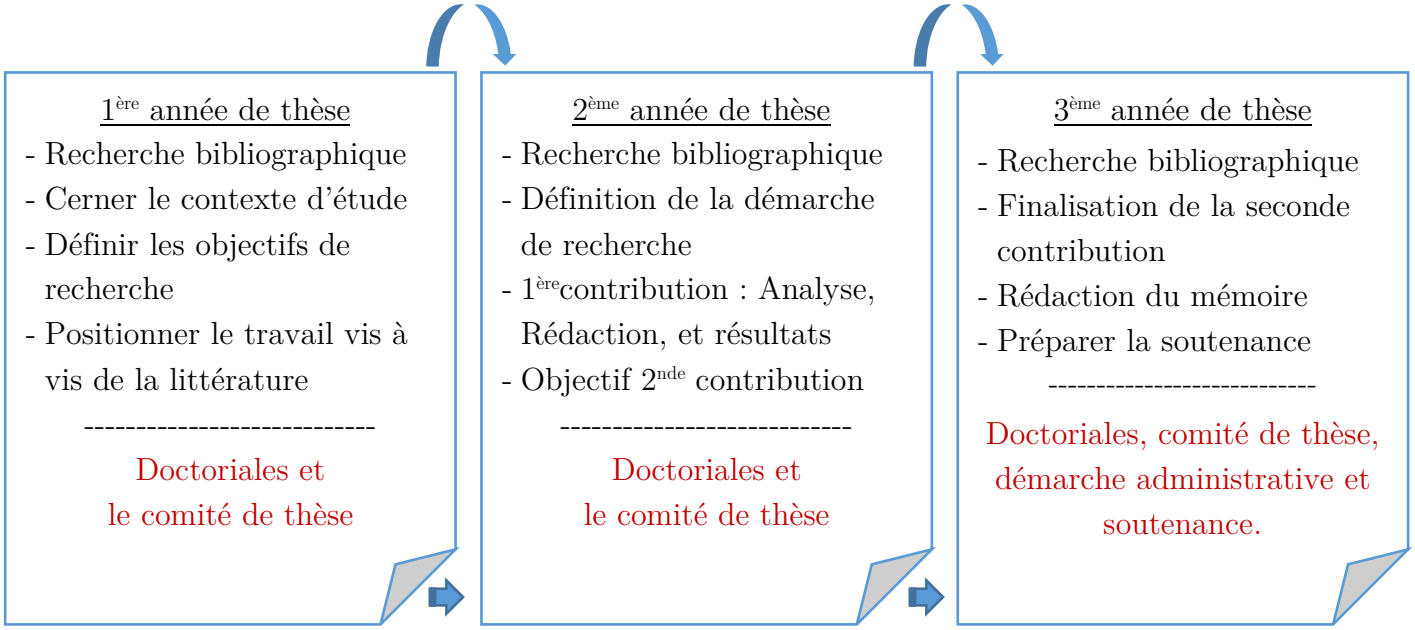

Figure 1.2 : Démarche méthodologique d'avancement du projet de thèse.

- (2) Structuration du manuscrit de thèse

Le mémoire thèse suit l'ordre chronologique selon les étapes de réalisation suscitées. Il est structuré en quatre chapitres en dehors de l'introduction générale et la conclusion générale. Les annexes sont constituées des articles scientifiques publiés et des tables du PVS (SYDAM World) et la circulaire instituant N°1615/MPMEF/DGD/ Du 21 juin 2003 instituant la mise en place d'une Base de Données des infractions douanières à partir du Procès-Verbal Simplifié dans le cadre d'une approche des contrôles douaniers basés sur l'analyse du risque.

Introduction générale : Le manuscrit débute par une introduction générale qui porte évidement sur l'introduction au projet de thèse ; dans cette section, nous donnons le but et les objectifs de notre projet de thèse. La notion de l'activité douanière abordée dans cette partie met en lumière tous les éléments importants à la compréhension du sujet. Puis expliquant le processus de l'approche méthodologique développée, nous présentons les deux contributions issues du travail de recherches de cette thèse.

Chapitre 1 : Généralités sur la fouille de données.

Ce chapitre est une introduction aux notions de fouille de données nécessaires à la compréhension du problème : Concept et principe des composantes de la fouille de données ainsi que des notions de base mathématique (*Contexte formel, correspondance de Galois et concept formel et la notion de la théorie des treillis*) Méthode d'exploration de données, les indicateurs intervenant pour la génération de règles d'association. Un accent particulier est mis sur les règles d'association et les méthodes de fouille de données.

Chapitre 2 : Analyses Bibliographiques : Problématique, applications et outils.

Le deuxième chapitre porte sur la revue littéraire ; la majeure partie de cette section concerne la fouille de données. Ce chapitre est subdivisé en deux sections.

- La première section présente un état de l'art sur les approches de fouille via leur application dans divers domaines d'activités.
- La deuxième section est un état de l'art sur les grandes familles des algorithmes de la fouille de données. Ces différentes approches algorithmiques de fouille sont développées pour résoudre les différentes applications en fonction de leurs principes méthodologiques.

Il sera aussi question des différentes techniques pour la préparation des données et l'amélioration des résultats pour certains algorithmes. Enfin, une analyse synthétique de ces deux états de l'art à la dernière section du chapitre permet de définir la stratégie d'approche des contributions que nous développons au chapitre 3 et au chapitre 4 de ce mémoire de thèse.

Chapitre 3 : Règles d'association sur la base de motif fréquent.

Ce chapitre est consacré à la première contribution de la thèse. En effet, par application du principe d'Apriori sur la base de motifs fréquents dans une base de données relatives aux infractions douanières dans le cadre des procédures douanières, nous découvrons des règles potentielles d'associations entre une opération douanière et une infraction dans le but d'extraire des connaissances régissant la survenue de la fraude.

Chapitre 4 : Analyse prédictive des comportements à risques liés à l'activité douanière.

La seconde contribution de la thèse est proposée dans ce chapitre. Dans le but d'aider l'administration douanière à anticiper les risques de fraude, nous proposons une extension de l'algorithme Apriori qui permet de faire une analyse prédictive des comportements à

risque lors des opérations de dédouanement. En effet, plutôt que d'extraire des règles d'associations au niveau des opérations de dédouanement et les natures des infractions douanières (chapitre 3), nous explorons la structure symbolique des données dans l'idée sous-jacente d'extraire de nouvelles règles d'association mettant en exergue le comportement des opérateurs et les infractions constatées.

Conclusion générale et perspectives : Cette thèse s'achève par la présentation d'une conclusion synthétisant les différents résultats obtenus, l'intérêt des travaux pour les administrations douanières et les recommandations.

ORGANIGRAMME DU MÉMOIRE DE THÈSE

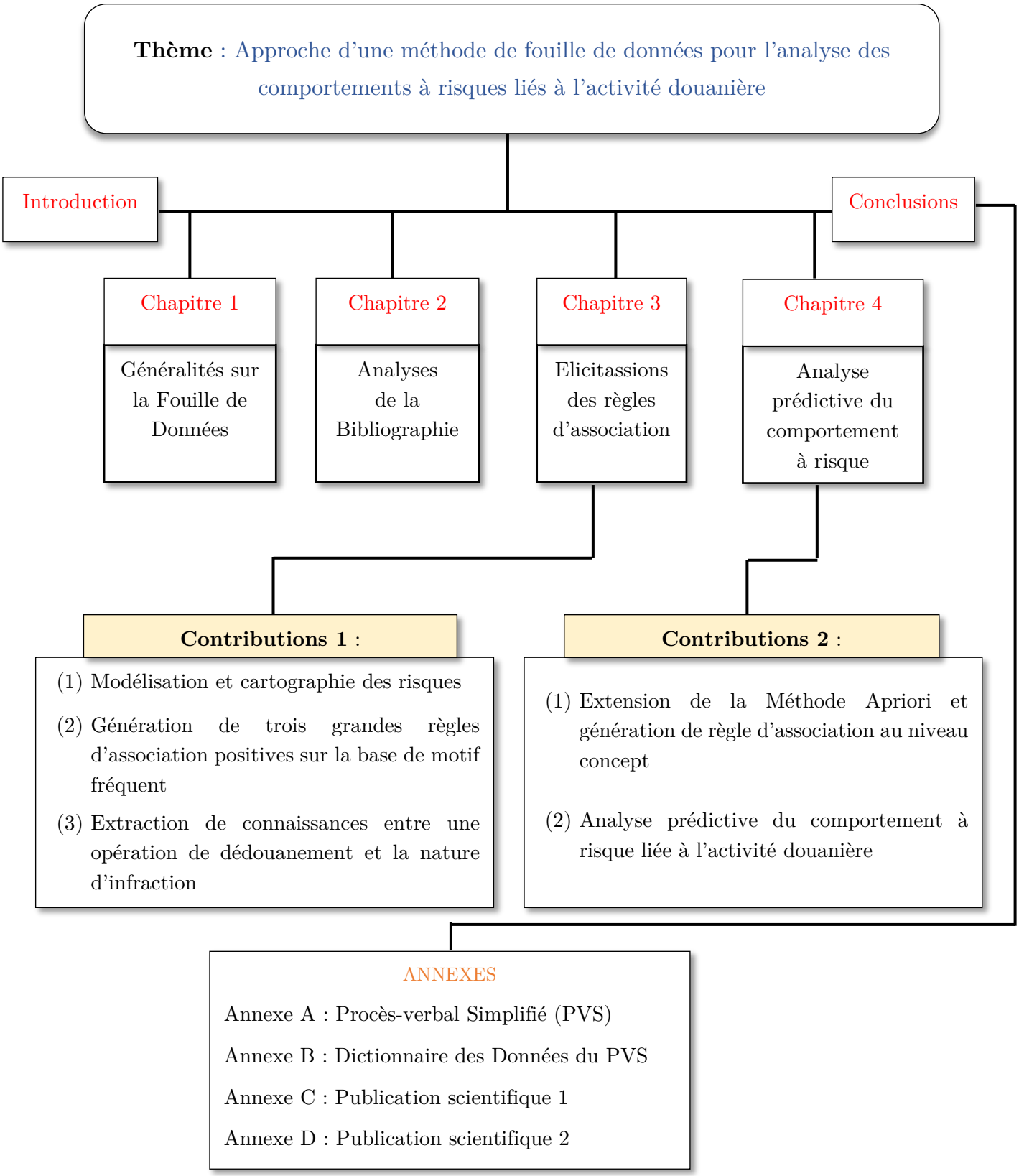

Figure 1.3 : Organigramme du manuscrit de thèse.

PUBLICATIONS SCIENTIFIQUES

Cette thèse a fait l'objet de deux articles scientifiques publiés au cours des 2èmes et 3ème année de thèse, soit en 2018 et 2019. ces articles sont fournis en annexes C et D de ce mémoire.

Article 1	Titre	Auteurs	Journal	Editeur
	Elicitation of Association Rules from Information on Customs Offences on the Basis of Frequent Motives	**Bi Bolou Zehero**, Etienne Soro, Yake Gondo, Pacôme Brou, Olivier Asseu	*Engineering,* http://www.scirp.org/journal/eng ISSN: 1947-394X, 1947-3931	*Scientific & Academic Publishing*
Citation	**Zehero, B.B.,** Soro, E., Gondo, Y., Brou, P. and Asseu, O. (2018) Elicitation of Association Rules from Information on Customs Offences on the Basis ofFrequent Motives. Engineering, 10, 588-605. https://doi.org/10.4236/eng.2018.109043			
Indexé par**Index Copernicus.** https://journals.indexcopernicus.com/search/formjml				

Article 2	Titre	Auteurs	Journal	Editeur
	Predictive Analysis of Risk Behaviors Related to Customs Activity Using a Diagram Algorithm	**Zehero Bi Bolou**, Brou Pacôme, Soro Etienne, Asseu Olivier and Daniel Bourget	Far East Journal of Mathematical Sciences (FJMS) http://www.pphmj.com ISSN: 0972-0871	*Pushpa Publishing House*
Citation	Zehero Bi Bolou, Brou Pacôme, Soro Etienne, Asseu Olivier and Daniel Bourget. (2019). predictive analysis of risk behaviors related to customs activity using a diagram algorithm, Far East Journal of Mathematical Sciences, Volume 110, Number 1, 2019, Pages 217-232. http://dx.doi.org/10.17654/MS110010217.			
Indexé par *:* **Scopus.**http://www.scimagojr.com/journalsearch.php?q=3900148513&tip=sid&clean=0 **Thomson ISI.**http://ip-science.thomsonreuters.com/mjl/				

CHAPITRE 1 : Concepts Généraux : la fouille des motifs

Résumé du chapitre 1 *: Dans ce chapitre 1, nous abordons les principes fondamentaux du Datamining. Il est subdivisé en quatre grandes sections : les notions mathématiques et statistiques de base de la fouille des données, le concept formel, les règles d'association, et la description méthodologique de la fouille de données.*

Sommaire : Pages

1.1 Introduction 35

1.2 Data Mining : Concepts et principes 36

1.3 Fondement mathématique : Notions de base 38

1.4 Règles d'Association et Implication 45

1.5 Méthodologie de fouille de données 59

1.6 Conclusion 63

1.1 Introduction

La découverte et la gestion de motifs font référence à un ensemble d'activités de prétraitement, d'extraction, de manipulation et de stockage de motifs à partir de données. Un motif (pattern) fait partie du résultat d'une fouille de données laquelle est une étape du processus d'extraction de connaissances à partir des données. En analyse formelle de concepts, le motif prend deux principales formes :

a. Des concepts formels décrivant des objets/individus avec leurs attributs communs et représentant des nœuds d'un treillis de concepts (Galois), et

b. Des règles d'association entre des groupes d'attributs, y compris des implications.

Cette partie du mémoire a pour but de brosser un portrait des éléments qui permettent de faire de la prédiction de données ou de nouvelles découvertes grâce aux données. En effet, l'exploration de données, également connue sous le nom de découverte de connaissances ou extraction de connaissances à partir de données peut prendre de multiples formes ; permettant de délivrer à des experts divers types de connaissances. À cet égard, les règles d'association, leurs variantes étendues plus particulièrement et les motifs graduels sont des modèles fréquemment fournis aux utilisateurs finaux pour en extraire des connaissances selon le domaine défini.

Ces types de connaissances consistent à mettre en évidence des schémas récurrents et relationnels dans les données. Ils se distinguent par le type de corrélation exprimée et la nature des données à partir desquelles ils sont extraits.

Le présent chapitre du mémoire est structuré comme ce qui suit :

- Concept et principe du Datamining (Fouille de données) ;
- Fondements mathématiques relatifs à la fouille de données ;

- Règles d'associations et implications ;
- Méthode de fouille de données.

1.2 Data Mining : Concepts et principes

1.2.1 Composantes du Data Mining

Un système type d'Extraction de Connaissance dans les Bases de Données (ECBD) [Han, 1998], [Simon, 2000], s'articule autour des composantes suivantes :

- Une ou plusieurs bases de données et leurs systèmes de gestion, pour le traitement des masses de données volumineuses.
- Une base de connaissances qui permet à la fois la gestion des connaissances et la résolution des problèmes liés au domaine des données. Le système utilise une base de connaissances (par exemple une ontologie du domaine) qui est enrichie grâce aux nouvelles connaissances inférées par le système.
- Un système de Fouille de Données pouvant s'appuyer sur des techniques symboliques comme l'extraction des règles d'association [Agrawal, 1994], la classification par treillis de Galois [Barbut, 1970], [Davey, 1994] ou l'induction par des arbres de décision [Breiman, 1984], [Quinlan, 1986].
- Et une interface se chargeant des interactions avec l'analyste et de la visualisation des résultats. L'analyste et l'expert sont chargés de guider les recherches et de valider les connaissances extraites.

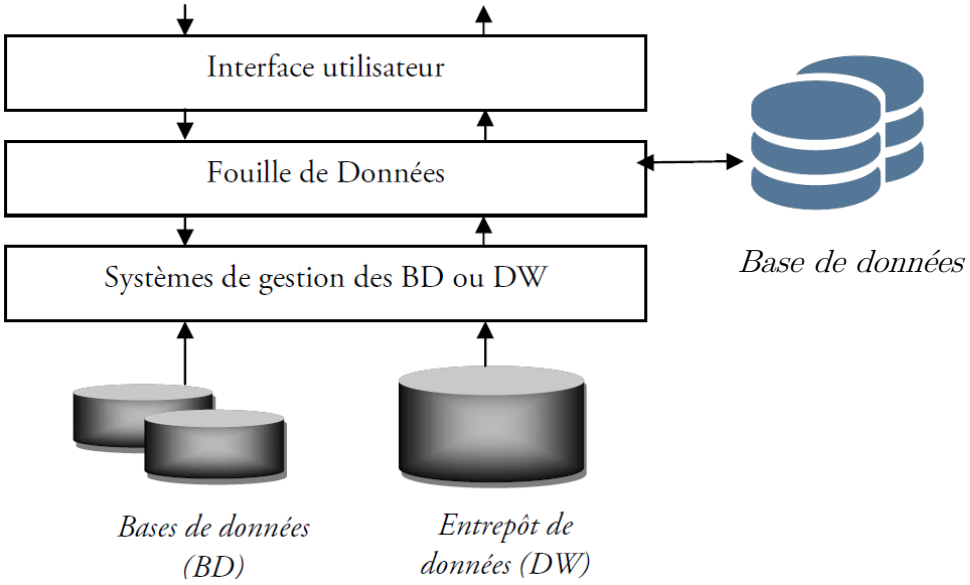

Figure 1.4: Architecture type d'un système d'ECBD [Han, 2000].

1.2.2 Extraction de Connaissances versus Fouille de Données (ECBD vs FD)

L'ECBD est un processus d'extraction de connaissances à partir de bases de données. Il consiste à analyser des données brutes pour en extraire des connaissances exploitables. Ces dernières vont permettre à un expert d'avoir une vision synthétique du domaine étudié. Le processus est dirigé par un analyste qui selon ses objectifs va appliquer des méthodes de fouille de données sur des données préalablement sélectionnées pour en déduire des modèles.

Il existe une confusion entre les concepts Fouille de Données et l'ECBD, certains auteurs les considèrent comme synonymes. Or, la FD n'est qu'une des étapes du processus d'ECBD. Elle correspond à l'application des méthodes et techniques d'extraction de connaissances.

La figure 1.3 illustre bien la différence entre les notions de *l'extraction de Connaissances dans les Bases de Données* (*ECBD*) et la *Fouille de de Données* (*FD*).

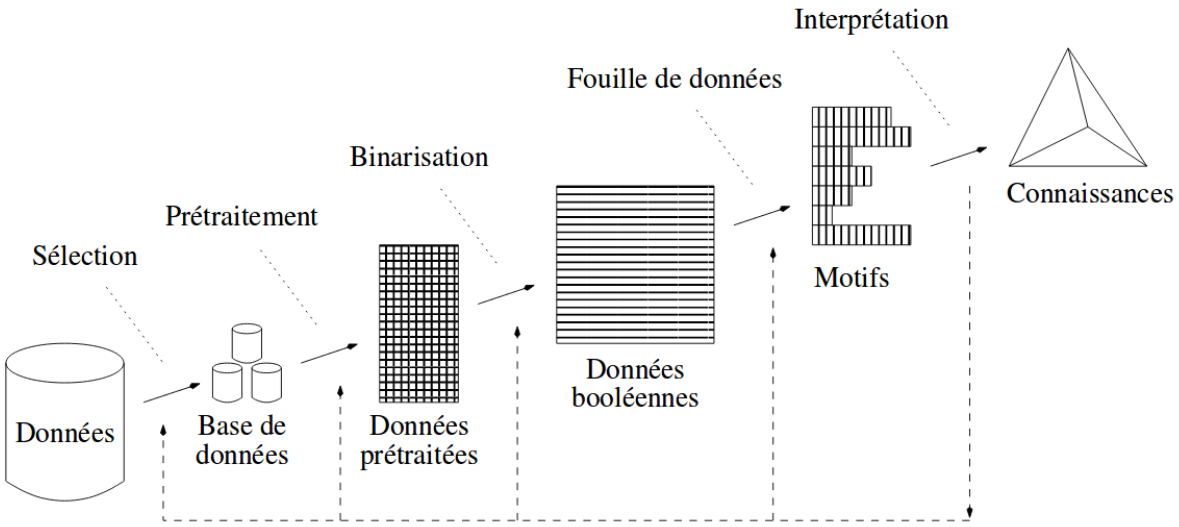

Figure 1.5 : Processus d'extraction de connaissances [Fayyad, 1996].

Dans les sections qui suivent, nous utilisons le cadre théorique de l'analyse de concepts formels présentés dans [Ganter, Wille, 1999], en rappelant de manière succincte les notions de base de ce cadre théorique de l'extraction de connaissances dans les bases de Données.

1.3 **Fondements mathématiques :** Notions de base

L'Analyse de Concepts Formels (ACF), appelée aussi Analyse Formelle de Concepts (AFC), est un formalisme qui constitue un pont entre les mathématiques, en particulier la théorie des ensembles ordonnés, et les applications d'analyse de données [Wille, 1982], [Ganter and Wille, 1999]. C'est un formalisme de représentation de la connaissance, basé sur la formalisation des concepts et la hiérarchie de concepts [Missaoui and Kwuida, 2011]. Un problème récurrent en AFC est le nombre de concepts qui peut être exponentiel en fonction de la taille du contexte. Pour contrôler le volume du contexte et celui du treillis de Galois correspondant, plusieurs techniques ont été proposées [Ayadi et al., 2009]. Une de ces techniques est de montrer seulement les concepts fréquents du treillis à la place du treillis complet [Kaytoue et al., 2011]. Ce dernier s'est avéré être un cadre théorique intéressant pour la fouille de données puisqu'il permet la génération de concepts (regroupement conceptuel) et de règles d'associations [Nehmé et al., 2005] et [Le Floc'h, et al., 2003].

1.3.1 Contexte formel

Un contexte formel est un triplet $\mathsf{B} = (O, I, M)$; B décrivant un ensemble fini O d'objets (ou transaction), un ensemble fini I d'items (ou attributs) et une relation (d'incidence) binaire M(c'est-à-dire $M \subseteq O \times I$. L'appartenance du couple ensemble(o, i) à M désigne le fait que l'objet $o \in O$ contient l'item $i \in I$.

Un contexte formel peut être représenté sous la forme d'un tableau où les lignes correspondent aux objets et les colonnes correspondent aux attributs. Les cases du tableau sont remplies comme suit : si le $i^{\text{ème}}$ objet g est en relation M avec le $j^{\text{ème}}$ alors la case intersection de la ligne i et la colonne j contient "×" sinon la case est vide. Nous illustrons dans le tableau 1.1 un exemple de contexte formel décrivant six transactions (clients) relatives à l'achat des articles $a, \ldots, f$ par des clients, un cas typique de l'analyse du panier de consommateur (Basket Market analysis).

Tableau 1.1: Contexte d'extraction B

O	a	b	c	d	e
1	×		×		×
2		×		×	
3			×	×	
4		×	×		
5	×	×			
6	×	×		×	×

Comme on le voit dans le tableau 1.1, le contexte d'extraction est représenté sous forme d'un tableau où les objets sont en lignes et les attributs en colonnes. L'intersection des lignes et des colonnes représente la relation binaire M entre les objets et les attributs.

1.3.2 Correspondance de Galois

Soit le contexte d'extraction$B = (O, I, M)$. Soit l'application Φ de l'ensemble des parties de O (c'est-à-dire l'ensemble de tous les sous-ensembles de O), noté par $P(O)$, dans l'ensemble des parties de I, noté par $P(I)$. L'application Φ associe à un ensemble d'objets $o \subseteq O$, l'ensemble des items $i \in I$ communs à tous les objets $o \subseteq O$

$$\Phi: P(O) \rightarrow P(I)$$

$$\Phi(0) = \{i \in I | \forall\, o \subseteq O, (o, i) \in M\}$$

Soit l'application Ψ de l'ensemble des parties de I dans l'ensemble des parties de O. Elle associe à un ensemble d'items $I \subseteq I$, l'ensemble d'objets $o \in O$ communs à tous les items $i \in I$.

$$\Psi : P(I) \rightarrow P(O)$$

$$\Psi(0) = \{o \in O | \forall\, i \subseteq I, (o, i) \in M\}$$

Le couple d'applications (Φ, Ψ) définit une correspondance de Galois entre l'ensemble des parties de O et l'ensemble des parties de I. Les applications $\gamma = \Phi \circ \Psi$ et $\omega = \Psi \circ \Phi$ sont appelées les opérateurs de fermeture de la correspondance de Galois [Ganter, Wille, 1999]. L'opérateur de fermeture γ, tout comme ω, est caractérisé par le fait qu'il est :

- Isotone :$I_1 \subseteq I_2 \Rightarrow \gamma(I_1) \subseteq \gamma(I_2)$;
- Extensive :$I \subseteq \gamma(I)$;
- Idempotent : $\gamma\big(\gamma(I)\big) = \gamma(I)$.

Définition 1.1 : Un motif est un sous-ensemble de I. On dit qu'un motif I est inclus dans l'objet o (ou que o contient I) si I et o sont en relation : $\forall i \in I, (o, i) \in M$.

Un motif de taille k est noté k-motif. Les motifs sont aussi appelés ensembles d'items *(« itemsets » dans la littérature anglo-saxonne). A titre illustratif, {item1, item2, item3, item5}* représente un *4-Itemset (où k=4)*

Définition 1.2 : Un motif $I \subseteq I$ est dit fréquent si son support relatif, $\text{Supp(I)} = \frac{|\Psi(I)|}{|O|}$,dépasse un seuil minimum fixé par l'utilisateur et noté *MinSupp.* Notons que $|\Psi(I)|$ est appelé support absolu de I.

Il est à noter que le support des motifs est anti-monotone par rapport à l'inclusion ensembliste, c'est-à-dire que si $I_1 \subseteq I_2$, alors $Supp(I_1) \geq Supp(I_2)$. Dans la suite et pour simplifier l'explication, *Supp(I)* désignera le support absolu du motif *I*.

1.3.3 Concept formel

Les paires de fermés reliées par cette connexion de Galois, détaillée dans la section précédente, forment les concepts formels définis comme suit.

Définition 1.3 : Un *concept formel* est une paire $c = (O, I) \in O \times I$, où O et I sont reliés par les opérateurs de la correspondance de Galois, c'est-à-dire $\Phi(O) = I$ et $\Psi(I) = O$. Les ensembles O et I sont appelés respectivement extension et intension de c.

1.3.4 Théorie des treillis : Notion de base

1.3.4.1 Ensemble ordonné

Définition 1.4 : (Relation binaire) Une *relation binaire R entre deux ensembles M et N est un ensemble de couples d'éléments* m, n *tels que* $m \in M$ *et* $n \in N$*, c'est à dire un sous ensemble de* $M \times N$. $(m, n) \in R$ *(Aussi noté par* mRn *signifie que l'élément m est en relation R avec l'élément n. Si M = N, on parle de relation binaire sur* M. R^{-1} *est la relation inverse de R, i.e. la relation entre N et M telle que* $nR^{-1}m \Leftrightarrow mRn$.

Définition 1.4 : (*Relation d'ordre (partiel)) Une relation binaire R sur un ensemble E est dite relation d'ordre partiel (ou simplement relation d'ordre) sur E si elle vérifie les conditions suivantes pour tous* $x, y, z \in E$:

1. $(x, x) \in R$, *(R est réflexive)*
2. *si* $(x, y) \in R$ *et* $x \neq y$ *alors* $(y, x) \notin R$ *(R est antisymétrique)*
3. *si* $(x, y) \in R$ *et* $(y, z) \in R$ *alors* $(x, z) \in R$ *(R est transitive)*

Une relation d'ordre R est souvent notée par $\leq$ (R^{-1} est noté par " $\geq$ ") et on dit que "x est plus petit que y" *lorsque* $x \leq y$.

Définition 1.5 (***Ensemble ordonné***) *Un ensemble partiellement ordonné (ou simplement ensemble ordonné) est un couple* $(E, \leq)$ *où E est un ensemble et "*$\leq$*" est une relation d'ordre sur E.*

Dans un ensemble ordonné$(E, \leq)$, deux éléments x et y de E sont dits comparables lorsque $x \leq y$ ou $y \leq x$, autrement ils sont dits **incomparables**. Pour deux éléments comparables et différents, $x \leq y$ et $x \neq y$, on note $x<y$. Un sous ensemble de $(E, \leq)$, dans lequel tous les éléments sont comparables est appelé **chaîne**. Un sous ensemble de $(E, \leq)$ dans lequel tous les éléments sont incomparables est appelé **anti-chaîne**.

Définition 1.6 (**Successeur, prédécesseur, couverture**) Soient $(E, \leq)$, un ensemble ordonné et $x, y \in E$. y est dit successeur de x lorsque $x<y$ et il n'existe aucun élément $z \in E$ tel que $x < z < y$. Dans ces cas, x est dit prédécesseur de y et on note $x \prec y$. Lorsque x est un **prédécesseur** de y on dit que x couvre y (et que y est couvert par x). La **couverture** de x est formée par tous ses successeurs.

Tout ensemble ordonné, $(E, \leq)$, peut être représenté graphiquement par un diagramme appelé "**diagramme de Hasse**" (ou diagramme de couverture) et obtenu comme suit :

1. Tout élément de E est représenté par un petit cercle dans le plan
2. Si $x, y \in E$ et $x \prec y$ alors le cercle correspondant à y doit être au-dessus de celui correspondant à x et les deux cercles sont reliés par un segment ;

À partir d'un tel diagramme on peut lire la relation d'ordre comme suit : $x < y$ si et seulement s'il existe un chemin ascendant qui relie le cercle correspondant à x à celui correspondant à y.

Définition 1.7 (*Principe de dualité des ensembles ordonnés*) Soit $(E, \leq)$ un ensemble ordonné. La relation inverse "$\geq$" de "$\leq$" est aussi une relation d'ordre sur E. "$\geq$" est appelée duale de "$\leq$" et $(E, \geq)$ est appelée le dual de l'ensemble ordonné $(E, \leq)$.
Le diagramme de Hasse de $(E, \geq)$ peut être obtenu à partir de celui de $(E, \leq)$ par une simple réflexion horizontale. De plus, il possible de dériver les propriétés duales de $(E, \geq)$ à partir des propriétés de $(E, \leq)$.

1.3.4.2 Treillis de concept

Définition 1.8 (*Majorant, minorant, supremum, infimum*) Soient $(E, \leq)$, un ensemble ordonné et S un sous ensemble de E. Un élément $a \in E$ est dit ***majorant*** de S lorsque $a \geq s \; \forall s \in S$. De façon duale, $a \in E$ est dit minorant de S lorsque $a \leq s \; \forall s \in S$.
Le plus petit majorant (respectivement minorant) de S, s'il existe, est appelé ***supremum*** *ou borne supérieure (respectivement infimum ou borne inférieure) de S et noté* $\bigvee S$ *(respectivement* $\bigwedge S$*). Dans le cas ou* $S = \{x, y\}, \bigvee S$ *et* $\bigwedge S$ *sont aussi notés par* $x \vee y$ *et* $x \wedge y$ *respectivement.*
Dans tous ensembles ordonnés, lorsque le supremum (*respectivement l'infimum*) existe, il est unique.

Définition 1.9 : La relation "$\leq$" permet d'organiser les concepts formels en un treillis complet $(K(O,I,M),\leq)$ appelé ***treillis de concepts*** ou encore ***treillis de Galois*** [Birkhoff, 1967] et noté par $\underline{K}(O,I,M)$ ou $\underline{K}(B)$. Le supremum et l'infimum dans $\underline{K}(B)$ sont donnés par :

$$\bigwedge_{j\in j}(\Phi_j,\Psi_j) = (\bigcap_{j\in j}\Phi_j\left(\bigcup_{j\in J}\Psi_j\right)'')$$

$$\bigwedge_{j\in j}(\Phi_j,\Psi_j) = (\left(\bigcup_{j\in j}\Phi_j\right)'',\bigcap_{j\in j}\Psi_j)$$

Le treillis de concepts est une représentation équivalente des données contenues dans un contexte formel qui met en avant les groupements possibles entre objets et attributs ainsi que les relations d'inclusion entre ces groupements. De plus, la représentation graphique du treillis de concepts, sous la forme d'un diagramme de Hasse, facilite la compréhension et l'interprétation de la relation entre les objets et les attributs d'une part et entre objets ou attributs d'autre part. L'avantage de cette représentation est qu'à partir d'un treillis de concepts il est toujours possible de retrouver le contexte formel correspondant et inversement.

Le diagramme de Hasse représentant un espace dans le cas d'un ensemble de cinq items $A = \{a,b,c,d;e\}$ est illustré dans la figure 1.1, où les nœuds sont les motifs et les arrêtes correspondent aux relations d'inclusion.

Id	Motif
1	{A,C,D}
2	{B,C,E}
3	{A,B,C,E}
4	{B,E}
5	{A,B,C,E}
6	{B,C,E}

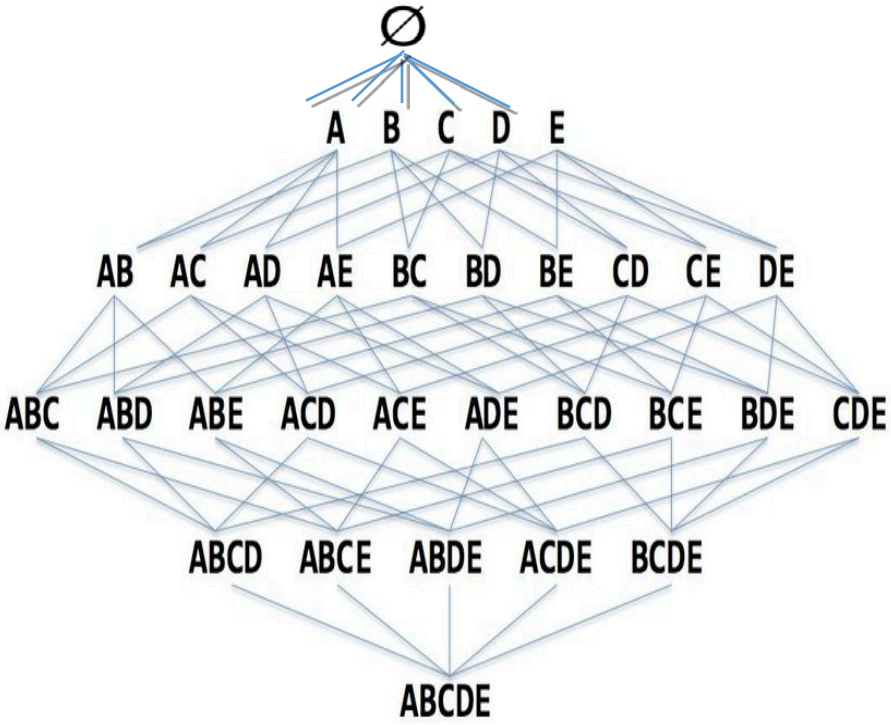

Figure 1.6 : Treillis des parties associé à A = {A,B,C,D,E}

1.4 Règles d'association et implications

Les Règles d'association ont été créées pour extraire de la connaissance à partir de données. La génération de règles d'association [Agrawal et al., 1993] est un aspect important de la fouille de données. Elle a été initialement introduite en analysant le panier du consommateur par Agrawal mais peut être valablement appliquée à divers domaines. Elle consiste à extraire des associations ou relations entre les attributs (*items ou produits*) d'un entrepôt de données. Ainsi, en découvrant les associations qui relient les différents attributs d'une base de données, on peut alors évaluer leur importance principalement avec deux indicateurs connus : **le support** et **la confiance** d'une règle d'association.

La définition de la règle d'association varie selon trois principaux courants initiés par les auteurs suivants :

- **Gras** définit des règles d'implication statistique pour aider les didacticiens à trouver des relations entre les acquisitions de notions élémentaires chez les élèves d'une classe [Gras, 1979],

- **Guigues et Duquenne** se sont plutôt intéressés à une représentation ordonnée de concepts avec les implications informatives [Guigues et Duquenne, 1986].
- Et enfin **Agrawal et Srikant** ont privilégié l'extraction optimisée de règles d'association dans de grandes bases de données [Agrawal et Srikant, 1996].

Par la suite, ces définitions ont connu des extensions dans plusieurs directions. La binarité des propriétés n'est plus obligatoire, on peut maintenant faire des RA avec des propriétés numériques [Guillaume 2000], [Cadot et Napoli. 2004]. Pour éviter l'explosion du temps d'extraction des règles, due à celle de la capacité de stockage des données, des algorithmes plus performants ont été proposés [Pasquier, 2000]. La sémantique des règles a été affinée grâce à de nombreux indices de qualité [Guillet 2004], ce qui aide l'utilisateur à choisir les règles les plus adaptées à ses besoins. La navigation ainsi que l'interrogation par un langage adapté ont été mises au point par Botta et al. [Botta et al. 2002] pour faciliter l'exploration de règles d'association.

Dans la littérature, plusieurs approches sont proposées pour améliorer l'efficacité des règles d'association. Sur ce point, nous invitons le lecteur à consulter [Ceglar 2006] pour une description complète sur les principes et les algorithmes. La recherche des règles d'association est un procédé important dans le Data Mining. L'objectif est de découvrir des associations ou des corrélations intéressantes entre des éléments dans ces grandes collections et bases de données.

Définition 1.10 : Une règle d'association r a la forme $X \rightarrow Y$ [*sup*, *conf*], où X et Y sont des sous-ensembles d'attributs appelés itemsets, $X \cap Y = \emptyset$, et *sup* et *conf* représentent respectivement le support et la confiance de la règle. Une corrélation de co-occurrence [Plasse et al., 2008], c'est-à-dire que les transactions ou requêtes qui contiennent l'ensemble des objets X ont tendance à inclure les objets de l'ensemble Y [Wang et al., 2012].

Le support de la règle $r: X \rightarrow Y$ [*sup, conf*] est *la probabilité*$(X \cup Y)$. Il représente la proportion des objets ayant simultanément les attributs X et Y. La confiance de r représente la probabilité conditionnelle Prob(Y / X). C'est donc la probabilité d'avoir Y lorsque X est présent dans le contexte $\mathbb{K}$.

1.4.1 Règles d'association à partir de données binaires

Les règles d'associations ont été créées pour extraire de la connaissance à partir de données, et sont, généralement, de la forme "Si<*Antécédent*>, alors<*Conséquent*>FinSi". Elles ont été introduites par Agrawal et al. (1993) avec pour but de découvrir des relations significatives entre attributs binaires (présence ou absence de l'attribut). Un exemple classique de règles d'associations est l'approche décrite par le panier de la ménagère par Agrawal et Srikant dans [Agrawal et Srikant 1994], donne une vue sur un ensemble d'achats effectués au supermarché. En considérant deux articles X et Y, la règle du type $X \rightarrow Y$ signifie que "si l'article X est présent dans le panier de la ménagère alors il y a aussi l'article Y. Par exemple, « si on achète de la viande, alors on achète du poisson ». Cette règle indique que les clients qui achètent de la viande ont également tendance à acheter du poisson. De façon classique, les règles d'association s'appliquent à un ensemble de données dites transactionnelles : chaque transaction contient une liste d'items. Dans l'exemple des ventes de supermarché, les items correspondent aux produits achetés et la transaction à un ticket de caisse. Cette base transactionnelle est représentée par une base de données binaires où les attributs correspondent aux items possibles. Ils prennent des valeurs binaires soit 0, Soit 1, indiquant respectivement soit l'absence ou la présence de cet item dans la transaction correspondante.

1.4.1.1 Représentation des données

Les données issues des différents documents et bases de données transactionnelles peuvent être représentées sous la forme d'une matrice booléenne à deux dimensions. Dans une telle base, chaque tuple représente une transaction tandis que les différents champs correspondent aux objets inclus dans la transaction. On note par n le nombre de transactions, par p le nombre d'articles, par 0 l'évènement d'absence de chaque article et par 1 sa présence dans la transaction. De ce fait on construira une matrice binaire de la base de données.

Ce tableau représente une matrice creuse a deux dimensions $n * p$. Avec $n = 5$ (Le nombre de transactions), et $p = 7$. (Le nombre d'items). T Représente les différentes transactions

Tableau 1.2: Exemple d'une Base de Données binaires

T	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
T_1	0	1	1	1	0	1	1
T_2	0	0	0	1	1	1	0
T_3	1	0	1	0	0	0	1
T_4	0	0	0	0	1	1	1
T_5	0	1	1	1	0	0	0

1.4.1.2 Critères de qualités

Pour extraire les règles d'association pertinentes, on se base sur des critères de qualité qui capturent différentes définitions de pertinence. Les plus classiques sont le support et la confiance [Agrawal et al., 1993] et [Agrawal & Srikant, 1994], dont nous rappelons les définitions ci-dessous, le tableau 1.1 présente une liste plus complète [Lallich & Teytaud, 2004] et [Lenca et al., 2004]. De nombreuses études comparatives de ces critères, que nous

ne détaillons pas ici, ont été menées [Hilderman & Hamilton, 2001] ; [Lenca et al. 2003; 2004]et[Lallich & Teytaud, 2004].

Soit$A$ un item,M un motif et n le nombre total de transactions dans la base, on note respectivement $n(A), n(M)$ et $n(\bar{A})$ le nombre de transactions qui contiennent respectivement A, le nombre de celles qui contiennent tous les attributs composant M, et le nombre de transactions qui ne contiennent pas $A(\bar{A}$ représente l'absence de l'item A).

Définition 1.11 : Support d'un motif

Le support d'un Itemset représente le nombre total des transactions d'une base de données comportant cet Itemset divisé par le nombre total des observations de cette base de données [Lallich et Teytaud, 2004]. Le support (*Supp*) d'un motif est défini par :

$$Supp\ (M) = \frac{n(M)}{n}$$

Le support d'un motif est donc le rapport de la cardinalité de l'ensemble des transactions qui contiennent tous les items de M par la cardinalité de l'ensemble de toutes les transactions. Il capture la portée du motif, en mesurant sa fréquence d'occurrence.

Ou encore (avec *Card* : Cardinalité)

$$Supp\ (M) = \frac{Card\ (M)}{n}$$

Définition1.12 : (*Motif-Itemset fréquent) On dit qu'un Itemset M est un itemset fréquent si et seulement si le support associé à cet Itemset est supérieur à un support minimum défini par l'utilisateur* [Lallich et Teytaud, 2004].

Remarque : Dans la suite, X et Y sont considérés comme des *Motifs-Itemset fréquent*

Définition 1.13(*Support d'une règle association*) *Les notions de support et de confiance ont été identifiées lors des premières études de recherche des règles d'association menées par Hajek et al. en l'occurrence la méthode GUHA* [Hajek et al., 1966].

Le support d'une règle $R = X \rightarrow Y$ est la proportion de transactions contenant à la fois là :

Il est défini par :$Supp\ (R) = Supp(X \cup Y)$

$$Supp\ (R) = \frac{Card(X \cup Y)}{n}$$

Avec Supp : la métrique du support

Remarque : La mesure du support est une mesure symétrique.

Elle évalue les règles $X \rightarrow Y$ et $Y \rightarrow X$ de manière équivalente.

Par exemple : la règle suivante "$Lait \rightarrow Pain$", littéralement **Si** Lait **alors** Pain. Ici le support représente le nombre de transaction dans lesquelles on trouve les Items « Lait » et « Pain », divisé par le nombre total de transaction.

*Définition 1.14 (*Confiance d'une règle association) La confiance d'une règle $\boldsymbol{R = X \rightarrow Y}$ *est définie comme la proportion de transactions contenant le conséquent parmi celles qui contiennent la prémisse. Elle peut être interprétée comme probabilité conditionnelle* $\boldsymbol{P(Y/X)}$ *et calculée à partir des supports* [Lallich et Teytaud, 2004].

$$Conf\ (R) = \frac{Supp(X \cup Y)}{Supp\ (X)}$$

$X \cup Y$ représente l'ensemble union contenant les éléments de la transaction X et les éléments de la transaction Y. Contrairement à la mesure de support, la mesure de causalité est imposée et elle capture sa précision. La confiance n'est pas symétrique. Elle évalue la qualité d'une règle où une relation de causalité est imposée et elle capture sa précision.

Il existe d'autres métriques de mesure de qualité de règles d'associations que nous ne détaillerons pas entre les lignes de ce mémoire de thèse. En effet, dans notre projet de thèse, nous nous sommes restreints aux mesures de support et de confiance pour deux raisons principales :

- Premièrement, notre travail est orienté vers la sémantique et l'interprétabilité des motifs et non vers les mesures de qualité ;
- Deuxièmement, les approches que nous proposons sont basées sur les algorithmes d'extraction fondés sur ces deux mesures de qualité.

Aussi, nous invitons le lecteur à consulter [Lallich et Teytaud, 2004] pour plus d'information sur les autres métriques de mesures présentés dans le tableau 1.3

Tableau 1.3: Principale mesures de qualité d'une règle d'association $X \rightarrow Y$

Mesure de qualité	Formule Mathématique
Confiance	$\frac{n(XY)}{n(X)}$
Confiance centrée	$\frac{n(XY)}{n(X)} - n(Y)$
Pearl	$n(X)\lvert\frac{n(XY)}{n(X)} - n(Y)\rvert$
Piatetsky-Shapiro	$n \times n(X)(\frac{n(XY)}{n(X)} - n(Y))$
Lœvinger	$\frac{\frac{n(XY)}{n(X)} - n(Y)}{n(\bar{Y})}$
Zhang	$\frac{n(XY) - n(X)n(Y)}{\max\{n(XY)n(\bar{Y}); n(Y)(n(Y)n(X\bar{Y})\}}$
Corrélation	$\frac{n(XY) - n(X)n(Y)}{\sqrt{n(X)n(\bar{X})n(Y)n(\bar{Y})}}$
Indice d'implication	$\sqrt{n}\frac{n(X\bar{Y}) - n(X)n(\bar{Y})}{\sqrt{n(X)n(\bar{Y})}}$
Lift	$\frac{n(XY)}{n(X)n(Y)}$
Suprise	$\frac{n(XY) - n(X\bar{Y})}{n(Y)}$
Conviction	$\frac{n(X)n(\bar{Y})}{n(X\bar{Y})}$
Sebage-Schœnauer	$\frac{n(X)}{n(X\bar{Y})}$
Multiplicateur de cote	$\frac{n(XY)n(\bar{Y})}{n(X\bar{Y})n(Y)}$
J-mesure	$n(XY)log\frac{n(XY)}{n(X)n(Y)} + (X\bar{Y})log\frac{n(X\bar{Y})}{n(X)P(Y)}$

1.4.2 Règles d'associations quantitatives

Le problème originel de la recherche de règles d'associations est l'extraction de corrélations à partir de données binaires. Or les bases de données réelles contiennent non seulement des variables catégorielles mais aussi des variables numériques, discrètes ou pseudo-continues. Les règles d'associations classiques ne peuvent donc pas leur être appliquées directement et ont été étendues aux règles d'associations quantitatives [Agrawal & Srikant, 1994] et [Miller & Yang, 1997] qui visent à exprimer des corrélations pour de telles données. Les approches proposées pour cette extraction dans la littérature peuvent être classées en trois catégories principales que nous détaillons ci-dessous et comparons ensuite dans le tableau 1.2. Nous commençons par les approches basées sur une discrétisation préalable, puis les approches guidées par des schémas de règles, et enfin les approches fondées sur un algorithme génétique.

- *Approches fondées sur une discrétisation préalable ;*
- *Approche guidée par des schémas de règles.*

Cette approche proposée par [Fukuda et al., 1996a ; 1996b] n'identifie pas une discrétisation des attributs numériques mais extrait des intervalles particuliers, dits intervalles d'intérêt, qui satisfont des contraintes de pertinence : elle est basée sur une optimisation de critères de qualité mesurant cette pertinence. L'évaluation des intervalles d'intérêts candidats dépend de la qualité des règles qu'ils induisent selon, par exemple, le support, la confiance ou le gain [Fukuda et al., 1996a ; 1996b].

Afin de limiter le coût de calcul, certaines approches sont fondées sur des schémas de règles restreints : un schéma de règle est une règle présentant dans chacun de ses membres gauche et droit des items catégoriels aux valeurs fixées et des items numériques dont les intervalles correspondants ne sont pas encore instanciés [Fukuda et al., 1996a ; 1996b], limitant par exemple le nombre d'attributs numériques dans la prémisse et la conclusion. Un tel schéma

de règles peut être illustré par l'exemple ci-contre : « *taille* $\in [t_1, t_2]$ *et (teint = jaune)*→ *(race = asiatique)*» ou les attributs teint et race sont instancies et les valeurs t_1 et t_2 de l'intervalle correspondant à l'attribut âge ne sont pas encore instanciées.

[Aumann et Lindell, 2003] et [Webb, 2001] proposent une autre vision du problème : des statistiques (moyenne, variance, écart-type, minimum etc.) sur des distributions des attributs numériques sont autorisées dans la partie droite d'une règle. Deux sortes de règles sont considérées : la première représente le cas où la prémisse de la règle est un ensemble d'attributs catégoriels et son conséquent un ensemble de statistiques sur les distributions de plusieurs attributs numériques ; la deuxième représente le cas où la prémisse de la règle contient un seul attribut numérique et son conséquent une statistique sur la distribution d'un seul attribut numérique.

- *Approche reposant sur un algorithme génétique*

L'optimisation est également la voie choisie dans les travaux de [Mata et al., 2002], qui propose d'utiliser des algorithmes génétiques : un individu est représenté par une liste de couples (attribut numérique, disjonction d'intervalles). La qualité des individus est évaluée par une mesure permettant d'optimiser le support des motifs, tout en veillant à ne pas retenir les domaines entiers des attributs numériques et à favoriser les motifs les plus spécifiques. Le seul critère optimisé dans cet algorithme est le support, ce qui limite l'applicabilité d'une telle approche.

Une autre approche basée sur l'algorithme génétique suivant une organisation classique a été proposée par [Nortet et al., 2006], [Salleb-Aouissi et al., 2013]. Contrairement à l'approche précédente qui optimise un seul critère qui pourrait être insuffisant, celle-ci cherche le meilleur intervalle pour chaque attribut optimisant le support, la confiance, ainsi que la mesure de gain. Cette approche est basée également sur les schémas de règles, et la

discrétisation obtenue varie donc pour chaque schéma tout en dépendant des attributs catégoriels et numériques qui le composent. Cet algorithme contient plusieurs paramètres à fixer : la taille de la population, le nombre de générations, les taux de mutation et de croisement. En outre, il n'est pas capable d'identifier plusieurs intervalles pertinents.

Tableau 1.4 : Tableau comparatif de trois règles d'associations quantitatives

Approche	Discrétisation	Optimisation	Limites
Discrétisation préalable	Discrétisation complète	Deux étapes	- Perte d'information
Schémas des règles	Intervalle d'intérêt	Une seule étape	- Schémas des règles - Format très limité
Algorithme génétique	Intervalle d'intérêt	Une seule étape	- Schémas des règles limitées à un intervalle - Nombreux paramètres

Ce paragraphe synthétise les travaux précédemment cités, classés selon différents critères listés dans le tableau 1.4 et présentes ci-dessous.

- Les méthodes effectuent-elles une discrétisation complète de l'univers pour identifier les intervalles souhaités ou extraient-elles seulement des intervalles d'intérêts ?
- Les méthodes utilisent-elles une optimisation en une seule étape ?
- La troisième ligne indique les limites de ces méthodes.

Les méthodes reposant sur une discrétisation a priori des attributs quantitatifs optimisent les intervalles d'intérêts en deux étapes, dont une étape de pré-discrétisation préalable induisant une perte d'information. Les approches basées sur les schémas de règles permettent quant à elles d'optimiser les intervalles d'intérêts en une seule étape, pendant la phase de génération des motifs fréquents, mais dans ce cas, le format des règles est souvent très limite. Les approches basées sur l'algorithme génétique optimisent également les intervalles

d'intérêts en une seule étape. Cependant, cette étape d'optimisation n'est pas effectuée pendant la phase de génération des motifs fréquents, mais pendant la phase de génération de règles. L'ensemble de ces approches sont limitées à l'identification d'un seul intervalle d'intérêt : elles n'utilisent pas la disjonction d'intervalles, comme cela est le cas dans les approches basées sur les schémas de règles. Elles reposent de plus sur plusieurs paramètres, pour lesquels il n'est pas aisé de trouver les valeurs optimales.

1.4.3 Règles d'associations floues

Alors que l'extension des règles d'association précédente prend en compte des données numériques, une autre extension vise à traiter des données floues. Considérons par exemple un attribut correspondant au salaire d'un employé. Dans le cas classique, cet attribut est décrit par des valeurs numériques. Dans le cas où, il peut être associé trois modalités floues « faible », « moyen », et « élevé ». L'attribut est ensuite décrit par ses degrés d'appartenance à ces modalités. Un exemple d'une telle règle étendue est : « les employés jeunes et de faible niveau d'études ont des salaires faibles » ou « employés », « niveau d'études » et « salaires » représentent les variables linguistiques et « jeune » et « faible » représentent leurs modalités floues respectives.

Définition 1.15 (Règle d'association floue) : Une règle d'association floue est de la forme générale$M_1 \rightarrow M_2$ *avec* $M_1 = (X, A)$ *et* $M_2 = (Y, B)$ *ou X, Y sont des attributs flous et A, B sont leurs modalités floues respectives.*

Les règles d'associations floues sont interprétées comme une généralisation des règles d'associations appliquées à des données floues, indiquant que la présence floue de M_1implique, au sens de la logique, la présence de M_2 [Hullermeier, 2001]. Ainsi, la règle « plus on est proche du mur, plus on freine fort » peut être considérée comme l'extension floue

d'une règle d'association concernant la présence binaire des attributs distance au mur et freinage.

Le support de la règle est alors calculé comme la somme des contributions de chaque objet a l'implication : une règle est valide si les degrés d'appartenance aux modalités floues impliquées dans la règle satisfont l'implication floue, pour chaque objet de la base de données individuellement.

Définition 1.16 (Support d'une règle d'association floue)

Formellement, le support d'une règle d'association floues est :

$$Supp\ M_1 \rightarrow M_2 = \sum_{o \in D} i\ (M_1(o), M_2(o))$$

Où i est un operateur d'implication résiduel, par exemple l'implication de Goguen définie par : $i\ (a, b) = min\left(1, \frac{b}{a}\right)$si $a \neq 0{,}1$ sinon.

Les auteurs Bosc et al.; et Hullermeier proposent aussi d'étendre le concept de découverte de règles d'associations, de façon à prendre en compte des propriétés graduelles et d'exprimer une contrainte sur les valeurs des attributs apparaissant dans la règle [Bosc et al., 2001] et [Hullermeier, 2001].

On peut noter que ce support ne s'applique pas à un motif, comme dans le cas des règles d'associations classiques, mais à une règle, d'une façon asymétrique qui permet de distinguer $M_1 \rightarrow M_2$ de $M_2 \rightarrow M_1$

Dans la littérature, il y a différentes formes de règles graduelles qui sont distinguées [Dubois & Prade, 1992]; [Hullermeier, 2001] suivant le type d'opérateur d'implication utilisé :

- Les *r*-implications modélisant les règles graduelles floues de la forme « plus X est A, alors plus Y est B » ;
- Les *s*-implications modélisant les règles floues de certitude de la forme « plus X est A, alors plus il est certain que Y est B.

Un autre type de règles, représentant un croisement de règles d'associations floues et de résumés linguistiques a été proposé par Bosc et al. [Bosc et al., 2001]. Il consiste à faire reposer l'interprétation d'une règle d'associations floue sur un calcul de cardinalités floues. Le principe est le suivant : comme dans le cas usuel, la validité de la règle $(X,A) \rightarrow (Y,B)$ dépend du nombre de données qui sont A d'une part et du nombre de données qui sont A et B d'autre part. La différence avec le cas usuel est qu'ici, il faut utiliser une cardinalité étendue puisque les ensembles de données considérées sont décrits par des modalités floues. Dans cette approche, la validité est définie comme le degré de nécessité de l'évènement « Q données vérifient la règle », où Q désigne un quantificateur ou tel que «la plupart » ou « très peu ».

1.4.4 Autre extension des règles d'associations : *Cas des motifs séquentiels*

Les extensions présentées ci-dessus considèrent des données différentes de celles considérées dans le cas classique. Il est important de noter qu'il existe une autre extension qui considère le même type de données que celles traitées dans le cas classique, mais enrichies par un attribut temporel. Il s'agit des motifs séquentiels : l'idée est de fouiller non plus les corrélations entre sous-ensembles de motifs, mais de fouiller les ordres répétitifs entre motifs. Un motif séquentiel est défini comme une liste ordonnée et non vide de motifs [Agrawal and Srikant, 1995]. De tels motifs sont par exemple de la forme : « les clients achètent du pain et du beurre, puis plus tard ils achètent du chocolat ».

De nombreux algorithmes efficaces ont été proposés pour extraire de tels motifs graduels tels que ceux proposés dans les travaux ci-après : [Agrawal et Srikant, 1995], [Masseglia et al., 1998], [Zaki, 2001],[Ayres et al., 2002],[Pei et al., 2004],[Chiu et al., 2004],[Zaki & Hsiao, 2005].

1.5 Techniques de fouille de données

Nous précisions que ces notes sont largement inspirées des travaux de *Devroye et al.* [Devroye et al., 1996], *Vladimir N* [Vladimir N, 1982 ; 1998], et *Hastie et al.* [Hastie et al., 2001]. La fouille de donnés est classiquement décrite comme un processus interactif de préparation des données (sélection de descripteurs, constitution d'une table, discrétisation), d'extraction de connaissances à l'aide d'algorithmes de calcul, de visualisation et d'interprétation des résultats, lors d'interactions avec l'expert. Les méthodes d'exploration proposent des solutions aux problèmes de recherche des règles d'associations. Ainsi, la fouille de données concerne l'étape algorithmiquement difficile de ce processus, qui produit des motifs potentiellement intéressants à partir des données. Les méthodes de forage de données fournissent à l'expert des solutions pour l'aide à la décision. On distingue généralement deux techniques de fouille de données :

(1) **Technique de fouille de données supervisée** : Chaque objet étudié est étiqueté par une valeur de classe. Par exemple, s'il s'agit de données médicales concernant des patients, la classe définit le degré d'atteinte de la maladie. Pour des produits de fabrication industrielle, la classe est déterminée par la qualité de fabrication,

(2) **Technique de fouille de données non supervisée** : Dans cette technique de fouille de données aucune classe n'est attribuée à priori. Suivant le type des données, les décisions concernent :

- *Séquence mining* : elle consiste à proposer une valeur de classe pour un objet dont la classe est inconnue. Un médecin peut ainsi adapter le traitement d'un patient en fonction de ses attributs ;

- *Clustering* : cette méthode permet de constituer des groupes homogènes d'objets, pour par exemple grouper des patients qui ont le même comportement ;
- *Règles associatives* : elles formulent les corrélations présentes dans les données et sont utilisées à des fins de classification ou de caractérisation de classe.

1.5.1 Technique de fouille de données supervisée

La classification est dite supervisée si toutes les données sont étiquetées préalablement.

- Modèle inductif où l'apprenant considère un ensemble d'exemples, et infère l'appartenance d'un objet à une classe en considérant les similarités entre l'objet et les éléments de la classe ;
- La plupart des algorithmes (classification, estimation, prédiction) utilisent l'apprentissage supervisé.

Il existe une multitude de problèmes qui entrent dans le cadre de la classification supervisée, parmi lesquels :

(1) *La reconnaissance et identification de caractères manuscrits* : X est une image en niveaux de gris, i.e. $X = [0,1]^k$ où k est le nombre de pixels. L'étiquette Y indique le caractère représenté par $X = 11$ pour reconnaître les 10 chiffres, $M = 37$ si l'on ajoute les lettres, $M = 63$ en comptant les majuscules, et plus encore si l'on tient compte des accents et autres caractères spéciaux).

(2) Plus généralement, *la reconnaissance de formes* : X est une image (éventuellement en couleurs, donc $X = [0,1]^{3k}$, et Y indique si l'image possède ou non une caractéristique donnée (contenir une voiture, un humain, etc. ; représenter un humain qui sourit, qui pleure, etc.) ; ici, $M = 2$ si la question posée est fermée (oui ou non), mais on peut aussi considérer des questions plus ouvertes.

(3) *La reconnaissance de parole* : X est un enregistrement sonore numérisé, $X = R$ avec très grand et Y indique qui parle (parmi un petit nombre de personnes possibles), ou bien ce qui est dit (*parmi une petite liste de phrases possibles*).

(4) La *catégorisation de textes* : X est un texte (X est l'ensemble des suites finies de caractères), Y indique qui a écrit le texte (*Par exemple, est-ce Shakespeare ? est-ce Corneille ou Racine ?*) ou bien quelle est la thématique principale du texte, etc.

(5) *La détection de spams* : X est le contenu d'un courriel$X \subset \{0,1\}^K$ avec K grand est l'ensemble des fichiers binaires de taille $<$ 5Mo), $Y = 1$ si c'est un spam et $Y = 0$ sinon.

(6) *L'aide au diagnostic médical* : X est un ensemble de caractéristiques du patient (*fréquence cardiaque, température corporelle, résultats d'examens médicaux, âge, sexe, antécédents personnels ou familiaux, etc.*), Y donne une information sur l'état de santé réel du patient (est-il en danger de mort ? faut-il l'opérer ? cela vaut-il la peine de lui faire passer un examen dangereux ou très coûteux ?). Dans le cadre de la cancérologie, on peut par exemple inclure dans X des données d'expressions de gènes (donc : $X = R$ avec très grand), la quantité d'intérêts Y indiquant si le patient est ou non atteint d'un cancer, ou bien de quel type de cancer il s'agit (*va-t-il y avoir des métastases* ?).

La recherche en apprentissage automatique a produit une large gamme d'algorithmes supervisés pour construire des classificateurs [Niharika et al., 2012].

Les méthodes reconnues concernent l'utilisation de *réseaux de neurones, les k plus proches voisins, arbres de décision et les réseaux de bayes.* Dans le chapitre 2, nous y consacrons un état de l'art à la section 2.5 de ces algorithmes.

1.5.2 Technique de fouille de données non-supervisée

La classification non supervisée désigne un corpus de méthodes ayant pour objectif de dresser ou de retrouver une typologie existante caractérisant un ensemble de n observations, à partir de p caractéristiques mesurées sur chacune des observations. Par typologie, on entend que les observations, bien que collectées lors d'une même expérience, ne sont pas toutes issues de la même population homogène, mais plutôt de k populations.

Les données peuvent se présenter sous différentes formes ; elles concernent n individus supposés affectés, pour simplifier, du même poids :

- Un tableau de distances (ou dissimilarités, ou mesures de dissemblance), $n \times n$, entre les individus pris deux à deux ;
- Les observations de p variables quantitatives sur ces individus ;
- Les observations, toujours sur ces n individus, de variables qualitatives ou d'un mélange de variables quantitatives et qualitatives.

D'une façon ou d'une autre, il s'agit, dans chaque cas, de se référer au tableau des distances deux à deux entre les individus (c'est-à-dire au premier cas). Le choix d'une matrice de produit scalaire permet de prendre en compte simplement un ensemble de variables quantitatives tandis que le troisième cas nécessite plus de développements

En classification non supervisée, l'objectif d'une méthode déborde le cadre strictement exploratoire. C'est la recherche d'une typologie, ou segmentation, c'est-à-dire d'une partition, ou répartition des individus en classes homogènes, ou catégories. Ceci est fait en optimisant un critère visant à regrouper les individus dans des classes, chacune le plus homogène possible et, entre elles, les plus distinctes possible. Cet objectif est à distinguer des procédures de discrimination, ou encore de classement (*classification en langue anglaise*) pour lesquelles une typologie est apriori connue, au moins pour un échantillon

d'apprentissage. Ainsi, nous sommes dans une situation d'apprentissage non-supervisé, ou clustering en anglais. Il existe de très nombreuses méthodes de classifications non supervisées, seule la description de quelques-unes fréquemment utilisées et appartenant à des types d'algorithmes différents seront présentés au chapitre 2 dans une revue de la littérature.

Les références de cette section sont basées sur [Forgy, 1965], [Hartigan et Wong, 1979] et [Huang, 1998].

1.6 Conclusion

Dans ce chapitre, les principes et concepts du Data Mining, les notions de base mathématique sur la fouille de données, les techniques de fouille de données notamment les règles d'associations ont été décrits afin de faciliter l'état de l'art dans le chapitre suivant.

CHAPITRE 2 : Analyse Bibliographique : Problématique, Applications et Outils.

Résumé du chapitre *: Ce second chapitre parcourt la littérature sur les travaux relatifs à la fouille de données et aux méthodes ou algorithmes d'extraction des données. Au surplus, les environnements libres des fouilles de motifs sont passés en revue. En fin de chapitre, une analyse des travaux de l'état de l'art est élaborée en positionnant nos travaux de thèse par rapport aux limites observées dans la littérature sur la fouille de données appliquée à l'activité douanière.*

Sommaire : Pages

2.1	Introduction	65
2.2	Problématique de l'étude du problème	65
2.3	Travaux majeurs	66
2.4	Représentations des données structurées	73
2.5	Algorithmes d'extraction de connaissances	75
2.6	Application de la fouille des données	92
2.7	Environnement libre des fouilles de motifs	99
2.8	Analyse des travaux présentés de la littérature	107
2.9	Question de recherches et positionnement de nos travaux	108
2.10	Conclusion	109

2.1 Introduction

La fouille de données est une technique d'exploration de données, qui est apparue avec l'explosion des quantités d'informations stockées dans les systèmes d'informations. Ainsi, la fouille de données vise à découvrir, dans ces masses de données les informations précieuses d'aide à la décision.

Dans notre démarche, l'étude bibliographique est subdivisée en quatre grands points :

- Présentation de la problématique relative la fouille de motifs.
- Différentes méthodes de fouilles de données ;
- Revue bibliographique sur les différents algorithmes de fouille de données.
- Le quatrième point fait l'économie des différents travaux de fouille de données réalisés dans différents secteurs d'activités.

2.2 Fouille de Données : Problématique sur l'étude

Nous présentons ici les notions de base relatives au problème de la fouille de motifs comme initialement introduit dans [Agrawal et al., 1993]. Nous discutons sa complexité temporelle et spatiale, et terminons par mentionner quelques-unes de ses applications typiques et outils libres y afférents. Historiquement, le problème de la fouille de motifs fréquents a été proposé dans le cadre de l'analyse du panier de la ménagère (*Market-Basket Analysis*), dont le but est d'analyser les tickets de caisse des clients dans l'optique de comprendre les habitudes de consommation des clients, d'agencer les rayons du magasin, d'organiser les promotions, de gérer les stocks à l'effet d'améliorer le profit. Cette connaissance des co-occurrences existantes entre produits est ensuite exploitée pour générer ce qui est appelé par la suite les règles d'association qui sont des formes d'implications directes entre des ensembles de produits. Ainsi, la mise en évidence de ces produits corrélés sert dans divers usages comme les campagnes de promotion, l'organisation des rayonnages. Bien qu'élémentaire, ce

problème a connu un succès fulgurant et a conséquemment envahi plusieurs autres tâches d'extraction de connaissances telles que : la classification, la segmentation et le Web Mining pour ne citer que les plus remarquables.

Dans le cadre de cette thèse de doctorat, nous utiliserons cette technique pour étudier la corrélation existentielle entre une opération de dédouanement et la nature de l'infraction dans une activité ordinaire de contrôle en douane dans le but de déterminer des règles d'associations, puis en se basant sur ces règles et l'analyse formelle de concepts (*exploration de la structure symbolique des données*), nous en déduirons la corrélation entre le comportement à risque des opérateurs économiques et les infractions.

2.3 Travaux majeurs

Ces dernières années, la quantité de données contenue dans les bases de données n'a cessé d'augmenter. Les informations s'y référant sont d'une telle envergure qu'il est très difficile de les utiliser à l'état brut. C'est en 1989, lors de la conférence de l'*International Joint Conférence On Artificial Intelligence (*IJCAI) qui se déroulait à Détroit, au Michigan, que plusieurs recherches d'envergure ont mis en avant l'idée d'une méthode pouvant être en mesure de faire ressortir des connaissances de systèmes complexes [IJCAI, 1989]. C'est également lors de cette conférence que la notion de « *découverte de connaissances grâce aux bases de données* » [Usama et Padhraic, 1996] a été mentionnée pour la toute première fois. L'idée sous-jacente de cette notion est de définir un processus pour la découverte d'éléments importants pouvant être, au premier abord, dissimulé sous renfermés des informations capitales pour des études de prédiction.

Ce n'est que quatre ans plus tard, soit en 1993 que les chercheurs Agrawal et al. ont proposé cette méthode dite du *problème de la fouille de motifs* (*fréquents*). Dans le cadre de l'analyse

du panier de la ménagère (*Market-Basket Analysis*), dont le but est de rechercher la collection de produits corrélés souvent achetés ensemble (*fréquence*), en examinant les tickets de caisse enregistrés par les supermarchés.

Les recherches autour de l'extraction de motifs s'attaquent à deux grands défis qui sont :

- La définition de méthodes et d'outils permettant d'appréhender de très grands volumes de données ;
- La sélection des motifs potentiellement intéressants.

L'extraction de connaissances sous contraintes de motifs est un paradigme permettant de découvrir des informations très précieuses. Les contraintes permettent à l'utilisateur de cibler les connaissances qu'il considère comme importantes en réduisant le nombre de motifs intéressants. Il existe des approches génériques pour l'extraction des informations sous la contrainte de motifs fréquentiels, de motifs ensemblistes et de motifs séquentiels [De Raedt et al., 2002] ; [Soulet et Crémilleux, 2005] ; [Pei et al., 2002] ; [Garofalakis et al., 1999] ; [Leleu et al., 2003]. L'extraction de connaissances sous contraintes exige l'extraction optimale des règles d'association qui induisent la qualité des connaissances extraites.

En effet, l'extraction de motifs peut renvoyer une collection importante de règles d'association à exploiter par un utilisateur. Ceci est dû à la présence d'un nombre très important de motifs inintéressants dont l'extraction dans des grands volumes de données est coûteuse pour l'exécution optimale des algorithmes d'extraction. Ainsi, les contraintes sont extrêmement utiles pour améliorer à la fois la qualité des motifs extraits et le processus d'extraction de connaissances.

Le framework fouille de motifs est un modèle fondamental appliqué à plusieurs domaines d'activités. En effet, ce problème repose sur une relation du type plusieurs-à-plusieurs entre deux ensembles : des objets (*transactions*) décrits par des propriétés (items). C'est pour

cette raison que ce modèle est souvent représenté par une matrice à deux dimensions : les objets dans les lignes et les items au niveau des colonnes. Nous rapportons dans les points suivants une sélection de ces applications subdivisées en deux grandes catégories :

- Applications communes de fouille de données (classification, segmentation, analyse d'association ou d'exception (outliers)), [Aggarwal et al., 2014]
- Applications génériques [Han et al., 2007] ; [Leskovec et al., 2014]; [Fournier-Viger et al., 2017].

Le développement d'outils d'analyse et de traitement de l'information numérique intégrant les fonctionnalités de fouille de données représentent de nos jours un domaine important de recherche en milieux académique et industriel. Les domaines d'activités de la bio-informatique et du génie biomédical ont perçu très tôt la pertinence d'intégrer des opérations de fouille de données dans leurs pratiques. Depuis quelques années, plusieurs projets de recherches en sciences humaines et sociales explorent les modalités d'application du datamining en tenant compte des particularités d'analyse propre à chaque discipline [Forest, 2006].

2.3.1 Fouille des données en météorologie et en astrologie

Les travaux de fouille de données appliquée à la météorologie sont basés sur les connaissances de phénomènes naturels, et ont pour objectif de prédire le temps [Jayanta, 2004].

En astronomie et en astrophysique, la fouille de données spatiales sert à la classification automatique d'objets spatiaux, et à la découverte des régions dignes d'intérêt, ou des objets rares de l'univers. [Christian Bohm et Claudia, 2010].

2.3.2 Fouille des données en Bio-informatique

La fouille des données en bio-informatique répond à l'identification de motifs pertinents à partir de l'analyse des données biologiques afin de prédire la structure secondaire des protéines et leur fonctionnalité.

Ainsi Atluri et al. et Naulaerts et al. [Atluri et al., 2009], [Naulaerts et al., 2015], [Naulaerts et al., 2017] présentent quelques applications typiques du paradigme fouille de motifs fréquents et règles d'association dans l'analyse de données biologiques. Il s'agit d'identifier des motifs pertinents pouvant prétendre à une interprétation dans un cadre biologique. Ces applications incluent l'annotation de données biologiques, dans le but de prédire leurs fonctionnalités, la découverte de motifs structuraux tels que les séquences ou les sous graphes, l'alignement de séquences.

2.3.3 Fouille des données pour la détection des profils communautaires

Avec l'arrivée du Web 2.0, on assiste à un foisonnement de services de réseautage social, qui mettent l'utilisateur au centre des préoccupations. Ces services permettent de partager des ressources (*YouTube, Flickr, Del.icio.us*), d'échanger des informations et de construire des relations personnelles ou professionnelles *(Facebook, LinkedIn*) ou encore de diffuser des news (*Twitter, blogs*). Les utilisateurs disposent ainsi de plusieurs espaces d'informations sur différents réseaux sociaux. Ces espaces permettent souvent l'accès à des informations complémentaires sur l'identité de l'utilisateur, ses relations avec les autres et les ressources qu'ils partagent. Ainsi, pour établir le lien entre les différents utilisateurs, [Cortis et al., 2012],[Raad et al., 2010]ont proposé le calcul d'une similarité sémantique entre les attributs des profils ; et quant à [Bartunov et al., 2012],[Buccafurri et al., 2012],[Jain et

al.,2013];[Narayanan et Shmatikov, 2009], ont étudié les propriétés topologiques des réseaux pour déterminer des liens entre les différents utilisateurs.

2.3.4 Fouille des données pour le suivi de trajectoires d'objets mobiles

La fouille de données appliquées au suivi de trajectoires d'objets mobiles a pour objectif, l'Aménagement du plan de circulation des véhicules dans les grandes agglomérations. La collecte de grandes quantités de données spatio-temporelles permet ainsi d'entrevoir de nouvelles applications pour le suivi de trajectoires. A titre d'illustration, le projet GeoPKDD [Giannotti et Pedreschi, 2008] a étudié l'aménagement du plan de circulation de grandes agglomérations en fonction des déplacements des véhicules. L'analyse d'objets en mouvement à également comme domaines d'applications la géographie socio-économique, le sport (le déplacement des supporters de football au stade) [Giannotti et al., 2011], l'analyse et le contrôle de la pêche, les prévisions météorologiques et l'analyse du mouvement (suivi de migration d'aigles [Li et al., 2011].Une autre approche est le suivi de l'ensemble des trajectoires d'un même objet par une seule séquence composée de segments; l'étude menée est le suivi de trajectoires d'objets mobiles pour l'extraction de motifs périodiques (Les orages). Dans le cas présent, la particularité est de suivre approximativement la même route à intervalles de temps réguliers (fréquence des pluies saisonnières) au début de l'été [Mamoulis et al., 2004], [Cao et al., 2005, 2007].

2.3.5 Applications du datamining dans les autres domaines d'activités

Dans la littérature, des travaux ont été effectués en intelligence artificielle et plus particulièrement, sur les techniques d'apprentissage issues du datamining [David M. Fram, 2008], [Martin Ester et al. 2004] et [Ruggieri et al. 2010].

Les domaines de recherches sont nombreux et variés :

- **Médecine** : Apport d'une méthode de fouille de données pour la détection des cas de cancer du sein dans les données du Programme de Médicalisation des Systèmes d'Information : une analyse formelle des concepts sur les 2001 données de PMSI et du registre du cancer de l'Isère Christophe Goetz
- **Finance, assurance** : crédit, prédiction du marché, détection de fraudes
- **Social** : données démographiques, votes, résultats des élections,
- **Marketing et ventes** : comportement des utilisateurs, prédiction des ventes, espionnage industriel,
- **Informatique** : IHM, Réseaux, Data-Warehouse, Data Mart, Internet (moteurs intelligents, profiling*)*

L'emploi de la fouille de données dans les applications du monde réel est en soi un défi. Chaque situation est particulière avec sa collecte de données ainsi que l'application des algorithmes [Shouning et Peihua, 2005].

2.3.6 Fouille des données et activité douanière

En raison de l'accroissement des flux commerciaux et donc de leur charge de travail, les administrations douanières les plus modernes ont développé depuis plusieurs années des approches structurées pour analyser le risque, seul moyen efficace pour limiter les contrôles intrusifs, répondre aux besoins des opérateurs privés et sécuriser leurs opérations. La prolifération des données issues de ces opérations s'est, en effet accompagnée de besoins croissants d'outils automatiques pour accéder à l'information pertinente par les administrations douanières en vue de détecter l'origine des fraudes au niveau des contrôles.

Ces données ont fourni un nouveau cadre applicatif à l'extraction des informations pour limiter les contrôles intrusifs. Ainsi, les administrations douanières s'appuient sur des systèmes d'analyses du risque qui reposent notamment sur une exploitation exhaustive et

systématique de l'information statistique disponible, comme seul moyen efficace pour, à la fois, faciliter les échanges et sécuriser leurs opérations [Harrison, 2007]. *Walsh*[Walsh, 2003] *et Widowson* [Widdowson, 2005] estiment que cette démarche devrait également être adoptée, pour les mêmes raisons, dans les administrations douanières des pays en voie développement qui doivent en plus faire face aux pressions de leurs autorités pour sécuriser les recettes.[Harrison, 2007]. *Walsh*[Walsh, 2003] *et Widowson* [Widdowson, 2005] s'inscrivent dans une démarche d'analyse de risque et la sélectivité des déclarations. Cependant, les critères de sélectivité sont déterminés par un « *Comité de sélectivité* » chargé de les identifier, les valider, et les paramétrer dans un module de sélectivité du système informatique. En conséquence, ces mesures de sécurité restent très dépendantes de l'appréciation humaine.

Geourjon et Laporte [Geourjon et Laporte, 2005] ont eux mis l'emphase sur l'apport d'une approche économétrique à partir d'une expérience menée dans cinq pays d'Afrique occidentale dans le cadre de la modernisation de leurs administrations douanières. A cet effet un système « expert » d'analyses de gestion des risques a été développé sur la base d'une exploitation systématique de l'information statistique. Cette approche économétrique a été empruntée de l'analyse du risque appliquée aux banques et assurances. Pour eux, la solution la plus propice pour le ciblage des contrôles dans les pays en développement doit donc avoir pour objectif d'éliminer au maximum l'intervention des agents de l'administration douanière pour limiter l'aléa moral. La sélectivité doit reposer sur une analyse de risques réalisée à partir des informations recueillies sur les fraudes constatées (*fraude avérée*), et non sur d'éventuels soupçons de fraude, et l'exploitation de ces informations est faite en utilisant des techniques d'analyses de données et d'économétrie. Ces régularités statistiques vont permettre d'établir des profils de risques par critère ; l'analyse statistique permet ainsi d'établir une échelle de risques « *quantitative* », contrairement à la méthode de sélectivité

traditionnelle. L'objectif est d'attribuer un score « global » à chaque nouvelle déclaration, obtenu en combinant les scores « individuels » des différents critères. Ce score « global » doit refléter au mieux la probabilité de fraude de la nouvelle déclaration. L'orientation vers l'un des circuits de dédouanement se fait en fonction de ce score et de seuils préalablement déterminés par l'analyse statistique [Laporte, 2011], [Geourjon et al., 2012].

Une autre méthode, comme la technique de scoring, utilisée dans de nombreux secteurs, a aussi été appliquée avec succès. En effet, une étude réalisée par *Grigoriou* [Grigoriou, 2011] vient en effet de mettre en évidence les avantages de celle-ci pour organiser les contrôles visant à s'assurer du respect des normes techniques, sanitaires et phytosanitaires par les administrations des douanes des pays en développement.

Tout au long de cette analyse bibliographique, la notion de risque dans le milieu des activités douanières est réalisée à partir des informations recueillies sur les fraudes constatées. *L'approche de méthode de fouille de données n'a pas été empruntée conformément à la procédure d'extraction de connaissances des données (ECD).*

2.4 Méthodes de Fouille de Données Structurées

Les méthodes de Fouille de Données Structurées peuvent être classées en trois catégories :

2.4.1 La fouille des graphes

Les données sont représentées sous la forme de graphes. Chaque objet est un graphe conceptuel. Les sommets du graphe représentent les entités ainsi que leurs attributs. Les relations entre ces entités correspondent aux arcs (orientés ou non orientés). Ce formalisme est plutôt adapté aux méthodes d'analyses descriptives telles que le clustering de graphes et la recherche des motifs fréquents [Cook et Holder, 2000], [Washio et Motoda, 2003].

2.4.2 Programmation logique inductive

La base de données est constituée d'une collection de faits exprimés sous la forme de prédicats logiques de premier ordre. Ces prédicats décrivent les entités ainsi que les relations entre elles [Džeroski, 1996],[Lavrač et Džeroski, 1994]

2.4.3 La fouille de données Multi-tables

Les données sont contenues dans une base de données relationnelle. Les tables représentent les entités. Les relations entre ces entités sont définies grâce aux contraintes de clés étrangères [Džeroski, 2003].

2.5 Algorithmes d'extraction de connaissances

Durant les deux dernières décennies, un travail considérable a été consacré aux aspects algorithmiques de la fouille de motifs fréquents, ce qui a donné naissance à un nombre phénoménal d'algorithmes et d'implémentations associées. Ainsi, nous avons groupé ces différents algorithmiques en trois grandes approches :

- L'approche par segmentation
- L'approche par classification
- L'approche naïve

2.5.1 L'approche par segmentation

La segmentation est une technique permettant de créer des regroupements de données qui se ressemblent dans le groupe d'appartenance et qui se dissocient des autres regroupements. Cette technique est souvent utilisée comme méthode d'exploration ou comme prétraitement des données.

Il existe différentes méthodes pour la segmentation en fouille de données. Ces méthodes peuvent utiliser différents types de données (*données primaires ou secondaires*) et une

multitude de format (*discret, continue, numérique, textuel*). Wedel et Kamakura [Wedel and Kamakura, 1998] ont classé les méthodes de segmentation en quatre groupes, ainsi que[Green, 1977] et [Wind, 1978], comme représenté dans le tableau 2.1.

Tableau 2.1: Classification des méthodes de la segmentation

	Apriori	Post hoc
Descriptive	Tableau de contingence Modèle log-linéaire	Méthode clustering
Prédictive	Tableau croisé Régression	AID, CART, ANN, Modèle de mélange

Dans les sections qui suivent, nous détaillons les éléments essentiels de cette classification.

2.5.1.1 L'approche Apriori

Selon une étude [Pearson, 1904] de marketing de la vente d'un produit ;le marché est segmenté en fonction des critères préexistants tels que : *l'âge, le sexe ou le statut économique et social.* Les méthodes de cette approche ont alors pour objectifs de découvrir ou de décrire les caractéristiques des clients dans les segments déjà connus. Les méthodes sont le modèle log-linéaire, tableau de contingence [Pearson, 1904], les tableaux croisés et la régression sont de type « Apriori ». Cette approche est très efficace dans le cas où les variables d'analyse sont bien définies. Cependant, il est parfois difficile d'identifier les variables à utiliser. Les méthodes « Apriori » sont donc utilisées soit dans des cas particuliers, soit lorsqu'on dispose apriori d'une certaine connaissance du domaine, soit par combinaison avec des méthodes « post-hoc » (hybrid segmentation). En référence, l'auteur Umesh [Umesh, 1987] a appliqué le modèle de régression pour expliquer les préférences de transport des consommateurs pour se rendre dans un centre d'achats.

2.5.1.2 L'approche Post hoc

Les méthodes de segmentation sont les plus utilisées dans l'approche post-hoc car elles sont faciles à mettre en œuvre et peuvent s'appliquer à divers types de données. La plupart des techniques de segmentation se divisent dans une des deux catégories principales suivantes : les méthodes hiérarchiques et les méthodes par partitionnement. L'idée principale de ces méthodes est basée sur une mesure de similarité (ou distance) pour grouper les données. Le but de la segmentation est de maximiser la similarité intra-classes et de minimiser la similarité inter-classes. La distance entre les objets peut être calculée par une des méthodes ci-dessous, dépendamment de types des données :

- Distance euclidienne de deux points en *n-dimension* méthodes de segmentation) :

$$d(x_i, x_j) = (\sum_{k=1}^{n} (x_{ik} - x_{jk})^2)^{1/2}$$

- Distance de Minkowsky : (l'extension de la distance euclidienne) :

$$d_p(x_i, x_j) = (\sum_{k=1}^{n} |x_{ik} - x_{jk}|^p)^{1/p}$$

- Distance de Manhattan :

$$d_M(x_i, x_j) = \sum_{1}^{n} |x_i - x_j|$$

La distance utilisée dépendra du type de données disponibles, de son format et de sa précision de la base de données. De plus ; Il faudra faire attention lors du calcul de ces distances à la normalisation des attributs, puisque les intervalles de variance des attributs peuvent être très différents, ce qui peut entraîner la dominance d'un ou de quelques attributs sur le résultat. Il est conseillé donc, de normaliser tous les attributs sur le même intervalle puis calculer la distance.

Le problème qui se pose lors du calcul de la distance entre les attributs catégoriels, c'est qu'on ne dispose pas d'une mesure de différence. La seule mesure qui existe, en l'absence de toute information sur la signification des valeurs, est l'égalité ou l'inégalité. La distance utilisée est alors :

$$d_c(x_i, x_j) = \frac{1}{n_c} \sum_{k=1}^{n_c} f(x_{ik}, x_{jk})$$

$$f(x_{ik}, x_{jk}) = \begin{cases} 1 \text{ si } x_{ik} = x_{jk} \\ 0 \text{ sinon} \end{cases}$$

Il faut enfin la normaliser avec les attributs numériques et le nombre d'attributs catégoriels. La distance entre deux données x_i et x_j composées d'attributs numériques et catégoriels, est donc :

$$d(x_i, x_j) = d_n(x_i, x_j) + d_c(x_i, x_j)$$

En se basant sur la distance entre deux attributs, plusieurs distances peuvent être calculées :

- Distance entre deux clusters : permet de mesurer la distance entre deux clusters pour une éventuelle fusion au cas où ils soient trop proches. Cette distance peut être prise entres les centres des deux clusters, entre les deux données les plus éloignées (ou plus proches) des deux clusters ou la distance moyenne de leurs données.

- Distance intra-cluster : c'est la distance moyenne entre les données à l'intérieur d'un cluster, elle peut être utile pour maintenir un seuil d'éloignement maximum dans le cluster au-dessus duquel on doit scinder ce cluster.

- Distance inter-cluster : c'est la distance moyenne entre les clusters, elle permet de mesurer l'éloignement moyen entre les différents clusters.

- Distance intra-clusters moyenne : permet avec la distance inter-clusters de mesurer la qualité du clustering.

La mesure de similarité peut être utilisée par un algorithme de segmentation pour trouver le partitionnement optimal des données. Parmi ces algorithmes on peut citer : les algorithmes de segmentations hiérarchiques et les algorithmes de segmentations par partitionnement qui sont traités dans les sections qui suivent.

- Les algorithmes de segmentations hiérarchiques

Les algorithmes de segmentation hiérarchiques sont récursifs et peuvent être soit de type par agglomération soit de type par division. Les méthodes par agglomération partent d'une partition où chaque donnée représente un segment et à chaque itération les deux segments les plus proches sont fusionnés jusqu'à ce que tous les points se trouvent dans un seul grand segment. Les méthodes par division créent une hiérarchie descendante en procédant par divisions successives. La distance entre deux segments peut être donnée par la distance entre les deux membres les plus proches, les plus éloignés ou entre leur centroïdes.

Les méthodes hiérarchiques retournent le résultat sous la forme d'un dendrogramme, qui non seulement indique les objets et les segments à chaque itération, mais aussi la valeur du critère choisi pour chaque partition rencontrée. Cela permet de déterminer le nombre de segment à utiliser dépendamment de l'objectif poursuivi. Cependant, ces méthodes sont difficilement applicables avec de grandes masses de données. Dans la segmentation du marché, [Greeno et al., 1973] ont utilisé la méthode hiérarchique pour identifier les segments en se basant sur les caractéristiques personnelles et comportementales des consommateurs.

- Les algorithmes de segmentation par partitionnement

Dans ces méthodes, le nombre de segments voulus (k) doit être connu, il peut être déterminé par des experts ou par itérations successives (en croisant différentes méthodes de segmentation).

Ces méthodes produisent les segments autour de noyaux choisis initialement, puis elles améliorent itérativement ces segments initiaux en se basant sur une fonction de coût à minimiser. Elles restent efficaces avec de grandes masses de données et des types des données variées : données énumératives, numériques et textuelles. Néanmoins, ces méthodes rencontrent deux problèmes majeurs :

- Détermination difficile du nombre de segments optimal (k) et le résultat dépend de l'initialisation (les noyaux de départ).
- Convergence de ces méthodes vers un optimum local.

Dans leur revue de littérature, [Punj et Stewart, 1983] ont conclu que les méthodes par partitionnement sont plus efficaces que les méthodes hiérarchiques dans la segmentation du marché. La méthode des K-means [Mc Queen, 1967] est la plus connue et la plus couramment utilisée dans la pratique. D'autres méthodes peuvent aussi être utilisées tels que : K medoids, CLARANS, EM-clustering [Jin and Han, 2017].

Des extensions pour les méthodes par partitionnement sont proposées. DeSarbo et al. ont proposé une méthode en considérant la pondération des variables [DeSarbo et al., 1984]. Les points indiquent l'importance de chaque variable. Cependant, cette méthode est difficile à implémenter. Dolnicar et Leisch, 2004 ont intégré la méthode hiérarchique à la méthode *k-means* (bagged clustering) pour améliorer la stabilité et l'interopérabilité des segments concernant les données binaires [Dolnicar et Leisch, 2004]. D'autres améliorations concernent le choix du point de départ et/ou la détermination du nombre de cluster. [Maulika et Bandyopadhyay, 2000] ont appliqué l'algorithme génétique pour choisir les noyaux au

départ. Il existe beaucoup de cas d'applications de ces méthodes en segmentation du marché. Nous citons ici trois exemples d'application de ces méthodes. En utilisant la méthode bagged clustering, *Dolnicar et Leisch* ont identifié cinq classes de touristes différents dans le marché du tourisme en Australie [Dolnicar and Leisch, 2004]. *H.Sung* a utilisé la méthode de k-moyenne pour identifier six groupes des voyageurs d'aventure aux États-Unis afin de déterminer les marchés potentiels [H.Sung, 2004]. Enfin, *Agard, Morency et Trépanier* ont appliqué la méthode k-moyenne couplée à un algorithme hiérarchique pour déterminer les comportements des clients en transport publique [Agard et al., 2006].

2.5.2 L'approche par classification

La classification supervisée est l'une des techniques les plus utilisées dans l'analyse des bases de données. Le but des classes est de trouver des dénominateurs communs au niveau des individus pour permettre de trouver à quelle classe appartient un nouvel individu arrivant [9]. Elle permet d'apprendre des modèles de décision qui permettent de prédire le comportement des exemples futurs. La classification supervisée consiste à inférer à partir d'un échantillon d'exemples classés.

Un système d'apprentissage effectue la recherche d'une telle procédure selon un modèle. Les systèmes d'apprentissage peuvent être basés sur des hypothèses probabilistes (classifieur naïf de Bayes, méthodes paramétriques) ; sur des notions de proximité (plus proches voisins) ; sur des recherches dans des espaces d'hypothèses (arbres de décision, réseaux de neurones). Les algorithmes de classification servent à regrouper des individus de mêmes propriétés. Le processus de classification se fait en deux étapes : Une étape d'apprentissage et une étape de classification.

- Une étape d'apprentissage (ou entrainement),

Durant cette étape, un classifier (*Une fonction, un ensemble de règles, ...*) est construit en analysant (ou en apprenant de) une masse de données d'exemples d'entrainement avec leurs classes respectives.

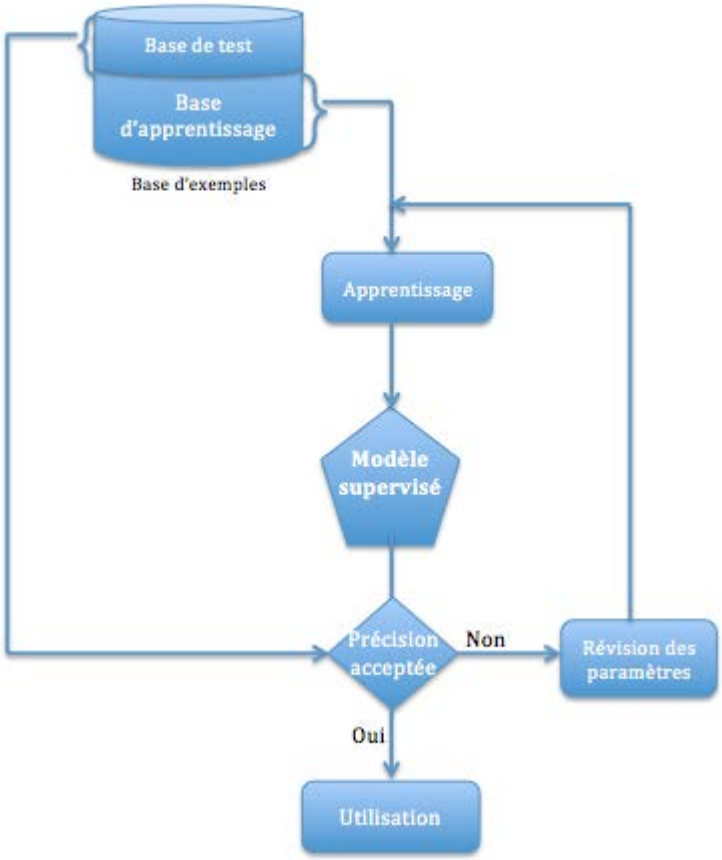

Figure 2.1: Processus de classification

Un exemple d'entrainement $X = (x_{1,}, x_{2,} \dots, x_{m,})$ est représenté par un vecteur d'attributs de dimension *m*. Chaque exemple est supposé appartenir à une classe prédéfinie représentée dans un attribut particulier de la base de données appelé attribut de classe. Puisque la classe de chaque exemple est donnée, cette étape est aussi connue par l'apprentissage supervisé.

- L'étape de classification (ou utilisation).

Dans l'étape de classification, le modèle construit dans la première étape est utilisé pour classer les nouvelles données. Mais avant de passer à l'utilisation, le modèle doit être testé pour s'assurer de sa capacité de généralisation sur les données non utilisées dans la phase d'entrainement. Le modèle obtenu peut être testé sur les données d'entrainement elles-mêmes, la précision (le taux de reconnaissance) est généralement élevée mais ne garantit pas automatiquement une bonne précision sur les nouvelles données. En effet, les données d'entrainement peuvent contenir des données bruitées ou erronées (outliers) qui ne représentent pas le cas général et qui tire le modèle vers leurs caractéristiques. Ce cas est appelé le sur-apprentissage ou en anglais ''*over fitting*'' et qui peut être évité en testant le modèle sur une base de données différentes appelée base de test. La base de test est un ensemble d'exemples ayant les mêmes caractéristiques que ceux de la base d'entrainement et qui sont écartés au départ de l'entrainement pour effectuer les tests. La méthode de prédiction utilisée dépend essentiellement du type d'information prédite c'est à dire le type de l'attribut de classe. Si l'attribut est catégoriel ou symbolique (appartient à un ensemble fini), il s'agit de classification. En revanche si cet attribut est continu (numérique) il s'agit d'un problème de régression.

Il existe de nombreux algorithmes de classification. Le plus connu est le C4.5 [Kantardzic, 2003]. Il s'agit d'un algorithme qui donne un arbre de décision comme modèle final. Cet algorithme est facile de compréhension et donne d'excellents résultats, surtout avec des attributs de type nominal. D'autres types de classification existent, comme *les k plus proches voisins, les réseaux de neurones*, *ou la classification bayésienne* que nous présentons dans les lignes du mémoire qui suivent.

2.5.2.1 k plus proches voisins

L'algorithme des k-plus proches voisins est un des algorithmes de classification les plus simples. Le seul outil dont on a besoin est une distance entre les éléments que l'on veut classifier. Si on représente ces éléments par des vecteurs de coordonnées, il y a en général plusieurs choix possibles pour ces distances, partant de la simple distance usuelle (euclidienne) jusqu'à des mesures plus sophistiquées pour tenir compte si nécessaire de paramètres non numériques comme la couleur, la nationalité.

Son fonctionnement est le suivant : on considère que l'on dispose d'une base d'éléments dont on connaît la classe. On parle de base d'apprentissage, bien que cela soit de l'apprentissage simplifié. Dès que l'on reçoit un nouvel élément que l'on souhaite classifier, on calcule sa distance à tous les éléments de la base. Si cette base comporte 100 éléments, alors on calcule 100 distances et on obtient donc 100 nombres réels. Si $k = 25$ par exemple, on cherche alors les 25 plus petits nombres parmi ces 100 nombres (Voir l'exemple à la figure 2.2).

Ces 25 nombres correspondent donc aux 25 éléments de la base qui sont les plus proches de l'élément que l'on souhaite classifier. On décide d'attribuer à l'élément à classifier la classe majoritaire parmi ces 25 éléments. Bien sûr, on peut faire varier k selon ce que l'on veut faire, on peut aussi complexifier la méthode en considérant que les votes des voisins ne sont pas de même poids, etc. Mais l'idée reste la même.

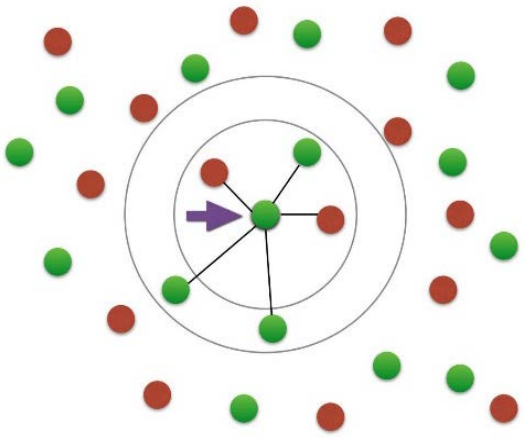

Figure 2.2: Schéma de construction d'un modèle de k=5 plus proches voisins

2.5.2.2 Réseaux de neurones

Les Réseaux de Neurones Artificiels (RNA) sont inspirés de la méthode fonctionnement du cerveau humain qui est totalement différente de celle d'un ordinateur. Le cerveau humain se base sur un système de traitement d'information parallèle et non linéaire, très compliqué, ce qui lui permet d'organiser ses composants pour traiter, d'une façon très performante et très rapide, des problèmes très compliqués tels que la reconnaissance des formes. Un réseau de neurones est une structure de réseau constituée d'un nombre de nœuds interconnectés par des liaisons directionnelles, chaque nœud représente une unité de traitement et les liaisons représentent les relations causales entre les nœuds. La figure 2.2 suivante représente une schématisation d'un neurone.

La figure 2.2 schématise un neurone k ainsi que ses éléments constitutifs de base :

- Un ensemble de connexions avec les différentes entrées x_i, pondérée chacune par un poids w_{ki},
- Un additionneur permettant de calculer une combinaison linéaire des entrées xi pondérées par les coefficients w_{ki},
- Un biais b_k qui permet de contrôler l'entrée de la fonction d'activation,
- Une fonction d'activation f permettant de délimiter la sortie y_i du neurone.
- Un ensemble de connexions avec les différentes entrées x_i, pondérer chacune par un poids w_{ki},
- Un additionneur permettant de calculer une combinaison linéaire des entrées x_i pondérées par les coefficients w_{ki},
- Un biais b_k qui permet de contrôler l'entrée de la fonction d'activation,
- Une fonction d'activation f permettant de délimiter la sortie y_i du neurone.

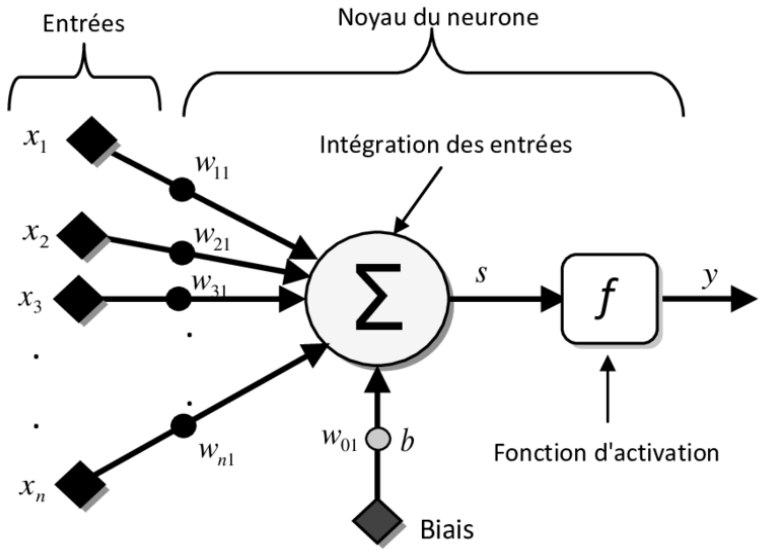

Figure 2.3: Modèle d'un neurone artificiel

Mathématiquement, la sortie y_k du neurone peut être exprimée par la fonction suivante :

$$y_k = f(w_{k1}x_1 + w_{k2}x_2 + \cdots + w_{kn}x_n + b_k)$$

L'architecture d'un réseau de neurones artificiels est définie par la structure de ses neurones et leur connectivité. Elle est spécifiée par le nombre d'entrées, de sorties, de nœuds et la façon selon laquelle sont interconnectés et organisés les nœuds. Une fameuse architecture des réseaux de neurones est celle basée sur des couches où les nœuds de chaque couche n'ont aucune connexion entre eux.

Cette architecture est utilisée dans presque 90 % des applications commerciales et industrielles. La figure suivante représente un réseau de neurones de quatre couches.

Les couches 1 et 2 s'appellent des couches cachées tandis que la couche 3 est la couche de sortie. La tâche principale des réseaux de neurones artificiels est l'apprentissage pour la classification, qui est réalisée par un processus itératif d'adaptation des poids w_i pour arriver à la meilleure fonction permettant d'avoir$f(x_i) = y_i, \forall i = 1 \ldots N$. Les valeurs des w_i sont initialisées aléatoirement, et corrigées selon les erreurs entre les y_i obtenus et attendus

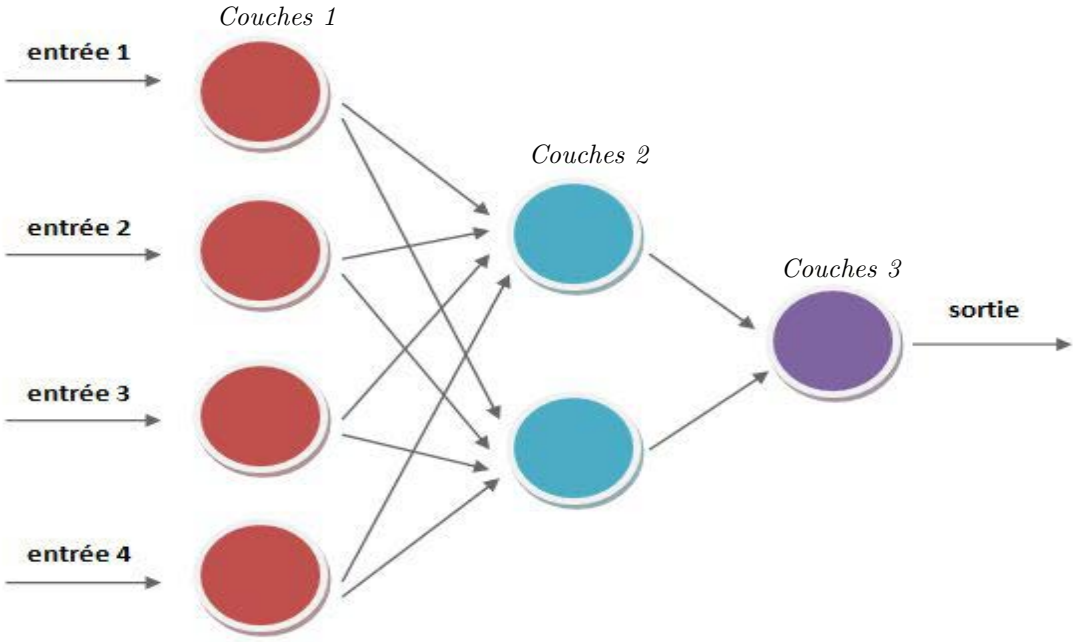

Figure 2.4 : Architecture d'un réseau de neurones artificiel

Dans un réseau de neurones multicouches, la correction se fait dans le sens inverse du sens de propagation des données ce qui est appelé la "*back propagation*". À chaque présentation d'un exemple d'apprentissage au réseau, on passe par deux étapes :

1. Dans l'étape de propagation, les valeurs du vecteur d'entrée (l'exemple) sont reçues dans la couche d'entrée et propagées d'une couche à l'autre jusqu'à la sortie où un vecteur de sortie (les y_i) est obtenu.

2. Dans la phase de *back propagation*, les w_i sont ajustés de la dernière couche jusqu'à la première de manière à rapprocher les y_i obtenus de ceux attendus

Ces deux étapes sont répétées avec chaque exemple d'apprentissage pour obtenir à la fin un réseau de neurones artificiels entrainés.

2.5.2.3 Classification bayésienne

Les techniques se basant sur les lois statistiques sont les premières qui ont été utilisées pour l'analyse de données. Elles consistent à prendre un sous ensemble d'une population et essayer d'arriver à des conclusions concernant toute la population. Ce sont des méthodes qui reposent sur la théorie de Bayes représentant une référence théorique pour les approches statistiques de résolution des problèmes de classification. Le principe de cette théorie est le suivant : Soit X un échantillon de données dont la classe est inconnue et qu'on veut la déterminer, et soit H une hypothèse (X appartient à la classe C par exemple). On cherche à déterminer $P(H/X)$ la probabilité de vérification de H après l'observation de X. $P(H/X)$ est la probabilité postérieure c'est-à-dire après la connaissance de X tandis que $P(H)$ est la probabilité à priori représentant la probabilité de vérification de H pour n'importe quel exemple de données. Le théorème de Bayes propose une méthode de calcul de $P(H/X)$ en utilisant les probabilités $P(H)$, $P(X)$et$P(X/H)$:

$$P(H/X) = [\ P(H/X).\ P(H)]/\ P(X)$$

$P(H/X)$ est donc la probabilité d'appartenance de X à la classe C, $P(H)$ la probabilité d'apparition de la classe C dans la population et qui peut être calculée comme le rapport entre le nombre d'échantillons appartenant à la classe C et le nombre total d'échantillons. $P(X/H)$ peut être considérée comme la probabilité d'apparence de chaque valeur des attributs de X dans les attributs des échantillons appartenant à la classe C :

$$P(H/X) = \prod P(\alpha_i = v_i/H)$$

Où α_i est le $i^{\text{ème}}$ attribut de X et v_i sa valeur. Cette astuce de calcul de $P(X/H)$ est basée sur la supposition d'indépendance entre les attributs. Bien que cette supposition soit rarement vérifiée, sa considération facilite le calcul et donne une idée approximative sur la

probabilité. Finalement *P(X)* est constante pour toute la population et indépendante des classes. Il ne reste donc que considérer la classe de *X*, celle maximisant le produit *P(X/H). P(H)*. Cette application est l'application la plus simple de la théorie de Bayes, elle s'appelle la classification naïve de Bayes.

La méthode naïve de Bayes est applicable uniquement en cas de vérification de l'indépendance entre les attributs, ce qui peut être contrôlé par la matrice de corrélation et ses valeurs propres. Aussi, les valeurs des attributs numériques doivent avoir une distribution normale.

Cette méthode reste une méthode simple et moins coûteuse en temps de calcul. Elle est aussi incrémentale c'est-à-dire que l'arrivée d'une nouvelle information (classe d'un nouvel enregistrement) ne nécessite pas de refaire tous les calculs pour la prendre en considération. Les connaissances apprises peuvent être renforcées sans avoir besoin de refaire tous les calculs.

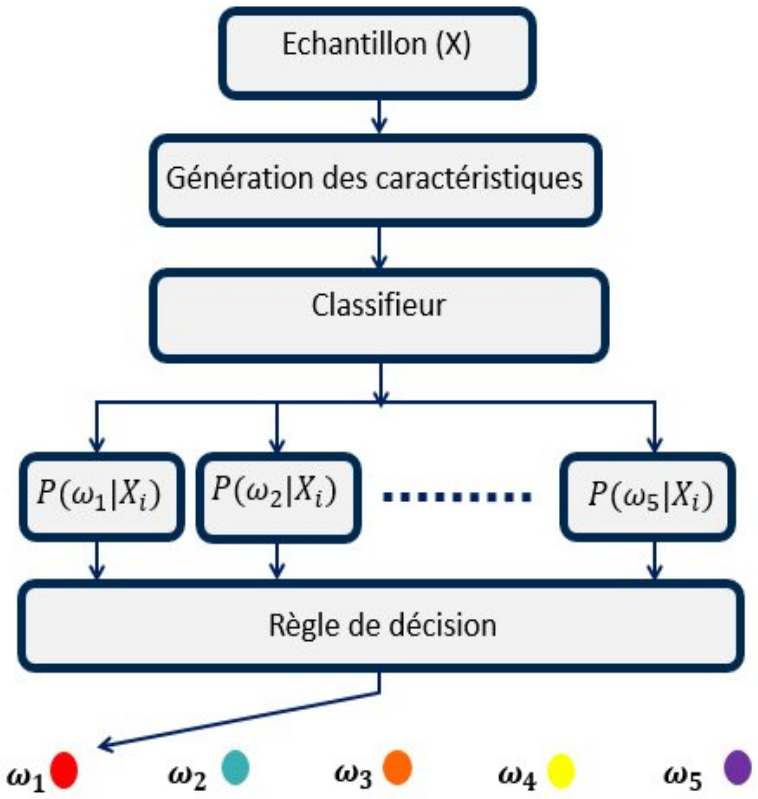

Figure 2.5 : Schéma de la classification bayésienne

2.5.3 L'Approche Naïve

Une Approche Naïve (AN) pour l'extraction des motifs fréquents consiste à parcourir l'ensemble de tous les motifs, à calculer leur nombre d'occurrences (support) et à ne garder que les plus fréquents. Cette approche proposée par Agrawal et ses co-auteurs en 1994 [Agrawal and Srikant, 1994], est un algorithme de base qui permet d'extraire des motifs fréquents dans une base ayant plusieurs milliers d'attributs et plusieurs millions d'enregistrements. L'idée repose sur une extraction par niveaux selon le principe suivant :

- Tout sous-motif d'un motif fréquent est fréquent
- Tout sur-motif d'un motif non fréquent est non fréquent.

Les algorithmes de référence basés sur cette approche sont nommés algorithmes d'extraction des itemsets fréquents (ou encore motifs fréquents).

2.5.3.1 Algorithme d'extraction des motifs fréquents

Les algorithmes d'extraction de motifs fréquents procèdent de manières itératives, en parcourant le treillis des motifs fréquents (itemsets fréquents) « par niveaux » en largeur, c'est-à-dire du bas vers le haut. Ils déterminent à chaque itération tous les motifs fréquents du niveau, une taille donnée. Pour chaque itération k, un ensemble de *k-itemsets* candidats *(k-itemsets fréquents potentiels*) est généré et les supports de ces candidats sont calculés lors d'un seul et même balayage, ce qui permet de limiter le nombre total de balayages réalisés. Les premiers algorithmes d'extraction des itemsets fréquents par niveaux sont intervenus dès 1993 pour étudier le comportement des consommateurs dans une base de transactions d'un supermarché [Agrawal et al., 1993], [Houtsma and Swami] en trouvant des corrélations entre différents produits consommable. Par la suite plusieurs autres algorithmes basés sur des propriétés et des méthodes identiques ont été proposés dans l'idée d'amélioration de l'efficacité de ceux-ci. On peut notamment citer, l'algorithme DHP (*Direct Hashing and*

Pruning) proposé par *Park et al.* [Park et al., 1995] qui utilise des tables de hachage afin de diminuer le nombre de candidats générés, l'utilisation de « bitmaps » hiérarchiques proposée par *Gardarin et al.* [*Gardarin et al.*, 1998] qui permet de diminuer le coût des opérations sur les ensembles d'itemsets, la parallélisation du calcul étudiée par *Zaki* [Zak, 1999] et l'utilisation de heuristiques non-déterministes proposée par *Gunopulos et al.* [Gunopulos et al., 1997]. D'autres auteurs ont plutôt mis l'emphase sur de nouvelles techniques pour extraire des motifs fréquents, c'est le cas des algorithmes AprioriTid[Agrawal and Srikant, 1994], Partition [Savasere et al., 1995], Sampling [Toivonen, 1996] et DIC (*Dynamic Itemset Counting*) [Brinet al., 1997].

2.5.3.2 Optimisation de l'algorithme Apriori

Les algorithmes basés sur le principe d'Apriori souffrent de la gestion du nombre de candidats qu'ils peuvent générer, surtout pour des valeurs de support relativement faibles. De récents travaux ont proposé une série d'algorithmes qui introduisent plusieurs optimisations et structures de données pour améliorer les performances du processus d'extraction de motifs fréquents. Ces algorithmes sont centrés essentiellement sur la réduction de la taille de l'espace de recherche dans le but de le stocker en mémoire et de réaliser moins d'entrées/sorties. Ils sont aussi focalisés sur la minimisation du coût de l'étape de calcul du support. Parmi ceux-ci, nous pouvons citer les méthodes proposées par [Park et al., 1995], [Brin et al., 1997], Zaki (1998); [Gardarin et al., 1998], [Bastide et al., 2000]; [Han et al., 2000], [Bykowski et Rigotti, 2001], [Bastide et al., 2002], [Calders et Goethals, 2002], [Boulicaut et al., 2003], [Geerts et al., 2005]. Dans l'objectif de limiter le nombre de candidats générés, d'autres travaux introduisent des représentations condensées pour l'extraction d'ensembles de motifs condensés dont la cardinalité est plus réduite, mais avec le même niveau de pertinence que l'ensemble de tous les motifs fréquents

[MannilaetToivonen, 1996]. Parmi ces représentations condensées, on peut citer les représentations closes [Pasquier et al., 1999],[Stumme et al., 2000],[Pei et al., 2000],[Zaki &Hsiao, 2002],[Uno et al., 2003],[Uno et al., 2005], les représentations par motifs maximaux [Zaki et al., 1997],[Lin et Kedem, 1998], [Lin et Kedem 1998],[Burdick et al., 2005], les représentations par motifs non-dérivables[Calders et Goethals, 2002],[Calders et Goethals, 2007], et les représentations par ensembles libres [Boulicaut et al., 2003].

Les approches citées ci-dessus sont marquées par leur effort algorithmique pour la réduction du temps de calcul de l'étape d'extraction des motifs intéressants. Ce succès obtenu est essentiellement dû à des prouesses de programmation avec la conjonction de la manipulation de structures de données compactes en mémoire centrale. Ils existent d'autres approches, permettant une réduction sans perte d'information, reposent sur un ensemble de résultats issus de la théorie de l'Analyse Formelle de Concepts (AFC) introduite par [Wille, 2009]. Le principe de ces approches est tout d'abord de déterminer l'ensemble minimal de règles d'association présentées l'utilisateur, tout en maximisant la quantité d'informations utiles véhiculées ; puis de disposer d'un mécanisme d'inférence qui, suite à la demande de l'utilisateur, permet de retrouver le reste des règles d'associations tout en déterminant avec exactitude leur support et leur confiance sans accéder à la base de données [Pasquier, 2000], [Gasmi et al., 2006]. Dans ce contexte, de nombreux algorithmes de fouille de règles d'associations basées sur les treillis des concepts ont été proposés, comme par exemple Touch et Talky-G [Szathmary et al., 2009]. Il faut noter que l'AFC est également utilisée dans les représentations condensées closes des motifs cités ci-dessus. D'autres algorithmes proposés s'appuyant sur l'architecture des processeurs multicœurs, qui proposent des optimisations basées à la fois sur la réduction de la base de données et sur les parallélismes multithreads. À titre d'exemple, nous pouvons citer les travaux des thèses de [Negrevergne, 2011] et de [Quintero et Flores, 2013]. dans ce mémoire thèse, notre intérêt n'est pas porté sur

l'optimisation des algorithmes d'extraction de motifs fréquents basée sur la minimisation du nombre de motifs extraits et du temps de leur extraction.

Le lecteur intéressé par une description complète sur la palette des algorithmes d'extraction des motifs fréquents peut se référer à la thèse de N. Pasquier [Pasquier, 2000]. Une typologie de structures de données pour l'implantation de ces algorithmes, avec des évaluations expérimentales, se trouve dans la thèse de Y. Bastide [Bastide, 2000].

2.6 Application de la fouille des données

L'extraction de motifs intéressants a connu récemment un développement impressionnant dû à une pression accrue des propriétaires de données sous-exploitées et à la réponse des chercheurs par de nombreux résultats théoriques et pratiques. A l'origine les données analysées provenaient du domaine de la vente et les motifs intéressants se présentaient sous forme de règles d'associations.

Le Datamining est une approche d'analyse de données, adaptée et utilisée dans plusieurs domaines d'activités.

- Assurances et santé

- Découverte d'associations des demandes de remboursements
- Identification de clients potentiels de nouvelles polices d'assurances.
- Détection d'associations de comportements pour la découverte de clients à risque.
- Détection de comportements frauduleux.

- Banques / Finances

- Détection d'usages frauduleux de cartes bancaires.
- Gestion du risque lié à l'attribution de prêts bancaires par le scoring.
- Découverte de relations cachées entre les indicateurs financiers.

- Détection de règles de comportements boursiers par l'analyse des données du marché.

  - Vente, distribution / Marketing

- La gestion de la relation client (GRC ou CRM) consiste en l'ensemble des activités visant à cibler, attirer et conserver les "bons" clients.
- Détection d'associations de comportements d'achat.
- Découverte de caractéristiques de clientèle.
- Prédiction de probabilités de réponses aux campagnes de mailing.

  - Ressources Humaines

Le Datamining est également utilisé dans les ressources humaines (RH) de certains ministères pour identifier les caractéristiques de leurs employés les plus performants. L'information obtenue (comme les universités fréquentées par des employés potentiels) peut contribuer aux efforts de recrutement des ressources humaines.

Ces dernières années, l'exploration de données a été largement utilisée dans les domaines de la science et de l'ingénierie, tels que la bio-informatique, la génétique, la médecine, l'éducation et l'énergie électrique.

  - Médical / Pharmaceutique

- Diagnostic assisté par ordinateur (CAD) par l'apprentissage de systèmes experts.
- Explication ou prédiction de la réponse d'un patient à un traitement.
- Identification des thérapies à succès (combinaison de prescriptions).
- Étude des corrélations entre le dosage dans un traitement et l'apparition d'effets secondaire.

  - La génétique humaine

Dans l'étude de la génétique humaine, le Datamining permet de répondre à l'objectif important de comprendre la relation de correspondance entre l'ADN et les maladies. En

effet, il vise à savoir comment les changements dans la séquence d'ADN d'un individu affectent les risques de développer des maladies courantes afin les prévenir ou de les traiter. Le datamining peut contribuer de manière significative et avec succès à l'explication ou la prédiction de phénomènes complexes dans les domaines médical et pharmaceutique.

- Ingénierie électrique

Dans le domaine de l'ingénierie électrique, les Datamining ont été largement utilisés pour la surveillance de l'état du matériel électrique à haute tension. Le but de surveillance de l'état est d'obtenir de précieuses informations par exemple, sur l'état de l'isolation (ou d'autres importantes des paramètres de sécurité).

- Aérospatiale

Le Datamining est également intégré aux données spatiales. L'objectif final est de trouver des modèles dans les données relatives à la géographie. Jusqu'à présent, l'exploration de données et de systèmes d'information géographiques ont existé en tant que deux technologies distinctes, chacune avec ses propres méthodes. L'immense explosion de données géoréférencées occasionnée par l'évolution de l'informatique, la cartographie numérique, la télédétection et la diffusion mondiale des systèmes d'informations géographiques mettent l'accent sur l'importance de développer une analyse et une modélisation géographique plus fines.

A la figure 2.6, nous présentons les divers domaines d'application de la technique de fouille de données

Les principaux indicateurs d'évaluations introduits pour l'extraction de connaissances dans un domaine sur la base des règles d'associations sont : Le support, la confiance, la densité sur la base des mots clés suivants : *Items, Itemset ou motif, transaction, base de données et motif fréquent.*

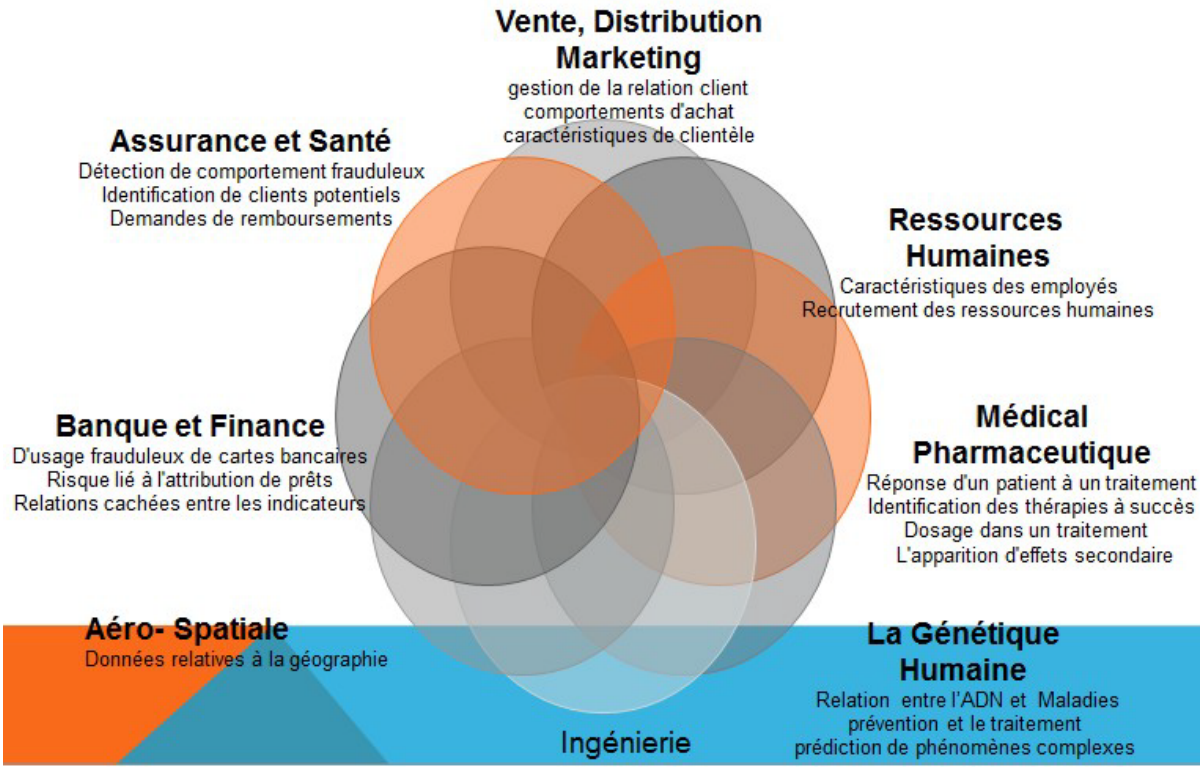

Figure 2.6 : Les différents domaines d'applications de la fouille de données

Les principaux indicateurs d'évaluation introduits pour l'extraction de connaissances dans un domaine sur la base des règles d'associations sont : Le support, la confiance, la densité sur la base des mots clés suivants : *Items, Itemset ou motif, transaction, base de données et motif fréquent.*

Définition des mots clés

Étant donné un ensemble fini non vide $A = \{a_1, a_2, \ldots, a_m\}$ de m éléments communément appelés items. Ces items peuvent désigner, comme déjà mentionné au chapitre 1, des articles commandés d'un supermarché, des pages Web visités en surfant, ou, comme dans cette thèse une opération douanière ou, une collection de variables, d'attributs, d'évènements.

- *Un motif* ou encore un itemset I est un sous-ensemble de A. Il est qualifié de k-motif lorsque son cardinal est égal à k. Dans le contexte du modèle du panier de la

ménagère, un *k*-motif symbolise donc l'ensemble des produits commandés par un client lors de ses courses.

- *Une transaction* t_i est un sous-ensemble non vide de A. Elle est identifiée par son identificateur unique *i.*
- *Un ensemble* (ou une base) de données D est un ensemble de n transactions, qu'on note comme un multiset$D = \{t_1, t_2, \ldots, t_n\}$.

2.6.1 Chaîne d'extraction des connaissances sur les données

Dans la communauté Extraction de Connaissances, le processus décrit par [Fayyad et al., 1996] est couramment référencé. Nous le généralisons ici à l'analyse de données. Ce processus a pour objectif de transformer des données de bas niveau sous d'autres formes plus compactes, plus abstraites ou plus utiles. Il comprend un ensemble d'étapes interactives et itératives [Fayyad et al., 1996].

Les données en entrée de ce processus sont diverses par nature : numériques, symboliques, booléennes, multi-dimensionnelles, multi-sources, etc. Elles peuvent également se distinguer par leur structure : données ensemblistes, arborescentes, séquentielles, sous la forme de graphes, etc. Elles peuvent être incomplètes, entachées d'erreurs. Elles peuvent être dynamiques, évoluer dans le temps, arriver en flots, etc.

Les attendus de ce processus peuvent prendre des formes diverses. L'utilisateur souhaitera, par exemple, faire face à un problème de classification, en associant une pathologie à un patient en fonction de ses symptômes. Dans un service hospitalier, on pourra chercher à prédire le nombre de lits occupés pour la semaine suivante afin de prévoir le personnel nécessaire. On pourra souhaiter détecter des anomalies dans le fonctionnement d'un équipement médical afin de déceler au plus vite les signes annonciateurs d'une panne.

Ce processus peut donc être très complexe et les étapes peuvent varier considérablement en fonction de la nature des données et des objectifs de l'application.

2.6.1.1 Sélection des données

Pour sélectionner les données, il faut tout d'abord déterminer les sources d'informations qui pourront être utiles. Couramment effectuée à l'aide de requêtes, cette première étape consiste à sélectionner, dans une base ou un entrepôt de données, les informations relatives au problème pour lequel on souhaite construire de nouvelles connaissances.

2.6.1.2 Pré-traitement

Les données sélectionnées sont souvent incomplètes, bruitées, de qualité hétérogène ou bien ne correspondent pas au format d'entrée des algorithmes de fouille. Elles sont donc nettoyées et formatées de façon à pouvoir appliquer une technique de fouille de données.

2.6.1.3 Analyse des données

C'est l'étape centrale du processus. L'algorithme est choisi selon le type des données et la problématique applicative. Les données sélectionnées et prétraitées sont explorées avec un ou plusieurs algorithmes. Ces algorithmes peuvent par exemple générer un ensemble de motifs, des règles ou un regroupement par classe. Même si l'étape d'analyse de données n'est qu'une partie du processus de fouille de données, elle est celle qui suscite le plus de travaux dans la littérature.

2.6.1.4 Restitution

Les informations extraites ne sont souvent pas directement interprétables. Cette phase consiste à traiter le format de sortie des algorithmes pour restituer les résultats, les rendre facilement visualisables et analysables par les utilisateurs. Une fois seulement les motifs validés, on obtient des connaissances.

2.6.2 Techniques de génération de motifs fréquents

Il existe plusieurs façons d'explorer les règles d'association, l'une de ces méthodes est la méthode naïve, on utilise alors toutes les combinaisons possibles des attributs et de leurs valeurs pour créer toutes les règles d'associations possibles. Cependant, cela devient problématique au niveau complexité computationnelle (*temporelle*) du fait de l'explosion combinatoire. En effet, le nombre de règles générées est énorme. On peut optimiser cette méthode en gardant juste les règles avec un support et une confiance minimum.

Il existe deux techniques d'extraction, *générer-et-élaguer* et *diviser-pour-régner*, présentées successivement ci-dessous.

2.6.2.1 Extraction par la technique générer-et-élaguer

Les algorithmes reposant sur cette technique parcourent en largeur l'espace de recherche par niveau et considèrent un ensemble de motifs d'une taille donnée lors de chaque itération. A chaque niveau k, un ensemble de candidats de taille k est généré et les motifs fréquents sont retenus pour en générer d'autres au niveau suivant par jointure. Les supports des motifs candidats sont calculés et les candidats qui ont le support inférieur à *MinSupp* sont élagués. Cet élagage est justifié par la propriété d'anti-monotonie du support

2.6.2.2 Extraction par la technique diviser-pour-régner

Les algorithmes reposant sur cette technique parcourent en profondeur l'espace de recherche et divisent la base de données en sous-ensembles de données, puis appliquent le processus d'extraction des motifs fermés [Pasquier et al., 1999c] récursivement sur ces sous-ensembles. Ce processus d'extraction repose sur un élagage de la base de données basé essentiellement sur une métrique statistique et des heuristiques.

- Le principe de cette technique est d'éviter l'inconvénient de la technique « générer-et-élaguer », à savoir la génération d'un nombre excessif de candidats. L'exemple principal dans cette catégorie est l'algorithme FP-Growth (Frequent-Pattern Growth) [Han et al., 2000] qui construit les motifs fréquents sans génération de candidats. Cet algorithme compresse tout d'abord les motifs fréquents représentés dans la base de données à l'aide d'une structure compacte appelée FP-Tree (Frequent-Pattern tree) dont les branches contiennent les associations possibles des items. Il fouille ensuite le FP-Tree, ce qui permet de générer tous les motifs fréquents possibles.

  -La technique « diviser-pour-régner » a été également implémentée dans d'autres algorithmes, comme par exemple Closet [Pei et al., 2000] inspire de l'algorithme FP-Growth en utilisant la même structure de données. En revanche, cet algorithme est basé sur l'approche d'extraction de motifs fermés fréquents. Plusieurs variantes de cet algorithme ont été proposées en gardant le même principe et en apportant des améliorations [Wang et al., 2003; Grahne & Zhu, 2003]. Ces variantes adoptent la même structure de données FP-tree [Han et al., 2000] qui permet de compresser la base de données et de fusionner plusieurs transactions, lorsqu'elles partagent un même item.

2.7 Environnement libre des fouilles de motifs

La fouille de motifs a conquis une attention capitale depuis son introduction. Ceci s'est clairement manifesté par la production d'un nombre phénoménal d'algorithmes où chacun prétend prendre le devant. En dépit de l'engouement observé, la communauté a enregistré, quand même, des avis divergents sur plusieurs convictions relatives aux comportements de plusieurs algorithmes, voire sur le même algorithme sous différentes configurations. Cet état

de fait a incité les chercheurs du domaine à l'organisation de rencontres scientifiques pour débattre et échanger les idées sur cette question. Les WorkShops FIMI'03 [Goethals and Zaki, 2003] et FIMI'04 [Bayardo et al., 2004] ont été organisés dans cette optique. Les objectifs visés furent d'apporter des réponses et contributions aux multiples défaillances constatées. Parmi ces dernières nous citons, en particulier, le manque d'implémentations publiques, de bases de données réelles, et d'études expérimentales sérieuses. Les organisateurs ont insisté surtout sur la nécessité d'adopter la stratégie open-source.

Dans cette section, nous explorons quelques outils et bibliothèques libres de fouille de données de l'angle de la découverte de motifs, que nous avons partiellement tirés de [Chen et al., 2007], [Naulaerts et al., 2015], [Altalhi et al., 2017] et [FournierViger et al., 2017]. Nous voulons en profiter pour, en plus, mettre en exergue certains outils généraux de fouille de données peu connus voire méconnus.

2.7.1 WEKA

WEKA (Waikato Environment for Knowledge Analysis) [Hall et al., 2009] est un environnement open source, multi-plateforme, extensible écrit en Java et développé à l'université de Waikato en New Zealand 11. Ce projet offre plusieurs algorithmes pour les tâches usuelles d'apprentissage et de fouille de données : classification, segmentation, analyse d'association, régression, et sélection d'attributs, avec une variété d'autres outils de prétraitement et de visualisation de données. Il est exploitable via une interface graphique ou en se servant de la ligne de commande. Le support de la fouille de motifs est, hélas, limité à seulement trois composants : les deux algorithmes d'extraction de motifs Apriori et FPgrowth, ainsi qu'un filtre d'association permettant différents filtrages de données avant d'attaquer un algorithme d'association. Toutefois, Weka possède plusieurs autres fonctionnalités intéressantes, ce qui lui a permis d'occuper les premiers rangs dans les études

d'évaluation et de comparaison à l'instar de celle présentée dans [Altalhi et al., 2017].

2.7.2 SPMF

SPMF (Sequential Pattern Mining Framework) est une bibliothèque de fouilles de données open source écrite en Java par Philippe Fournier-Viger et ses collaborateurs [Fournier-Viger et al., 2016]. Cette suite d'algorithmes est entièrement spécialisée en fouille de motifs ; elle est pour cela considérée la suite la plus riche sur le sujet [Fournier-Viger et al., 2017], car elle inclut 138 implémentations de différents algorithmes de fouille d'itemsets, de séquences, classifications, etc. SPMF est conçue de façon à offrir une intégration simple à d'autres projet de fouille de données. Elle est offerte aussi comme une application indépendante en deux modes : par interface graphique ou via la ligne de commande. Notons que le site de cette bibliothèque inclut plusieurs autres données utiles telles qu'une documentation, des bases de données réelles et synthétiques, des études de performance et une mailing-liste, etc.

2.7.3 KNIME

KNIME (Konstanz Information Miner) [Berthold et al., 2009] est un environnement interactif d'analyse de données, modulaire en pipeline à travers des composants prédéfinis appelés nœuds, extensible et multi-plateformes basé sur Eclipse. Fondé par Michael Berthold comme nouveau venu du monde exploration de données et bien qu'originellement orienté bio-informatique, il connaît un succès aussi bien dans les sphères professionnelles que celles d'enseignements et de recherches. KNIME offre des solutions d'entrée/sortie, de préparation, d'analyse et de visualisation avancées simples et rapides avec une intégration parfaite à d'autres langages et technologies et une variété de sources de données. Compte tenu de ses atouts, il est sélectionné parmi les quatre premiers outils de fouilles de données [Altalhi et al., 2017]. Outre l'intégration aux bases de données via des ports dédiés et à des outils connus

d'exploration de données tels que Weka et R, KNIME offre deux options pour la fouille de motifs : l'Item Set Finder node et Association Rule Learner implémentant les algorithmes : Apriori, FPGrowth et bien d'autre, et l'Association Rule node qui produit à la fois les motifs fréquents et les règles d'association.

2.7.4 Rattle et R

«The R Analytical Tool To LearnEasily» ([Williams, 2011] est une interface graphique orientée d'onglets pour la fouille de données en utilisant le langage de programmation statistique R |R Core Team, 2013]. Elle est un logiciel libre et inclut de multiples fonctionnalités de la préparation, la transformation et le résumé des données à la création de modèles supervisés ou non (dits paradigmes dans Rattle), et de présentation graphique et évaluation. Une propriété remarquable de cet outil est que l'ensemble des actions opérées en interagissant avec l'interface peuvent être capturées sous la forme d'un script R invocable alors de façon indépendante de l'outil. Rattle support la fouille de motifs via l'onglet Associate du paradigme Unsupervised, il implémente l'algorithme Apriori avec les mesures d'intérêt suivants : le support/confiance, le Lift et le Leverage. Par ailleurs, notons que le langage R inclut ARules [Hahsler et al., 2011] une bibliothèque dédiée complètement à l'extraction de motifs fréquents (tous, fermés et maximaux) et les règles d'association associées.

2.7.5 Tanagra

Tanagra est un logiciel open source écrit en C++ par *Ricco & Rakotomalala* [Rakotomalala, 2005]. Fonctionnant exclusivement sous Windows et seulement via une interface graphique, il propose des modules pour l'analyse exploratoire et statistique de données, ainsi que l'apprentissage automatique. Tanagra offre une implémentation de l'algorithme Apriori pour

l'extraction de motifs fréquents. Un paramétrage simple permet de limiter la sortie aux motifs fréquents fermés ou maximaux. L'absence de maintenance et de mise à jour depuis décembre 2013, ainsi que l'imputabilité et l'inaccessibilité de son site sont des points faibles de cet outil.

2.7.6 Mahout

Mahout est une bibliothèque open source d'apprentissage automatique et de fouille de données développée en Java par la fondation Apache à partir de 2008 comme une partie du projet du moteur de recherche Lucerne de la même fondation [Owen et al., 2011]. Elle est destinée à des projets énormes impliquant des données très volumineuses. La force de Mahout est attribuée à ses capacités de mise à l'échelle, de parallélisme et de distribution fondées sur le paradigme MapReduce de Hadoop [Dean and Ghemawat, 2008]. Ce logiciel d'apprentissage automatique n'offre aucune interface graphique et est utilisable exclusivement par la ligne de commande et exige plusieurs autres outils, qui sont à adapter par les développeurs.

Elle a offert initialement des algorithmes de classification, de segmentation et de recommandation. Actuellement, la fouille de motifs y est intégrée à travers PFP [Li et al., 2008] une implémentation parallèle de l'algorithme FPGrowth.

2.7.7 Orange

Orange est un environnement libre livré sous la licence GPL, destiné à l'apprentissage automatique et l'exploration et la visualisation de données [Demšar et al., 2013]. Écrit en Python par le laboratoire de Bio-informatiques de l'université de Ljubljana (Slovénie), il adopte la philosophie de la programmation visuelle basée sur des composants. Des widgets sont donc disponibles pouvant être connectés pour créer des workflows qui constituent

l'usage standard. Les utilisateurs expérimentés peuvent user sa bibliothèque Python pour différentes sortes de traitements et de modifications de composants. Le support de la fouille de motifs est limité dans Orange à deux variantes de l'algorithme Apriori.

2.7.8 ELKI

ELKI (Environment for deveLoping KDD-Applications supported by Indexstructures) est un logiciel de fouille de données open source indépendant de toute plateforme écrit en Java [Achtert et al., 2008], dont la finalité est la recherche en algorithmique. Bien maintenu et à jour, cet outil est spécialisé dans les tâches non supervisées, plus particulièrement l'analyse de clusters et la détection d'outliers. Il est conçu dans l'optique d'être paramétrable et extensible afin de permettre des expérimentations et des évaluations rapides et simples. Outre les fameuses techniques de segmentation et de détection d'outliers, ELKI offre deux algorithmes de classification et les trois algorithmes de fouille de motifs les plus célébrés : Apriori, Eclat et FPGrowth.

De nombreux outils de fouille de données existent ; chacun offre des fonctionnalités intéressantes. Les présenter tous, n'est pas l'objectif de ce mémoire. Une description et une comparaison de certains de ces outils en considérant plusieurs caractéristiques peut être trouvée dans [Altalhi et al., 2017], [Fournier-Viger et al., 2017]. Trois autres bibliothèques dignes d'être mentionnées dans ce contexte sont : DMTL [Chaoji et al., 2008], iZi [Flouvat et al.,2008].

Parmi les logiciels libres, ci-dessus énumérés quelques-uns sortent du lot :

KNIME [archive] (prononcer NAÏM), acronyme de Konstanz Information Miner1, est un logiciel libre édité par un laboratoire de l'université de Constance dénommé Nycomed Chair for Bio-informatics and Information Mining2,3. Il intègre notamment tous les modules

d'analyse de Weka et permet de créer des scripts en langage R. Ces deux logiciels sont décrits ci-dessous. KNIME s'exécute sur Linux, Windows et Mac OS. Comme tous les logiciels libres, KNIME est extensible.

R4 est un langage et un environnement permettant d'effectuer des calculs statistiques et de créer leurs graphiques. Sous licence GNU, R est semblable au langage S et à son environnement créé aux Laboratoires Bell par John Chambers et ses collègues. R peut être considéré comme une autre mise en œuvre de S. Il y a quelques différences importantes, mais beaucoup de code écrit pour S s'exécute inchangé sous R. R fournit un large éventail de techniques statistiques et graphiques telles que la modélisation linéaire et non linéaire, les tests statistiques classiques, l'analyse des séries chronologiques, la classification et le clustering, entre autres. Il peut être fortement étendu par des programmes développés par la communauté. Le langage S est souvent le véhicule de choix pour la recherche en matière de méthodologie statistique, et R fournit une voie open source à la participation à cette activité. Un des atouts de R est la facilité avec laquelle des graphiques bien conçus, de qualité digne de publication, peuvent être produits, contenant des symboles mathématiques et des formules si besoin est. Un grand soin a été accordé à la prise en charge des options par défaut pour les choix mineurs dans la conception des graphiques, mais l'utilisateur conserve le contrôle complet de ces options. R est publié selon les termes de la licence GNU sous forme de code source. Il se compile et s'exécute sous une grande variété de plates-formes UNIX et de systèmes similaires, y compris FreeBSD et Linux, Windows et Mac OS.

Orange5 est un logiciel libre créé à l'université de Ljubljana en Slovénie. Ce logiciel est doté d'une interface homme-machine conviviale. Il est développé en C++ et en Python. Chaque algorithme se présente sous la forme de widgets pouvant avoir une entrée et une sortie ; ils sont agencés dans une fenêtre6.RapidMiner est un logiciel libre distribué par la société

Rapid-I7, basée à Dortmund en Allemagne. Il intègre le Business Intelligence dont les principales fonctionnalités sont l'ETL, l'OLAP, la production d'états et l'exploration de données et les techniques classiques comme les SVM, l'ACP, les arbres de décision et les réseaux neuronaux. Ce produit est aussi distribué en version commerciale.

Tanagra 8 est un logiciel libre d'exploration de données développé sous la direction de *Ricco Rakotomalala* du laboratoire ERIC de l'Université Lumière Lyon 2. Il permet d'effectuer les traitements d'analyses factorielles telles que l'ACP, l'AFC, l'ACM, la régression PLS, de classification non supervisée avec l'algorithme des *k-means* et l'algorithme hiérarchique ascendant. Il permet aussi d'importer des fichiers au format weka6.

Weka est un logiciel libre de fouille de données créé par l'université de Waikato (Nouvelle-Zélande). C'est une collection d'algorithmes d'apprentissage automatique mis en place pour effectuer des tâches d'exploration de données9. Les algorithmes peuvent soit être appliqués directement à un ensemble de données soit être appelés directement par un code Java développé par une équipe informatique indépendante par exemple. Weka contient des outils pour les prétraitements des données, la classification, la régression, le clustering, les règles d'association et la visualisation. Il est également bien adapté au développement de nouveaux schémas pour l'apprentissage automatique. C'est un logiciel open source publié sous la LGPL6.

2.8 Analyse des travaux présentés de la littérature

Comme nous l'avons indiqué tout au long de cette analyse bibliographique, les recherches autour de l'extraction de motifs s'attaquent à deux grands défis qui sont la définition de méthodes et d'outils permettant d'appréhender de très grands volumes de données et la sélection des motifs qui sont potentiellement intéressants. Ainsi, dans la littérature, nous constatons que plusieurs approches méthodologiques sont proposées pour extraire des

connaissances dans les bases de données. S'intéressant de près à la gestion du risque dans le système de contrôle douanier, on constate que les études menées ont plutôt mis l'emphase sur des outils statistiques. En effet, l'exploitation des informations est faite en utilisant des techniques d'analyse de données et d'économétrie ; la sélectivité de ces données repose sur une analyse de risques réalisée à partir des informations recueillies sur les fraudes constatées (*fraude avérée*), et non sur d'éventuels soupçons de fraude, et à ce propos les travaux de [Truel, 2010], [Hintsa et al.,2011] [Geourjon et al., 2010] confirment ces études. Elles sont statiques et figées car les règles définies sont peu souvent actualisées.

L'émergence de nombreuses méthodes de fouille des données ont favorisé leur application à divers domaines d'activités, et nous citons ces quelques travaux de recherches : l'analyse du comportement d'achat des consommateurs dans un supermarché [Agrawal et al. 1994], l'analyse de la pertinence des traces patients pour dans le domaine médical pour étudier l'activité du patient, dans ses composantes comportementales et cognitives. [Mokeddem, 2016], l'extraction de connaissances sur les comportements potentiellement à risques de navires [Bilal, 2013], les prévisions météorologiques et l'analyse du mouvement suivi de migration d'aigles [Li et al., 2011]. À la problématique de recherche des motifs fréquents introduite par Agrawal et Srikant [1995] dans le contexte du panier de la ménagère, on constate que cette méthode a été appliquée avec succès à de nombreux domaines comme la biologie [Wang et al., 2004b],[Salle et al., 2009], la fouille d'usage du Web [Pei et al., 2000],[Masseglia et al., 2008], la détection d'anomalies [Rabatel et al., 2010], la fouille de flux de données [Marascu et Masseglia, 2006] ou la description des comportements au sein d'un groupe [Perera et al., 2009].

2.9 Question de recherches et positionnement de nos travaux

Malgré l'efficacité démontrée des algorithmes de datamining dans plusieurs secteurs sus référencés, il n'existe pas d'équipes de recherches ayant tenté d'adapter les modèles existants pour les exploiter dans un contexte d'analyse stratégique lié aux risques de fraude dans l'activité douanière, en d'autres termes la littérature ne mentionne pas une application formelle des approches de fouille de données aux activités douanières. D'où tout l'intérêt que nous portons à ce projet de recherche. C'est dans ce contexte que se pose la question de recherches suivante : Quel moyen utilisé pour analyser les données issues des opérations de contrôle douanier et en extraire des connaissances ?

Dans cette optique, nous proposons les techniques du Datamining pour découvrir des connaissances afin d'anticiper et analyser le comportement à risque lié à l'activité douanière étant donné que la fouille de données concerne des données d'observation par opposition à des données expérimentales qui peuvent être utilisées dans des domaines connexes comme la statistique. Toutefois le succès de la démarche ou approche proposée repose sur la bonne volonté des intervenants à faire remonter l'information concernant les risques de fraude. Si aucun intervenant ne révèle d'information relative à un nouvel état d'un moyen de réduction du risque ou à un nouvel incident, l'information ne sera jamais enregistrée dans la base de données. Alors, les règles décrivant le risque ne seront jamais actualisées. Conséquemment, il en sera autant pour les facteurs de risques et les causes potentielles de fraude, ainsi que les probabilités associées. Dans pareil contexte, des décisions dépassées risquent d'être prises pour l'anticipation des risques. Une culture de sécurité et une confiance mutuelle dans l'administration douanière sont primordiales afin d'encourager la remontée d'information relatives aux infractions douanières pour brosser un portrait plus juste du risque et améliorer l'efficacité des moyens de réduction du risque.

2.10 Conclusion

Au terme de cette analyse bibliographique, on constate qu'il existe dans la littérature une multitude d'approches pour fouiller les données. Les différents modèles développés varient selon le domaine d'activités et les résultats attendus à cet effet. Dans la majeure partie des travaux effectués sur la question de gestion du risque lié à l'activité douanière, les résultats obtenus émanent de l'emprunt des outils statistiques. Ce qui a permis de poser les questions relatives à notre travail de thèse de doctorat, et positionner notre travail de thèse vis-à-vis de ces questions. Dans les chapitres 3 et 4 suivants, nous présentons les contributions apportées pour juguler et décrire les comportements à risques liés à l'activité douanière.

CHAPITRE 3 : Règles d'Associations sur la Base de Motifs Fréquents

Résumé du chapitre *: le troisième chapitre est notre première contribution ; il s'agit de l'apport des règles d'association à partir de l'algorithme classique (APRIORI) appliqué sur la base de motifs fréquents, à une masse de données d'infractions douanières issues des transactions douanières afin d'extraire des connaissances. Trois grandes règles d'association issues de données binaires ont été mis en évidence : les règles de prévision, les règles de ciblage et les règles neutres ont permis d'établir une association entre une opération de dédouanement et la nature de l'infraction.*

Sommaire : Pages

3.1	Introduction	111
3.2	Motivation et problématique	111
3.3	Problème d'apprentissage : Les règles d'associations	114
3.4	Observations sur les travaux de la littérature	115
3.5	Approche méthodologique : Définition du problème	118
3.6	Modélisation du modèle mathématique de la cartographie des risques	137
3.7	Espace des données à explorer	140
3.8	Expérimentation : Mise en contexte et Résultats	144
3.9	Comparaison du modèle Econométrique à l'approche fouille de données	154
3.10	Conclusion	155

3.1 Introduction

Dans le chapitre précédent, nous avons présenté l'apport des méthodes et algorithmes de fouille de données ou datamining dans plusieurs secteurs d'activités. Les résultats de recherche obtenus dans ces secteurs indiquent que les mêmes approches, avec les ajustements nécessaires, pourraient être incorporées dans le processus d'analyse des risques de fraude en matière douanière pour ainsi fournir des indicateurs pour la prise de décision. Dans un premier temps pour anticiper les risques de fraude, et dans le second, faire de la prédiction sur le comportement à risque des opérateurs lors des opérations de dédouanement. Dans cette section, nous présentons notre première contribution de notre projet de thèse. En effet, nous proposons une méthode apriori pour générer des règles d'association à partir de données relatives aux infractions douanières sur la base d'une corrélation entre Type et Nature d'une fraude afin d'en extraire des connaissances.

3.2 Motivation et problématique

La mobilisation des recettes douanières dans les pays en voie de développement constitue une priorité tant au regard de l'équilibre des finances publiques qu'en matière de réduction de la pauvreté. En raison du contexte de réduction de l'assiette des recettes douanières corolaire des intégrations économiques, de la libre circulation, des processus de démantèlements tarifaires, d'accords de partenariats économiques et de fraudes à grandes échelles, les douanes dans le cadre de la mobilisation des recettes doivent recourir à des méthodes robustes d'analyse et de gestion des risques pour une efficacité du contrôle douanier.

Compte tenu du volume important des échanges commerciaux sur le plan mondial, les administrations douanières les plus modernes s'appuient sur le développement technologique des dispositifs de collectes de données numériques pour stocker de très grandes quantités de

données pour analyser le risque de fraude. Ce système (analyse du risque) est alors fréquemment utilisé pour la recherche, l'évaluation et la planification à d'autres fins en termes d'analyses et de prévisions des infractions dans les administrations douanières. Ce système est également un moyen efficace en vue de lutter contre les contrôles intrusifs répondant ainsi aux exigences des opérateurs privés pour sécuriser leurs transactions [Harisson, 2007] ; cependant, elle est uniquement basée sur le renseignement fourni lors de contrôles en vue de lutter contre les mauvaises pratiques [Truel, 2010]. En effet, dédouaner une marchandise, ne veut pas dire payer les droits et taxes afférents, mais plutôt, l'accomplissement de toutes les formalités douanières pour l'affectation d'un régime douanier à ladite marchandise et ce même en l'absence de paiement de droit de douane. On voit ainsi, qu'adaptant à chaque contexte, l'analyse du risque nécessite chaque fois une démarche spécifique [Gates, 2006]. De plus, c'est une aventure risquée pour les recettes, car cette méthode néglige le comportement à risque des agents des douanes [Geourjon et Laporte, 2004].

Vu le nombre important de transactions douanières et la multiplicité des risques, l'analyse du risque doit évoluer pour faire place à de nouveaux défis. Parmi les travaux de la littérature traitant de ces questionnements dits de système de surveillance des infractions douanières ; une première tentative a été de proposer une approche économétrique capable de cibler au mieux les déclarations douanières qui présentent un risque réel de fraude. Ce modèle économétrique développé par Laporte [Laporte, 2011] permet de déterminer les critères de risques pertinents pour expliquer la fraude à partir de l'analyse statistique de l'historique des déclarations et de calculer la probabilité de fraude pour toute nouvelle déclaration [Laporte, 2011]. Le modèle est présenté comme ce qui suit :

$$\Pr(fraude_i = 1) = \alpha + \beta_1 fq\_critère1_i + \beta_2 fq\_critère2_i + \cdots + \beta_N fq\_critèreN_i + \varepsilon_i$$

Avec : Pr : *la probabilité* ; fraude_i : *variable binaire* 0-1 pour l'opération *i* (1 si la fraude est constatée et 0 sinon) ; fq_i : *la fréquence de fraude* pour chaque critère de risque associé à l'opération *i et* ε_i *: l'écart aléatoire* et les paramètres à estimer. L'insuffisance dans ce modèle est qu'il ne tient pas compte de la nature de l'infraction. Pour résoudre ce problème, Laporte propose deux autres modèles à partir d'un modèle de probabilité linéaire : PROBIT ou LOGIT plus approprié pour estimer un modèle dont la variable expliquée est binaire en théorie mais la valeur prédite ne peut pas être interprétée comme une probabilité de fraude car n'appartenant pas à l'intervalle [0,1]. Nous pouvons également citer d'autres méthodes proposées comme la technique de scoring pour avoir une approche plus structurée en évaluant de manière efficace le risque et orienter les déclarations dans les différents circuits de contrôle des administrations de douanes des Pays En Voie de Développement (PED). Geourjon [Geourjon et al. 2012] ont montré dans un article de recherches la pertinence de cette technique à partir d'une expérience réalisée au Sénégal. Ils mettent en évidence la technique du scoring relativement simple permettant dans les douanes des Pays en Développement d'évaluer le risque pour limiter efficacement les contrôles, et que leur développement contribue à la modernisation des administrations [Geourjon et al. 2012]. Une étude réalisée par Grigoriou prône la technique du scoring pour organiser les contrôles tout en s'assurant du respect des normes techniques, sanitaires et phytosanitaires [Grigoriou, 2011].

Nous constatons que les différentes méthodes recensées mettent en évidence des progrès obtenus en termes de facilitation dans le processus de contrôle. Cependant trop de questions restent en suspens quant à leur uniformité dans les différentes administrations douanières. Les travaux de Geourjon et al. ont montré que chaque administration a adopté une démarche propre et adaptée à son contexte et à ses besoins [Geourjon et al. 2012]. Vu que l'analyse et la gestion du risque pour orienter les déclarations dans les différents circuits de contrôle

reposent principalement sur une exploitation des données, nous proposons une approche intégrée qui exploite l'existant en matière de fouille de données. L'idée sera d'explorer des données historiques et d'exploiter les relations usuelles entre ces données dans le but de découvrir les connaissances ayant conduit aux infractions douanières. Ces connaissances vont servir à l'identification automatique des infractions liées aux activités douanières à partir des bases de faits (*Procédure de dédouanement, enquête douanière, matérialisation du contrôle*)

Pour illustrer notre démarche, si nous cherchons le mot "Fraude" dans Google, nous obtenons 60 000 000 réponses nous guidant vers des sites contenant ce mot. Supposons que nous soyons assez rapides pour consulter une page toutes les trois secondes, il nous faudra un peu plus de 1000 ans pour toutes les visiter. Cette tâche est irréalisable. Il nous faut donc un moyen pour non seulement stocker et rechercher des informations, mais également pour les analyser et les interpréter pour aider à la prise de décision. Il apparait la nécessité de mettre en place un Système Intelligent d'Aide à la Décision (SIAD) pour identifier des fraudes et la découverte à priori des situations d'infractions. C'est dans ce contexte bien précis que se situe notre problématique suivante : *Comment éliciter des règles d'association à partir des données de l'activité douanières sur la base de motifs fréquents pour analyser le risque douanier ?*

3.3 Problème d'apprentissage : Les règles d'associations

Bien avant l'essor que connait actuellement le domaine des technologies de l'information et de la communication, la problématique d'extraction de connaissances des données a été de tout temps posés. Le développement des technologies tant en matière de stockage de données que du traitement de l'information, a rendu cette tâche d'extraction de connaissances davantage plus ardue [Kantardzic, 2003]. En effet, on assiste non seulement à une croissance

exponentielle du volume d'informations stockées au sein de nos organisations mais également à une complexification de ces données [Tan, et al., 2006]. La fouille de données est définie comme le processus non trivial d'extraction d'informations implicites, nouvelles et potentiellement utiles à partir de grands volumes de données [Fayyad et al., 1996]. Elle propose d'utiliser un ensemble de techniques et algorithmes qui ont pour objet la découverte de motifs et des connaissances à partir de grandes quantités de données [Berry and Linoff, 2011]. La fouille de données constitue l'étape clé du processus de découverte de connaissances. Même si l'étape de fouille de données n'est qu'une partie du processus général pour la découverte de connaissance, elle est celle qui suscite le plus de travaux dans la littérature. Nous pouvons citer les techniques et méthodes mises en œuvre pour guider le processus et arriver à une extraction efficace de connaissances au sein des entrepôts de données. Lesquels ont été regroupés sous l'appellation ECD pour Extraction de Connaissances à partir des Données ou KDD (*Knowledge Discovery in Databases est le processus global de fouille de données comportant plusieurs étapes dont la fouille de données ; Par abus de langage ECD et KDD sont souvent confondus au sein de la littérature*). L'extraction de règles d'associations fait partie intégrante d'un processus d'extraction de connaissance de données (ECD). C'est une technique non supervisée de datamining qui permet à partir des données d'un ensemble apparaissant fréquemment dans un entrepôt de données, d'extraire des connaissances.

3.4 Observations sur les travaux de la littérature : *Limites et pistes d'amélioration*

La facilitation des échanges accentuée par la mondialisation a entrainé une croissance rapide de la taille des volumes de données stockées dans les Bases de Données disponibles dans les administrations douanières. Face à ce volume croissant de marchandises, l'analyse des

risques s'est avérée indispensable pour dégager les risques de fraudes en vue de limiter le contrôle inclusif et optimiser le contrôle douanier.

L'approche d'analyse des risques basée sur la statistique descriptive et l'économétrie ont permis la modernisation du système d'information des administrations douanières dans les PED [Goujon et al. 2012], cependant, force est de reconnaitre que, la méthode de la statistique descriptive ne permettait que de découvrir des irrégularités statistiques dans les situations de fraude sur une période donnée. Les résultats obtenus ne permettaient que de définir la probabilité pour que toute nouvelle déclaration présente une irrégularité (Voir Figure 3.1).

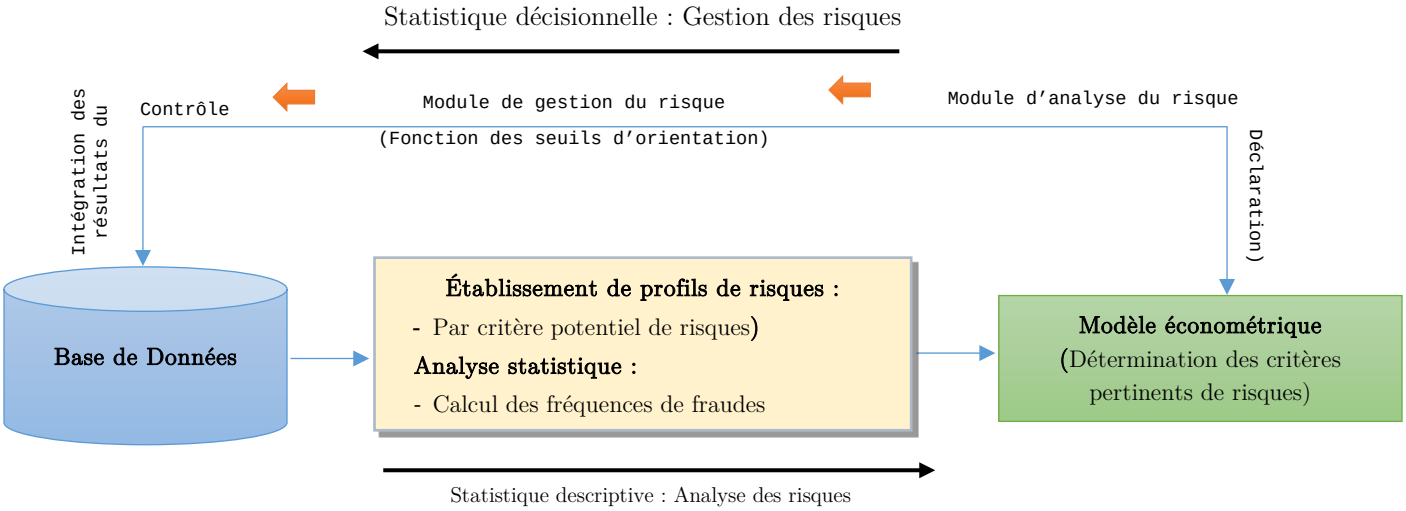

Figure 3.1: Modèle économétrique pour la gestion du risque en douane

L'approche économétrique développée par Laporte, présente des limites pour anticiper les risques de fraudes. En effet, le nombre de variables du modèle devrait être augmenté ce qui va entrainer l'analyse de données nouvelles et de nouveaux paramètres à intégrer au modèle à chaque nouvelle information. Une nouvelle approche méthodologique (fouille de données) s'impose au traitement des données d'infractions douanières. En effet, il s'agit de découvrir

des règles d'expertise pour aider à l'analyse des risques liés aux opérations douanières *(Voir Figure 3.2).*

Deux aspects sont relevés pour motiver cette action :

- Extraire une règle générale à partir de données observées sur la base des motifs fréquents.
- Découvrir de nouvelles règles de connaissances après l'analyse de ces données.

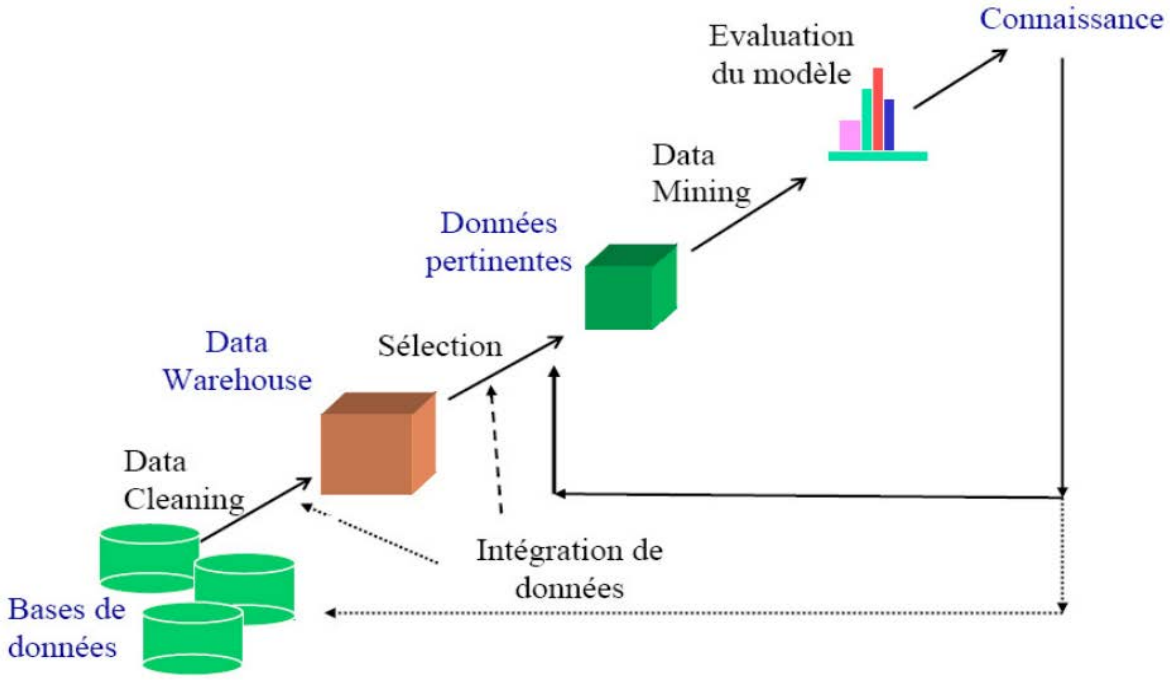

Figure 3.2 : Processus d'analyse des données

Le cloud Computing a entrainé la collecte de grandes quantités de données. La découverte de motifs dans les données est l'une des problématiques phare en fouille de données. Ainsi, rechercher des motifs fréquents a été proposé pour faciliter l'extraction des règles d'associations fréquentes [Agrawal et al. 1993, 1996]. Cette approche donne ainsi une meilleure abstraction des trajectoires et diminue la taille des données pour l'analyse.

Cao et al. se sont intéressés à l'extraction de motifs périodiques dans des BD climatiques. Les objets étudiés par exemple des orages, ont la particularité de suivre approximativement la même route à intervalles de temps réguliers, C'est à dire très fréquemment, il y a des pluies saisonnières au début de l'été [Cao et al., 2005, 2007].

Afin d'extraire des règles de connaissances dans une base de données relative aux trajectoires de bus, Fisher et al. ont mis en évidence certains motifs qui à priori sont des groupes d'objets partageant le même type de déplacement (*direction, vitesse*) [Fisher et al., 2005].

L'extraction de règles de connaissances à partir de motifs fréquents a été largement étudiée dans la littérature. Les travaux présentés dans ce document ne sont pas exhaustifs. C'est dans ce contexte d'étude que se situent les travaux de ce chapitre du mémoire de la thèse où nous étudions l'apport de la méthode des motifs fréquents pour extraire des connaissances à partir d'une masse de données relative aux infractions douanières.

3.5 Approche méthodologique : Définition du problème

L'ECD (*Extraction de Connaissance à partir des Données*) désigne le processus non trivial d'extraction d'informations implicites, précédemment inconnues et potentiellement utiles à partir des données [Frawley et al., 1991]. Les travaux de recherches effectués dans cette partie de notre mémoire de thèse s'inscrivent dans le domaine de l'extraction de connaissances à partir des données. Les connaissances extraites peuvent prendre différentes formes mais nous allons nous intéresser, dans ce manuscrit, seulement aux connaissances extraites sous la forme de règles d'associations. L'extraction de règles d'association consiste à découvrir des relations entre les variables d'un entrepôt de données, et cette extraction de connaissances s'appuie sur la fréquence des motifs.

L'approche générique proposée est basée sur un processus itératif non supervisé qui va extraire des motifs fréquents à partir d'une masse de données des infractions douanières les

unes après les autres permettant ainsi l'exploration pas à pas des données. L'idée est de découvrir des règles associatives adaptées au contexte douanier pour identifier et résoudre les problèmes liés aux fraudes et infractions douanières.

Cette approche va fonctionner sur la base de recherches de structures intrinsèques, des relations, ou affinités entre les données. En d'autres termes, il s'agit de trouver des tendances et corrélations qui résument les relations entre données (Larose, 2005 ; Tan, et al., 2006 ; Hornick, et al., 2007) L'objectif est de découvrir des règles d'associations pour aider à la détection des situations à risques (Fraudes, infractions). Le processus itératif comprenant plusieurs étapes qui s'étend de la spécification du problème à l'interprétation et l'évaluation des résultats. Sur la figure 3.1, nous schématisons le processus de l'ECD en nous inspirant des travaux de thèse de Frédéric Pennerath [Pennerath, 2009]

Comme indiqué sur a figure 3.3, les étapes de l'extraction de connaissances sont partitionnées en huit étapes :

1. Étape 1 : *Spécification de la problématique* qui consiste à définir clairement la question ouverte visée par l'ECD
2. Étape 2 : *Processus de la sélection des données*. Elle consiste à conserver uniquement les données qui vont permettre de répondre à la question. Dans notre cas ici présent la structure de données permet une représentation séquentielle des fraudes. Ainsi, il faut donc effectuer un tri afin de ne garder que le sous-ensemble de données réellement utile à la résolution du problème.
3. Étape 3 : *Processus de prétraitement des données*, C'est une phase de nettoyage qui a pour objectif d'améliorer la qualité des données. Plusieurs méthodes vont être appliquées afin de supprimer le bruit, de compléter les données manquantes, de détecter et corriger les données incohérentes ou encore de gérer la présence de redondances.

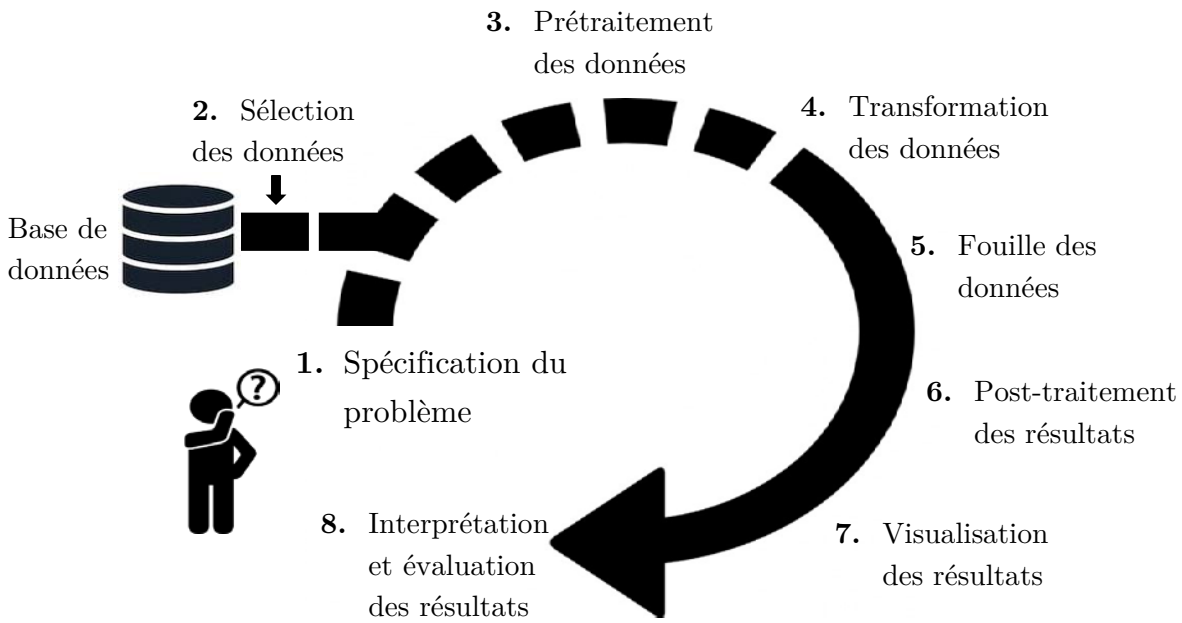

Figure 3.3: Les différentes étapes de l'Extraction de Connaissances des Données

4. Étape 4 : Transformation des données, puisque dans certains cas il faut formater les données pour pouvoir appliquer l'algorithme d'extraction de connaissances. En effet, pour s'exécuter, un algorithme nécessite un certain type de fichiers (*fichier csv, fichier arff, base de données MySQL…*), mais également un certain type de données (*binaires, continues, discrètes, etc.*). À titre d'exemple, dans le logiciel Weka [Hall et al., 2009], l'implémentation de l'algorithme d'extraction de règles d'association FP-Growth [Han et al., 2000] nécessite des données binaires pour s'exécuter tandis que celle de l'algorithme Apriori [Agrawal and Srikant, 1994] peut également travailler sur des données discrétisées.
5. Étape 5 : **Fouille des données** qui est l'étape centrale de l'ECD. Elle permet d'extraire des connaissances inconnues et utiles à partir des données en appliquant des algorithmes automatiques ou semi-automatiques. Il s'agit de trouver ici tous les « patrons » ou itemsets fréquents, qui apparaissent dans l'entrepôt de données

avec une fréquence supérieure ou égale à un seuil défini par l'utilisateur, appelé *Minsup (Valeur minimal de la mesure du Support (voir la formule dans le tableau 1.2).*

6. Étape 6 : *Post-traitement* est la sixième étape et permet d'améliorer la qualité des résultats générés à l'étape 5. En effet, et notamment dans le cas des règles d'association, les résultats sont parfois retournés en quantité trop importante et rendent difficile l'analyse et l'identification de ceux qui sont réellement intéressants. Afin de résoudre ce problème, un post-traitement utilisant différentes mesures de qualité va aider l'expert à mieux évaluer les résultats : il va falloir générer l'ensemble des règles associatives, à partir de ces patrons fréquents, ayant une mesure de confiance supérieure ou égale à un seuil défini par l'utilisateur, appelé *Minconf* et choisir les motifs les plus représentatifs pour établir des règles de connaissances tout en supprimant la redondance présente dans ceux-ci.
7. Étape 7 : *Visualisation des résultats*. Cette étape permet de représenter sous forme visuelle des résultats conformément aux attentes des utilisateurs.
8. Étape 8 : *Interprétation et l'évaluation des résultats*. Cette dernière étape va consister à expliquer les résultats obtenus avec les données de départ conformément aux objectifs. Si les conclusions de cette étape ne sont pas en accord avec ce qui est attendu, le processus peut revenir à n'importe quelle étape afin de continuer et d'affiner l'analyse en cours.

Dans un monde qui s'ouvre de plus en plus facilement aux nouvelles technologies, il devient très facile de collecter et de stocker des informations. Ainsi, l'analyse ces données statistiques constitue une clef importante pour la découverte de nouvelles connaissances mais la difficulté provient de la nature des données statistiques, qui sont fortement corrélées et denses, ce qui pose souvent le problème d'efficacité [Stewart and White, 1991], [Brin et al., 1997].

Dans le cadre de cette contribution, nous allons nous focaliser sur la fouille de données (*5ème étape du processus de l'ECD*) qui consiste à l'extraction de connaissances inconnues et utiles à partir des données. Les connaissances extraites peuvent prendre différentes formes mais nous allons nous intéresser, dans ce manuscrit, à la représentation des données sous la forme de graphes où chaque objet est un graphe conceptuel. Les sommets du graphe représentent les entités ainsi que leurs attributs car ce formalisme est plutôt adapté aux méthodes d'analyse descriptive et la recherche des motifs fréquents tels que traités dans notre première contribution.

La découverte de ces règles aura différents objectifs en fonction des données analysées.

3.5.1 Définition du problème

Les connaissances extraites peuvent prendre différentes formes mais nous allons nous intéresser, dans cette contribution, seulement aux connaissances extraites sous la forme ***de règles d'association.***

Remarque : Pour permettre au lecteur de comprendre le vocabulaire utilisé sur les termes : *Base de données transactionnelle Item, motif, k-motif, transaction, règle d'association, prémisse, antécédent ;* nous l'invitons à se référer au chapitre 1 du mémoire de cette thèse, précisément à la section 1.3

Définition *3.1* - (*Base de données transactionnelle*) *:* Les bases de données considérées ici sont de simples tables contenant l'information, éventuellement construites par jointures à partir de plusieurs relations. L'exemple du tableau 3.1 répertorie les valeurs de trois attributs multivalués X_1, X_2 et X_3 pour 8 objets d'étude, appelés également *n-uplets*. Dans cet exemple, les deux premiers attributs X_1 et X_2 sont de type symbolique ou qualitatif car leur domaine de définition est discret. A contrario, le dernier attribut X_3 est numérique ou quantitatif.

Tableau 3.1: Modèle de base de données au format attribut /valeur

	Attributs (a)		
Objet (o)	X_1	X_2	X_3
o_1	+	→	.
o_2	-	→	.
o_3	+	→	.
o_4	+	←	b_i
o_5	-	→	.
o_6	-	→	.
o_7	+	←	.
o_8	-	←	.

Avec$b_i \in \mathbb{R}^+$avec $i \in \{1,2;3;4,5,6,7,8\}$

Vu que nous travaillons sur l'extraction de motifs ensemblistes, où un motif est un ensemble d'attributs booléens ; Cela nécessite de discrétiser les attributs numériques, afin de disposer de données booléennes. Il sort du cadre de ce mémoire de discuter précisément des méthodes de discrétisation qui permettent d'obtenir de tels contextes booléens à partir d'attributs multivalués ou continus. Cette étape de prétraitement des données est difficile dans le cas d'attributs numériques : il faut regrouper ensemble des valeurs différentes qui expriment la même information ou définir des intervalles. Les connaissances des experts se révèlent indispensables pour effectuer les bons choix lors de cette opération délicate. Nous supposons donc obtenir (tableau 3.2) une matrice booléenne qui indique pour chaque objet les attributs qu'il contient. Ce format est usuellement qualifié de transactionnel. Dans ces contextes booléens, un attribut est souvent appelé item et un objet transaction.

Tableau 3.2: Modèle de base de données au format transactionnel

	Attributs (a)						
Objet (o)	.	.	.	a_i	.	.	.
.	×		×	×			
.		×	×	×			
.	×		×	×			
o_j	×			×	×		
.		×	×		×		
.		×	×		×		
.	×			×			×
.		×		×			×

Une base de données booléenne B est notée sous la forme d'un contexte formel $B = (A, O, R)$ où $A = \{a_1 \ldots . a_m\}$est l'ensemble des attributs, $O = \{o_1 \ldots . o_n\}$celui des objets et M une relation binaire entre A et O.

Positionnement du problème

Définition 3.1 (*Item, motif et k-motif*) : Soit $I = \{i_1, \ldots, i_p\}$ un ensemble de p d'items, où chaque item est une variable binaire de la base de données. Un ensemble d'items est appelé un motif, et plus spécifiquement on dit que X est un *k-motif* s'il est composé de k items : $X = \{i_1, \ldots, i_k\}$.

Remarque : On utilisera les minuscules pour représenter les items et les majuscules pour représenter les motifs dans la suite du manuscrit. Par conséquent, x est un item et X est un motif

Définition 3.2 (*Notion de transaction*) : Soit $D = \{t_1, \ldots, t_n\}$,un ensemble de n transactions t, où chaque transaction t est un ensemble d'items tel que $t \subseteq I$. Une transaction t de D contient X, un ensemble d'items de I, si $X \subseteq t$.

Définition 3.3 (*règle d'associations*) : Une règle d'associations est une implication de la forme $X \Rightarrow Y$, où $X \subseteq I$, $Y \subseteq I$, *et* $X \cap Y = \emptyset$.

Une règle $X \Rightarrow Y$ indique que les transactions possédant le motif X ont tendance à posséder le motif Y. Cependant, il n'existe aucune relation de causalité entre X et Y : la présence de X ne cause pas la présence de Y.

Définition 3.3 (*Prémices, antécédent, conclusion et conséquent*) : La partie gauche de la règle est appelée la prémisse ou l'antécédent et la partie droite est la conclusion ou le conséquent.

Pour une règle $X \Rightarrow Y$, X est donc la prémisse ou l'antécédent et Y est donc la conclusion ou le conséquent.

Définition 3.4 (*Support, support relatif, support absolu, support minimum : MinSup et motif fréquent*) : Le support représente la fréquence de la règle ou la portée de la règle. Par extrapolation, on utilise la probabilité que la règle soit présente que l'on nomme support relatif par rapport au support absolu qui correspond aux nombres d'occurrences. La règle $X \Rightarrow Y$ a donc un support s si $s\%$ des transactions de D contiennent $X \cup Y$ également note par XY simplification d'écriture. Autrement dit, le support correspond à la probabilité que la prémisse et la conclusion Y soient vraies. Par $D_{X \cup Y}$, nous indiquons l'ensemble de toutes les transactions qui contiennent $X \cup Y$, $D_{X \cup Y} = \{t \in D \mid X \cup Y \subseteq t\}$. Le support s de $X \Rightarrow Y$ est calculé comme $s = sup\ (X \Rightarrow Y) = sup(X \cup Y) = sup(XY) = P(XY) = |DX \cup Y| // n$. Un motif X qui respecte le support minimum (i.e. $sup(X) > MinSup$) est dit fréquent.

Définition 3.5 (*Confiance et confiance minimum : MinConf*) :

La confiance représente la force de la règle. La règle $X \Rightarrow Y$ a une confiance c si $c\%$ des transactions de D qui contiennent X contiennent également Y. Autrement dit, la confiance est la probabilité conditionnelle que la conclusion Y soit vraie sachant que la prémisse X est vraie : c'est-à-dire $P(Y|X)$. La confiance de $X \Rightarrow Y$ est calculée comme c $= Conf(X \Rightarrow Y) = Sup(XY)/sup(X)$. Il existe également un seuil pour la confiance minimum *MinConf* afin

de permettre aux utilisateurs de ne sélectionner que les règles plus pertinentes (*i.e. conf(X ⇒ Y) >MinConf*).

Définition 3.6 (*Règles d'association valide*) : Une règle d'association est dite valide si ses valeurs pour le support et pour la confiance sont supérieures aux seuils minimaux fixés par l'utilisateur : *MinSup* et *MinConf.*

Le problème qui nous intéresse est que : étant donné *une base de données transactionnelle T* et *un seuil de support minimal (MinSup), un seuil de support minimal* (*MinConf.*); Comment extraire tous les itemsets fréquents de T?

C'est pourquoi la recherche d'algorithmes efficaces de telles règles a été un problème majeur de cette communauté. Nous allons maintenant expliquer l'algorithme Apriori qui peut être employé pour l'extraction efficace des règles d'associations.

3.5.2 Algorithme Apriori

Dans cette section, nous allons expliquer l'algorithme que nous avons adopté. Pour les références bibliographiques, le lecteur est invité à consulter [Agrawal et al., 1993] et [Agrawal and Srikant, 1994)].

L'algorithme Apriori se décompose en deux phases à savoir : 1. Générer tous les motifs fréquents ; 2. Générer toutes les règles d'associations valides à partir des motifs fréquents. Des optimisations vont être apportées dans les deux parties de l'algorithme, à savoir dans la recherche des motifs fréquents et dans la recherche des règles. Commençons par analyser ces optimisations et pour cela, étudions les méthodes utilisées dans les deux étapes du processus. Nous déroulerons ensuite Apriori.

- (1) **Génération des motifs fréquents** : L'algorithme Apriori utilise une propriété du support qui va permettre de ne pas parcourir tout l'espace de recherches et par

conséquent va accélérer le processus d'extraction des motifs fréquents. Le support est une mesure anti-monotone.

Définition 3.5 (Mesure anti-monotone) : Une mesure M est dite anti-monotone si et seulement si : $\forall X,\ Y \subseteq I,$ *si* X *(*Y *et* M *(*Y*)* alors $M(X)$.

Autrement dit, une mesure est anti-monotone si lorsqu'elle est vérifiée pour un motif, elle est forcément vérifiée pour un sous-ensemble englobant ce motif. Il existe également des mesures monotones comme nous le verrons dans la suite.

Définition 3.5 (Mesure monotone) : Une *mesure M est dite monotone si et seulement si* : $\forall X,\ Y \subseteq I,$ *si* X *(*Y *et* $M(X)$ *alors* M *(*Y*).*

Autrement dit, une mesure est monotone si lorsqu'elle est vérifiée pour un motif, elle est forcément vérifiée pour un sur-ensemble englobant ce motif. Cette propriété définit donc que le support de tout sur-ensemble Y d'un motif X est inférieur ou égal au support de X, c'est-à-dire que $\forall Y, X, sup(Y) \leq sup(X)$.

Par conséquent, tous les sur-ensembles d'un motif non fréquent sont non fréquents. Par exemple, si c'est un motif non fréquent, aucun sur-ensemble de C ne peut être fréquent comme par exemple AC ou BC. Cette propriété va permettre d´élaguer un *k-motif* lorsqu'au moins un de ses sous-ensembles de taille *(k-1)* n'est pas fréquent. La figure 3.1représente le treillis de l'ensemble des motifs à étudier si C n'est pas fréquent.

La vérification du support de C nous permet d'élaguer 7 motifs dans l'étude et qui sont les suivants : $\{AC, BC, CD, ABC, ACD, BCD, ABCD\}$. Cette propriété va donc avoir une incidence sur l'ordre dans lequel on génère les motifs. Pour éviter la vérification du support sur un maximum de motifs, il faut donc générer les motifs par ordre croissant de taille.

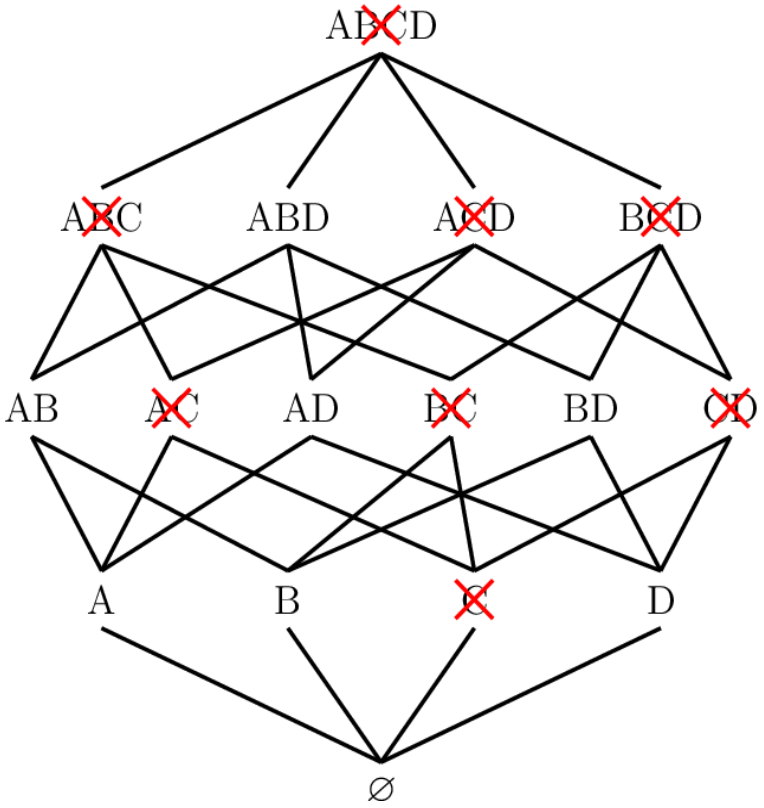

Figure 3.4 : Treillis des motifs potentiellement fréquents quand C n'est pas fréquent

Nous abordons maintenant le fonctionnement de l'approche Apriori. Toutefois, précisons tout d'abord le prérequis que nécessite cette méthode : les motifs doivent être écrits dans l'ordre lexicographique afin d'éviter de considérer plusieurs fois le même motif. AB et BA représentent donc le même motif en dehors des situations avec référence temporelle.

Nous présentons le processus algorithmique de génération des motifs fréquents dans la suite de cette section.

Tableau 3.3 : *Algorithme 1* - Génération de motifs fréquents

	Entrées : Base de données Transactionnelle : D, support minimum : $MinSup$ Sortie : Ensemble des motifs fréquents
1.	$F = \emptyset$
2.	$C_1 = \{i \in I\}$
3.	Pour $(k = 1; C_k \neq \emptyset; k++)$
4.	$F_k = \emptyset$
5.	Pour tout motif candidat $X \in C_k$ Faire
6.	*s= support (D,X)*
7.	Si $s \geq MinSup$ Alors
8.	$F_k = F_k \cup \{X\}$
9.	FinSi
10.	FinPour
11.	$C_{k+1} = \text{candidats}\ (F_k)$
12.	$F = F \cup F_k$
13.	FinPour
14.	Retourner F

Processus algorithmique du tableau 3.3 :

Cet algorithme commence (ligne 1) par initialiser l'ensemble C_1 des 1-motifs candidats à l'ensemble de tous les items i de la base de données D passée en paramètre. Le processus suivant (lignes 2 à 11) va être réitéré jusqu' à ce que l'on n'obtienne plus de candidat ($C_k = \emptyset$) à partir de l'ensemble F_{k-1} des motifs fréquents X de niveau inférieur (ligne 8). En effet, la génération des candidats effectuée par la fonction candidats (Voir l'algorithme 2), repose sur la propriété anti-monotone du support et va donc s'effectuer uniquement à partir des motifs fréquents afin d'éviter de vérifier le support de certains candidats que l'on sait par avance trop faible. Pour un niveau k donné, on commence par initialiser l'ensemble des k-motifs fréquents F_k à l'ensemble vide (ligne 3). Ensuite pour tous les candidats X de C_k (ligne 4), on calcule le support avec la fonction support (ligne 5) qui interroge la base de données. Puis, si le motif X est fréquent (ligne 6), on le stocke dans l'ensemble F_k qui servira,

une fois l'ensemble des motifs X de C_k parcouru, à générer les candidats de niveau supérieur (ligne 8). La dernière étape (ligne 11) consiste à ajouter l'ensemble des k-motifs fréquents F_k à l'ensemble des motifs fréquents F. Lorsqu'il n'existe plus de candidats à analyser, on retourne l'ensemble des fréquents F (ligne 12).

Maintenant, nous exposons sur la fonction candidats, correspondant à l'algorithme 2, permettant de générer le prochain niveau de candidats.

Processus algorithmique du tableau 3.4 :

Cette fonction va prendre en paramètre l'ensemble des k-motifs fréquents F_k et va retourner les candidats à analyser à la prochaine itération. L'algorithme commence par générer l'ensemble C_{k+1} des candidats potentiels (ligne 1) en combinant l'ensemble F_k avec lui-même. Deux *k*-motifs peuvent être combinés pour créer un nouvel *(k+1)-motif* candidat si et seulement s'ils ont en commun les (*k-1*) premiers items.

Tableau 3.4 : *Algorithme 2* - Génération des (*k+1*) - motifs candidats

	Entrées : Ensemble des k-motifs fréquents :F_k
	Sortie : Ensemble des (*k+1*)- motifs candidats : C_{k+1}
1.	$C_{k+1} = F_k \bowtie F_k$
2.	Pour tout motif candidat potentiel $X \in C_{k+1}$Faire
3.	Pour tout k-motif $Y \subset X$ Faire
4.	Si$Y \notin F_k$Alors
5.	$C_{k+1} = C_{k+1} \setminus X$
6.	FinSi
7.	FinPour
8.	FinPour
9.	Retourner C_{k+1}

Par exemple, en se référant à la figure 3.4, AB et AC ont leur premier item A en commun et peuvent donc se combiner pour créer le motif ABC. Cependant, les motifs AC et BC ne peuvent pas se combiner car même si C est commun, il n'est pas le premier item. La seconde étape (lignes 2 à 5) est l'élagage des candidats dont tous les sous-ensembles ne sont pas fréquents. Pour se faire, il faut vérifier pour chaque ($k+1$)-motif X du nouvel ensemble C_{k+1} (ligne 2), que chaque sous-motif Y de taille k (ligne 3) composant X est bien fréquent. Si le sous-motif Y n'est pas fréquent (ligne 4) alors le motif X analysé doit être retiré de l'ensemble C_{k+1} (ligne 5). Une fois que tous les motifs X sont analysés, on retourne l'ensemble des ($k+1$)-motifs candidatsC_{k+1} (ligne 6).

La première étape du processus algorithmique étant finie, nous passons à la seconde étape qui consiste à générer les règles à partir des motifs fréquents dans la première phase de l'algorithme (*Algorithme 1 et Algorithme 2*).

- (2) *Génération des règles valides* : Pour la génération des règles, on utilise une propriété de la confiance pour élaguer certaines règles sans avoir à calculer leur confiance.

Propriété 3.1 (Propriété de la confiance) : Cette propriété est définie comme suit :
$\forall (X, Y, Z)$tel que :$Z \subsetneq Y \subsetneq X$,$\mathrm{Conf}(Z \Rightarrow X \setminus Z) \leq \mathrm{Conf}(Y \Rightarrow X \setminus Y)$ *Conf*(Z X\Z).

Preuve :

$$\mathrm{Conf}(Z \Rightarrow X \setminus Z) \leq \mathrm{Conf}(Y \Rightarrow X \setminus Y)$$
$$\Leftrightarrow \frac{\mathrm{Sup}(Z \Rightarrow X \setminus Z)}{Sup\ (Z)} \leq \frac{\mathrm{Sup}(Y \Rightarrow X \setminus Y)}{Sup\ (Y)}$$
$$\Leftrightarrow \frac{\mathrm{Sup}(X)}{Sup\ (Z)} \leq \frac{\mathrm{Sup}(X)}{Sup\ (Y)}\ , \text{puisque } Z \subsetneq Y\ .$$

Cette propriété de la confiance induit deux situations :

1. Si la confiance de la règle $Z \Rightarrow X|Z$ n'est pas valide alors la confiance de la règle $Y \Rightarrow X|Y$ ne le sera pas non plus.

2. La proposition contraposée, si la confiance de la règle $Z \Rightarrow X|Z$ est valide alors la confiance de la règle $Y \Rightarrow X|Y$ le sera également.

Cette propriété va avoir une incidence sur l'ordre d'études des différentes règles pour un même motif. Pour éviter la vérification de la confiance sur un maximum de règles, il faut commencer par analyser les règles qui possèdent un motif en conclusion le plus petit possible. Regardons maintenant comment ces optimisations ont été utilisées dans l'algorithme Apriori. Nous passons à la seconde phase de l'algorithme dite de ''*génération des règles* ''. Pour générer l'ensemble des règles d'association, C'est l'algorithme nommé : *algorithme de génération des règles* (voir Algorithme 3) qui recherche les règles d'associations possédant un seul item, puis fait appel à un autre algorithme nommé *algorithme des autres règles* (voir algorithme 4) pour générer les autres règles, c'est-à-dire celles possédant plusieurs items.

Processus algorithmique du tableau 3.5 :

L'algorithme de génération des règles (voir tableau 3.5) prend en paramètre l'ensemble des fréquents F ainsi que le seuil minimum de la confiance *MinConf*. Cet algorithme se déroule en deux phases. La première phase (lignes 1 à 9) consiste à générer, à partir des motifs fréquents les règles possédant un seul item en conclusion. On récupère dans l'ensemble E_1 tous les items qui composent le motif X en cours d'étude (ligne 3). Ensuite, pour chaque item Y de l'ensemble E_1 (ligne 4), on calcule la confiance de la règle $X|Y \Rightarrow Y$ (ligne 5). Si la confiance de la règle est supérieure ou égale au seuil *MinConf* (ligne 6), alors la règle est ajoutée à l'ensemble des règles dites valides R (ligne 7).

Tableau 3.5 : *Algorithme 3* - Génération des règles d'associations

	Entrées : Ensemble des motifs fréquents F, support minimum : *MinConf*
	Sortie : Ensemble des règles d'associations valides R
1.	$R = \emptyset$
2.	Pour tout k-motif $X \in F$ tel que $k > 1$Faire
3.	$E_1 =$ *ensemble des* $1 - motifs \subset X$
4.	Pour tout $Y \in E_1$ Faire
5.	c= $\mathsf{Conf}(Y \Rightarrow X \setminus Y)$
6.	Si c $\geq$ MinConfalors
7.	$R = R \cup \{X/Y \Longrightarrow Y\}$
8.	Sinon
9.	$E_1 = E_1/Y$
10.	FinSi
11.	FinPour
12.	$R = R \cup$ *AutresRègles* (X, E_1, MinConf)
13.	Retourner R

Si la règle n'est pas valide (ligne 8), le motif Y va être retiré de l'ensemble E_1 (ligne 9). Une fois tous les items Y parcourus, l'ensemble des règles possibles possédant un seul item en conclusion sera généré pour un motif donné et E_1 contiendra uniquement les conclusions qui ont permis de générer les règles valides. Ce nouvel ensemble E_1 sera utilisé dans la fonction autresRègles (ligne 10) pour générer les autres règles, c'est-à-dire celles possédant plusieurs items en conclusion. La fonction « *autresRègles* » repose sur la propriété anti-monotone du support et va nous éviter de calculer la confiance de certaines règles que l'on sait par avance trop faible.

Dans la dernière partie de ce processus algorithmique, nous présentons la fonction récursive nommée '' *AutresRègles* '' qui met en évidence d'autres règles d'associations de plus d'un item.

Tableau 3.6 : *Algorithme 4* - Génération des règles d'associations de plus d'un item

	Entrées : *k*-motif fréquent, ensemble des m-motifs en conclusion : E_m Confiance minimum : *MinConf*
	Sortie : Ensemble des règles d'associations valides R' possédant un (*m+1*) motifs en conclusion pour le motif *X*
1.	$R = \emptyset$
2.	Si$k > m + 1$ Alors
3.	E_{m+1} = Candidat (E_m)
4.	Pour tout Y $\in E_{m+1}$ Faire
5.	c= $\mathrm{Conf}(X \setminus Y \Rightarrow Y)$
6.	Si c≥ $MinConf$ Alors
7.	$R' = R' \cup \{X \setminus Y \Rightarrow Y\}$
8.	Sinon
9.	$E_{m+1} = E_{m+1} \setminus Y$
10.	FinSi
11.	FinPour
12.	C_{k+1} = candidats (F_k)
13.	$R' = R' \cup$ AutresRègles $(X, E_{m+1}, MinConf)$
14.	FinSi
15.	Retourner *R'*

Déroulement de la fonction « *AutresRègles ()* » : La fonction récursive « *autresRègles* » (Voir algorithme 4) va retourner l'ensemble des règles valides possédant plusieurs items en conclusion pour chaque motif *X* passé en paramètre. Le second paramètre de la fonction est l'ensemble $\mathbf{E_m}$des m-motifs conclusion Y pour lesquels la règle X\Y ⇒ Y est valide.

En effet, l'algorithme commence par vérifier s'il est possible de générer d'autres règles à partir du motif *X* (ligne 2). En effet, il faut vérifier que la taille *k* du motif *X* passé en

paramètre est strictement supérieure aux tailles (*m+1*) des futures conclusions. Puis, on appelle la fonction candidats (ligne 3) afin de générer les conclusions de taille (*m+1*) à partir des conclusions de taille m qui ont mené à des règles valides à l'itération précédente. Ce nouvel ensemble de conclusions est stocké dans l'ensemble $\mathtt{E_{m+1}}$. Ensuite, pour chaque item Y de l'ensemble $\mathtt{E_{m+1}}$(ligne 4) on calcule la confiance de la règle X\Y ⇒ Y (ligne 5). Si la confiance de la règle est supérieure ou égale au seuil MinConf (ligne 6) alors la règle est ajoutée à l'ensemble des règles valides R' (ligne 7). Si la règle n'est pas valide (ligne 8), le motif Y va être retiré de l'ensemble $\mathtt{E_{m+1}}$(ligne 9). Une fois tous les items Y parcourus, l'ensemble des règles possibles possédant (m+1) items en conclusion est généré pour un motif X donnée et $\mathtt{E_{m+1}}$ contiendra uniquement les conclusions qui ont permis de générer les règles valides. Ce nouvel ensemble $\mathtt{E_{m+1}}$ sera à son tour utilisé dans la fonction récursive autresRègles (ligne 13) pour générer les règles possédant (*m+2*) items en conclusion.

Une fois les règles d'associations générées, nous déterminons celles qui sont valides en introduisant un nouvel indicateur de mesure. Dans la section suivante, nous présentons les règles d'association dite ''*valides*'' (Le Lift) dans le but afin de sélectionner les RA qui fournissent des informations pertinentes étant donné que, nous voulons dans cette contribution découvrir des connaissances pour une prédiction et ciblage. Dans la section suivante, nous présentons les règles valides.

3.5.3 Règles valides

L'extraction de règles d'associations est l'une des techniques les plus populaires de la fouille de données. Ce problème surnommé analyse du panier de la ménagère a été introduit pour la première fois en 1993 [Agrawal et al., 1993] pour analyser des bases de données de la grande distribution. Depuis lors, ce problème a été intensément étudié pour son utilité dans de nombreux domaines d'applications tels que les systèmes de recommandations, la bio-

informatique ou encore les diagnostics médicaux [Agrawal et Srikant, 1994b], [Agrawal et Srikant, 1994c]. Dans ces règles, la cible n'est pas prédéfinie. La partie droite peut être une conjonction de conditions d'attributs. Alors, elle peut être considérée comme une propriété probabiliste relative à la co-occurrence d'événements qui satisfont des contraintes statistiques sur la base de données comme *un support* et *une confiance* minimale.

En termes de probabilité, le *support (supp)* de la règle $X \rightarrow Y$ représente $p(X \cap Y)$ et la *confiance (Conf)* représente la probabilité conditionnelle $p(Y/X)$ où par abus d'écriture, on note respectivement X et Y, l'ensemble des exemples vérifiant l'antécédent X et le conséquent Y. Une règle peut avoir d'excellents supports et confiance sans être pour autant « intéressante » ; Dans ce cas, il nous faut un critère afin de limiter la prolifération des *règles (Car s'il y'a* m *items, il y'a* $\sum_{k=2}^{m} \binom{m}{k} (2^k - 2)$,règles associatives possibles). C'est dans cette optique que nous introduisons un nouveau paramètre qui est un indicateur de pertinence des règles associatives : le Lift qui est une mesure de performance de la règle d'association en vérifiant si les résultats obtenus ne sont pas le fruit du hasard [Le Bras et al., 2011]. Son interprétation est la suivante : Si la mesure est supérieure à 1, cela indique une corrélation positive : la règle est considérée comme intéressante ; Si la mesure vaut 1, sa corrélation est nulle, la mesure dans ce cas ne sert à rien et Si la mesure est inférieure à 1, la corrélation est négative.

Le calcul du lift est défini comme suit : $Lift = Conf\ (X \rightarrow Y)|N$

Dans ce manuscrit de thèse, les contraintes qu'une règle d'associations $(X \rightarrow Y)$ doit respecter afin d'être considérée comme valide sont décrites dans le tableau 3.7.

Tableau 3.7: Règles d'associations valides

$(X \rightarrow Y)$
$Supp(X \rightarrow Y) \geq MinSup$
$Conf(X \rightarrow Y)) \geq MinConf$
$Lift = Conf\ (X \rightarrow Y)\|N > 1$

3.6 Modélisation de la cartographie du risque douanier

La gestion des risques ne constitue pas pour les douanes un concept nouveau. En puisant dans les renseignements, les informations et les expériences, la douane a adopté au fil du temps des procédures destinées à lutter contre le non-respect ou le contournement de la législation douanière. Les méthodes traditionnelles de contrôle douanier, comprenant une intervention à 100 % ou la sélection de pourcentages élevés d'importations ou des critères de sélection purement aléatoire, ne constituent pas les meilleurs modèles en matière de gestion des frontières et ne répondent pas aux attentes actuelles de l'administration douanière, de la communauté et des entreprises.

Ainsi l'une des tâches principales de la douane consiste à évaluer les risques inhérents à la circulation des marchandises. On entend par « risques » les facteurs qui pourraient avoir une répercussion sur les objectifs douaniers. Afin de réaliser ces objectifs, il est important de bien connaître les risques encourus ainsi que l'impact potentiel de ces risques sur lesdits objectifs. Les administrations douanières doivent choisir d'aborder le contrôle des opérateurs économiques sous l'angle de la gestion des risques. L'idée est d'axer les activités de contrôle douanier sur les risques plutôt que sur des éléments ou des déclarations sélectionnés de manière aléatoire.

La gestion des risques devrait être considérée comme un complément indispensable de la gestion stratégique de la lutte contre les infractions dans le cadre des opérations de dédouanement. Elle devrait également être pleinement intégrée dans le processus de dédouanement pour constituer le fondement permettant d'anticiper sur les risques éventuels.

3.6.1 Modèle mathématique de la cartographie des risques

nous présentons dans cette section du mémoire de thèse, un modèle mathématique qui permet de réaliser la cartographie des risques douaniers nécessaire à une analyse optimale de ces risques.

3.6.1.1 Définition et notations des différentes variables binaires en 0-1 et les ensembles

Pour modéliser la cartographie des risques douaniers aux fins de leur analyse et gestion, les hypothèses suivantes sont à considérer :

Soient :

- n le nombre d'opérateurs économiques qui effectuent n opérations de dédouanement
- $I= \{1, 2, 3,\ldots, n\}$, l'ensemble des opérateurs connus des services douaniers
- $J= \{1, 2, 3,\ldots, n\}$, l'ensemble des risques liés aux marchandises dans le cadre des opérations de dédouanement
- $K= \{1, 2, 3,\ldots, m\}$, l'ensemble des risques internes à la douane liés au management
- $M= \{1, 2, 3,\ldots, k\}$, l'ensemble des risques aléatoires
- x_{ij} : variable binaire associée aux risques encourue sur les marchandises lors d'une opération de dédouanement
- $x_{ij} = \begin{cases} 1: \text{si une infraction est constactée au cours d'une opération de dedouanement} \\ 0 : \text{ sinon} \end{cases}$
- y_k : variable binaire associée aux risques internes à la douane liés au management
- $y_k = \begin{cases} 1: \text{si une infraction est constactée lors d'un management en interne} \\ 0 : \text{ sinon} \end{cases}$
- γ_k : le coût associé au management interne dans l'administration douanière selon l'importance du risque interne

$$\gamma_k = \begin{cases} \text{si k} = 1, \text{ risques internes liés à la procédure} \\ \text{si k} = 2, \text{ risques internes liés à la reglementation} \\ \text{si k} = 3, \text{ risque internes liés à l'encadrement} \end{cases}$$

- β_{ij} : le coût associé à l'affectation entre un opérateur et une infraction selon l'importance du risque liée à la marchandise
- ε_m : Variable binaire de bouclage du modèle liée à la survenance d'un risque aléatoire
- $\varepsilon_m = \begin{cases} 1 : \text{si il y'survenance de risque aléatoire} \\ 0 : \text{sinon} \end{cases}$

3.6.1.2 Fonction critère

La fonction critère est la fonction-objectif ou encore la fonction définissant les différents risques. Elle est la somme composée de trois coûts en terme du risque, qui sont :

- La fonction du risque liée aux marchandises dans le cadre des opérations de dédouanement : f_1
- La fonction du risque liée aux management interne de l'administration douanière : f_2
- La fonction liée aux variables de bouclage en cas de survenance du risque aléatoire : f_3
  - *La fonction du risque liée aux marchandises dans le cadre opérations de dédouanement*

    Cette fonction traduit le risque encouru au cours d'une opération de dédouanement. Elle est modélisée par une affectation entre un opérateur et une marchandise pondérée par une valeur (coût) nommé par la variable β_{ij}, et qui représente la relation entre un opérateur et une infraction selon l'importance du risque lié à la marchandise.

$$f_1 = \sum_{i \in I} \sum_{j \in J} \beta_{ij} x_{ij} \qquad (3.1)$$

  - *La fonction du risque liée aux management interne de l'administration douanière*

    Cette fonction représente le risque interne lié à la gestion interne de l'administration douanière. C'est un modèle linéaire avec des variables binaires dont les valeurs sont restreintes à l'intervalle discret [0;1].

$$f_2 = \sum_{k \in K} \gamma_k y_k \tag{3.2}$$

- *La fonction de survenance du risque aléatoire*

C'est une fonction de bouclage du modèle avec une variable binaire. Elle représente la fonction du risque aléatoire.

$$f_3 = \sum_{m \in M} \varepsilon_m \tag{3.3}$$

3.6.1.3 Contraintes du problème de la cartographie des risques

- Contraintes des affectations
  - Un opérateur économique commet au moins une infraction au cours d'une opération de dédouanement

$$\sum_{j \in I} x_{ij} \geq 1\,, \quad i = 1, \dots, n \tag{3.4}$$

  - Une infraction est commise par un et un seul opérateur économique au cours d'une opération de dédouanement

$$\sum_{i \in I} x_{ij} = 1, j = 1, \dots, n \tag{3.5}$$

- Contrainte de non-négativité
  - Propension du risque lié aux dédouanement de marchandises

$$\beta_{ij} \in \mathbb{R}^+ ; \qquad \forall\, i \in I \;;\; \forall j \in J \tag{3.6}$$

$$\sum \beta_{ij} = 1 \qquad \forall\, i \in I\,, \forall\, j \in J \tag{3.7}$$

  - Propension du risque interne lié aux managements

$$\gamma_k \in \mathbb{R}^+ ; \qquad \forall\, k \in K \tag{3.8}$$

$$\sum \gamma_k = 1 \qquad \forall\, k \in K \tag{3.9}$$

- **Définition et borne des variables décisionnelles**
  - Variable binaire associée aux risques encourue sur les marchandises lors d'une opération de dédouanement

$$x_{ij} \in \{0,1\}; \qquad \forall\, i \in I \text{ et } \forall\, j \in J \tag{3.10}$$

- Variable binaire associée aux risques internes à la douane liés au management

$$y_k \in \{0,1\}; \qquad \forall\, k \in K \tag{3.11}$$

- Variable binaire associée à la survenance du risque aléatoire

$$\varepsilon_m \in \{0,1\}, \qquad \forall\, m \in M \tag{3.12}$$

3.6.1.4 Modèle mathématique de la cartographie

Le modèle mathématique de la cartographie des risques douaniers est la combinaison des différentes fonctions de risques ($\sum f_i = f_1 + f_2 + f_3$) dont l'objectif est d'avoir le contour des différents risques liés à l'activité douanière afin d'y apporter des éléments de mitigation.

$$F_{(risque)} = \sum_{i\in I}\sum_{j\in J}\beta_{ij}x_{ij} + \sum_{k\in K}\gamma_k y_k + \sum_{m\in M}\varepsilon_m \tag{3.13}$$

s.c

$$\sum_{j\in J} x_{ij} = 1, \qquad i = 1,\dots,n \tag{3.14}$$

$$\sum_{i\in I} x_{ij} \geq 1\,, \qquad j = 1,\dots,n \tag{3.15}$$

$$\sum \beta_{ij} = 1 \qquad \forall\, i \in I\,, \forall\, j \in J \tag{3.16}$$

$$\sum \gamma_k = 1 \qquad \forall\, k \in K \tag{3.17}$$

$$x_{ij} = \{0;1\} \qquad i = 1,\dots,n\,;\; j = 1,\dots,n \tag{3.18}$$

$$y_k = \{0,1\} \qquad k = 1,\dots,m \tag{3.19}$$

$$\varepsilon_m = \{0;1\} \qquad m = 1,\dots,p \tag{3.20}$$

La relation (3.13) représente le modèle de la cartographie des risques en lien avec :

- Les risques liés au dédouanement de marchandises et aux opérateurs ;
- Les risques internes liés au management dans l'administration douanière ;
- Les risques aléatoires.

La relation (3.14) assure qu'une infraction est commise par un et un seul opérateur économique. La relation (3.15) stipule qu'un opérateur peut commettre au moins une infraction. Les relations (3.16) et (3.17) représentent les propensions des risques respectivement liés aux marchandises dans le cadre du dédouanement et au management interne de l'administration. Enfin, les relations de (3.18) à (3.20) sont les variables décisionnelles binaires intervenant dans la fonction cartographie des risques liés aux activités douanières.

Ainsi, la méthode de la cartographie des risques est un outil essentiel pour l'évaluation du profil d'un opérateur au cordon douanier. En effet, la douane peut aussi gagner à utiliser la méthode de cartographie des risques de façon systématique pour apprécier les risques. Une approche systématique est de plus nécessaire pour déterminer comment un opérateur agréé doit être contrôlé et évalué par la suite. La méthode a pour objet de classer les risques par ordre de priorité en évaluant leur probabilité et leur impact sur les objectifs douaniers. Elle permet de structurer et de faciliter l'appréciation des risques. En combinant la cartographie des risques et les mesures prévues, on obtient une approche structurée ciblée sur le recensement des risques, l'évaluation des risques, le contrôle et l'évaluation à des fins d'amélioration continue. Le processus de cartographie des risques se déroule en cinq étapes : comprendre les activités d'un opérateur (1), clarifier les objectifs douaniers (2), recenser les risques qui sont susceptibles d'avoir une répercussion sur les objectifs douaniers (3),

évaluer la significativité des risques (4) et répondre aux risques ; (5) comment prendre en charge les risques.

Dans le cadre de la gestion des risques, la douane tient compte des mesures que les opérateurs eux-mêmes ont prises pour prévenir les risques dans leurs processus d'entreprise. Les administrations douanières veulent concentrer leurs capacités limitées sur les risques qui ne sont pas couverts, ou qui ne le sont pas suffisamment, par les mesures adoptées par les opérateurs économiques. Afin de pouvoir suivre cette approche, il est nécessaire de bien connaître l'opérateur économique, ses processus d'entreprise et les mesures qu'il a prises pour réduire les risques liés aux processus fiscaux et aux processus non fiscaux (chaîne d'approvisionnement). La douane doit par conséquent évaluer l'organisation, les processus, les procédures, l'administration, de l'opérateur économique. Autrement dit, c'est l'organisation administrative de l'opérateur économique et son système de contrôle interne qui doivent être évalués.

Afin de déterminer le traitement douanier réservé à une marchandise lors d'opérations d'importation ou d'exportation, il existe trois notions essentielles qu'il faut connaître et maîtriser : *l'espèce tarifaire, l'origine et la valeur en douane*. Ces informations seront à transmettre aux autorités douanières lors de chaque passage en douane, via la déclaration douanière. C'est donc en anticipant et en maîtrisant ces données que l'on peut supprimer le risque douanier, et ainsi organiser et gérer les flux à l'international au mieux des intérêts de l'entreprise.

Afin que l'approche utilisée pour la cartographie des risques soit la plus simple possible et, partant, la plus efficace possible, il est souhaitable de la diviser en deux phases : tout d'abord, la phase de cartographie interne (exécutée par la douane), suivie de la phase de cartographie

commune (avec l'opérateur), au terme de laquelle la douane doit localiser les risques et décider de la manière d'y répondre.

3.7 Espace des données à explorer

Nous allons présenter dans cette section la base de données d'où sont issues les données à partir desquelles l'extraction de connaissances a été effectuée pour l'analyse de comportements potentiellement à risques dans l'objectif de tester et valider notre méthodologie.

L'interprétation des connaissances générées à l'aide des règles d'associations à partir des d'indicateurs tels : *le support, la confiance et le lift* permettent de qualifier un comportement comme étant à risque.

3.7.1 Données des infractions douanières issues du Procès-Verbal Simplifié du Système de dédouanement de la république de Côte d'Ivoire

Nous avons fait l'acquisition des données d'infractions douanières recensant les infractions sur la période de 2015 à 2018. Ces données ont été extraites du PVS (SYDAM World).

Ces données recensent les infractions ou fraudes douanières constatées lors des opérations de dédouanement sur toute l'étendue du territoire de Côte d'Ivoire. Cette masse de données d'une taille de 10.3 Mo, contient 6854 cas de fraude sur la période indiquée plus haut. Dans notre étude nous nous sommes limités qu'aux opérations de dédouanement où des cas d'infractions sont constatés. Le modèle physique de données du PVS, est représenté sur la Figure 3.5

Nous présentons en annexe du mémoire de thèse, le dictionnaire des données ainsi que les différentes tables pour aider le lecteur dans la compréhension et la signification des différentes données présentées à la figure 3.5

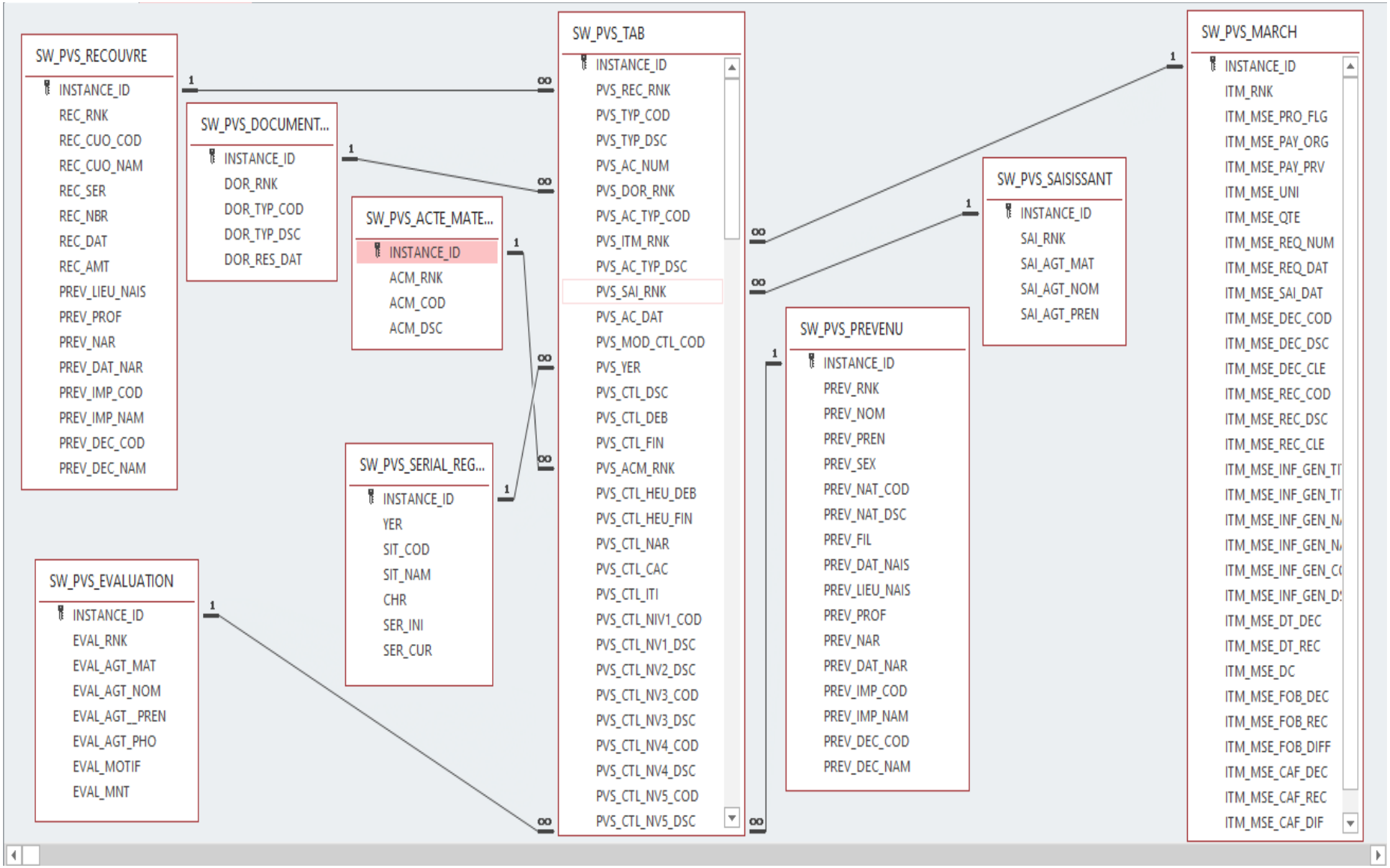

Figure 3.5 : Modèle conceptuel de la base de données du Procès-Verbal Simplifié (PVS)

3.7.2 Acquisition de données : Nettoyage

Une phase d'acquisition des données à explorer est nécessaire dans toute approche de fouille de données. De nos jours, d'importantes bases de données sont constituées à partir de flux continus d'informations ou de mises à jour régulières. Ces données peuvent contenir des bruits, des données manquantes, des incohérences ou des imprécisions.

Contrairement à l'analyse statistique, les données utilisées (En datamining) sont souvent créées à d'autres fins, elles sont donc souvent inexploitables directement par la fouille de données. Pour une exploitation de ces données, il faut un nettoyage des données ou (Cleaning data en anglais, voir la figure 3.3) car la qualité des connaissances extraites par fouille de données dépend beaucoup de la qualité et de la quantité des données en entrée. En effet, plus il y a de données (cas observés) meilleure est la précision des connaissances. Une analyse statistique des valeurs d'attributs de ces données est nécessaire étant donné que la qualité des résultats d'analyses exploratoires dépend généralement plus de la préparation des données que de la méthode exploratoire utilisée. La première étape de toute investigation dans les données est le calcul des statistiques univariées pour connaître la distribution des variables et identifier les anomalies. Il est important de manipuler les incohérences et les valeurs manquantes avec intelligence car elles peuvent véhiculer des informations. Ainsi une étape de préparation de ces données est nécessaire pour permettre leur exploitation.

3.7.3 Sélection des données

Nous avons sélectionné du PVS les données qui décrivent les fraudes ou infractions (*Code de l'infraction, description de l'infraction, code nature de l'infraction, code titre de l'infraction, description de la nature de l'infraction, etc.*), les marchandises (*Code de la marchandises, provenance, description de la marchandise, droit et taxe déclarés de la marchandise, Droit et taxes reconnus, Valeur FOB déclarée, Valeur FOB reconnue, valeur CAF déclarée, valeur CAF reconnue, valeur sur le marché intérieur, droit de compromis*

etc.) et la description des Prévenus (appelés Opérateurs (code du prévenus, nom et prénom du prévenus, description, Filiation du prévenu etc.). La sélection des variables sur lesquelles va porter notre analyse va réduire le nombre de variables à considérer, le nombre de règles générées et ainsi faciliter l'interprétation des résultats. Cette sélection de données va constituer le contexte d'exploration sur lequel va porter l'extraction de règles d'association dans le but de trouver les relations d'implications entre les différents facteurs de situations.

Remarque :

(1) Si la valeur Fob déclarée par l'opérateur < à la valeur Fob reconnue par les services de douane, alors il y'a une infraction ou fraude constatée, dans un tel cas un droit compromis est imputé au prévenu coupable de la fraude.

(2) Selon le Modèle Relationnel de la Base de Données formelles des opérations de douane, et les différentes clés (code), à partir du code du procès-verbal simplifié (PVS), nous pouvons identifier la nature de l'infraction et le type de la marchandise sur laquelle la fraude a été identifiée ainsi que la description du prévenu (le fraudeur).

3.8 Expérimentation : *Mise en contexte et résultats*

Le problème d'extraction de règles d'associations à partir des données consiste à générer toutes les règles d'association valides. La découverte de ces règles va avoir différents objectifs en fonction des données analysées. Dans le cadre de notre thèse de doctorat, nous appliquons l'approche Apriori sur une masse se de données recensant des opérations douanières. L'objectif est de procéder à une analyse des risques douaniers à partir de l'exploration historique des données d'infractions dans le but de découvrir les connaissances régissant la survenue de la fraude. Cette masse de données souvent hétérogènes et complexes génère ainsi de nouveaux besoins auxquels les méthodes d'extraction de connaissances doivent

pouvoir répondre. L'appréciation des infractions passe inéluctablement par une bonne identification des risques

La méthode de recherche de motifs fréquents s'appuie sur la notion formelle de motif. Cette phase consiste à extraire du contexte tous les ensembles d'attributs binaires$m \subseteq A$, appellé itemsets, qui sont fréquents dans le contexte $\mathcal{B}$. Soit $\mathcal{B}$ une base de données décrivant un ensemble d'objets $O = \{o_1, o_2, \dots, o_N\}$ (*Les transactions),* par un ensemble A fini d'attributs$A = \{a_1, a_2, \dots, a_n\}$, appelé aussi Item. Pour identifier et sélectionner un item, nous considérons une relation $\mathcal{R}$ du type 0-1 (booléen) entre un objet O et un item a notée $o\mathcal{R}a \in \{0,1\}$. On notera la base de données, le triplet $\mathcal{B} = (O, A, \mathcal{R})$. Les transactions sont représentées par les opérations douanière (OD). Les Item sont les infractions relatives aux fraudes. Ainsi si une infraction est constatée sur une opération douanière, la relation $\mathcal{R}$ prend la valeur de 1 sinon 0. Par conséquent, la Base de Données est modélisée par une matrice booléenne ou les lignes et les colonnes correspondent respectivement aux objets et aux attributs spécifiquement les infractions (Voir le tableau 3.8).

Dans le tableau 3.8, l'infraction de type 3 est associée aux opérations douanières (OD) N°2 et N°3 ou encore que les Opérations Douanières (OD) ne comportent pas d'infraction de type *n-1.*

L'extraction des règles d'association va consister à extraire des règles sur la base de principaux paramètres : le Support et la Confiance.

Tableau 3.8 : Exemple représentatif d'une Base de données binaires

*O***R***A*	*Infraction Type 1*	*Infraction Type 2*	*Infraction Type 3*	…	*Infraction Type n-1*	*Infraction Type n*
OD_1	1	0	1	…	0	1
OD_2	0	1	1	…	0	0
…	…	…	…	…	…	…
OD_N	1	1	0	…	0	1

C'est une implication entre deux itemsets à laquelle sont associés le support, qui définit la portée de la règle, et la confiance, qui définit la précision de la règle dans le contexte d'extraction. Pour éliciter des règles associatives, nous recherchons des généralisations de motifs qui apparaissent fréquemment dans la masse de données en vue de trouver des régularités sous forme d'éléments associés.

Enfin, pour faciliter l'exploitation de ces règles de découvertes, nous les catégorisons en trois groupes :

- (a)*Règle de prévision* : Ce sont des règles utiles contenant des informations de qualité. L'antécédent est connu à priori contrairement à son conséquent. Dans ce cas la confiance de la règle est supérieure à plus de 50%.
- (b)*Règle de ciblage* : Ce sont des règles de connaissances générales qui identifient les relations entre les différentes attributs (motifs). L'antécédent et le conséquent de la règle sont connus mais pas la relation d'implication entre les deux parties.
- (c)*Règle neutre* : Ces règles ne donnent pas d'informations nouvelles

Remarque : Une règle dénote l'interaction entre deux évènements (A = Une opération de dédouanement et un risque de fraudes lié au dédouanement) où leurs actions sont généralement dépendantes, ce qui peut conduire à un risque de fraude. Nous nous sommes inspirés pour cela de l'application du panier de la ménagère [Agrawal et al., 1993].

3.8.1 Résultats et Analyses

Les expérimentations ont été menées sur une plate informatique Intel Core ™ i7-3540M 3.00 GHz avec 8 Go de RAM, sous Linux Ubuntu. L'algorithme Apriori a été implémenté dans le package 'Arules' du progiciel R. Le langage de programmation est Python via la bibliothèque PyFIM. Les mesures tiennent compte des index, de l'espace mémoire non utilisé et de celui qui est libéré lors de la suppression ou du déplacement des données.

- **Génération des règles d'associations**

Nous présentons dans cette section, un exemple de code informatique permettant l'extraction de règles avec les indicateurs Support, Confiance et Lift.

- *Pour obtenir les règles ayant un support d'au moins 20% et une confiance supérieure à 60%, il suffit de lancer la commande :*

rules< −apriori(Adult,parameter=list(support=0.2,confidence=0.6))

- *Si on choisit de se focaliser sur la prévision de règles ayant en membre droit l'item '' Fausse_declaration de valeur'' et trier par confiance :*

rules< −apriori (*Adult,parameter=list(support=0.2)*)

rules.Fausse_declaration de valeur < −subset(rules, subset = rhs %in% " Fausse_declaration de valeur ")

rules.Fausse_declaration de valeur < − sort(rules. Fausse_declaration de valeur, by = "confidence")

inspect(rules. Fausse_declaration de valeur)

- *Pour spécifier des propriétés des règles recherchées, on fait appel à la fonction* ***subset ().****On peut combiner des tests aussi dans l'appel de subset() avec la mesure d'intérêt Lift .*

Par exemple subset = rhs %in% " Fausse_declaration de valeur " & lift >1.5.

- ## Résultats des expérimentations

Les différents tests informatiques réalisés ont permis de mettre en évidence plusieurs règles intéressantes dans le tableau 3.9.

Tableau 3.9:Résultats des implémentations

N°	***Customs operation category***	***Infringement-type***	***Supp.***	***Conf.***	***Lift***
1.	Contrôle de change	Fuite de capitaux	0,10	0,61	1,07
2.	Dédouanement marchandises	Fausse déclaration de valeur	0,35	0,57	2,36
3.	Contrôle marchandises	Fausse déclaration d'origine	0,35	0,59	3,02
4.	Dédouanement marchandises	Détournement	0,14	0,41	1,5
5.	Dédouanement marchandise	Fausse déclaration d'espèces	0,35	0,55	1,8
6.	Dédouanement de marchandises	Sous-évaluation et fausse facturation	0,37	0,71	1,1
7.	Dédouanement de marchandises	Fausse désignation de marchandises	0,32	0,54	2,01
8.	Dédouanement de marchandises	Contrebande de marchandises	0,33	0,55	1,12
9.	Transactions non déclarées	Falsification d'un cachet douanier	0,19	0,28	2,13

10.	Dédouanement de marchandises	Substitution de marchandises dans le cadre du régime d'importation temporaire	0,37	0,49	1,04
11.	Dédouanement de marchandises	Contrebande de marchandises par fausse déclaration	0,18	0,65	3,59
12.	Contrôle marchandises	Fausse désignation de la qualité	0,34	0,58	2,39
13.	Contrôle_change	Fausse déclaration d'espèces	0,27	0,55	1,17
14.	Dédouanement marchandises	Faux étiquetage de marchandises	0,43	0,68	1,28
15.	Dédouanement marchandises	Liquidation fictive du régime de transit douanier	0,24	0,23	1,08
16.	Dédouanement marchandises	Importation sans déclaration	0,21	0,51	3,12
17.	Contrôle de marchandises	Falsification d'une facture	0,41	0,55	2,89
18.	Contrôle de marchandises	Fausse déclaration du nom de l'exportateur	0,38	0,41	1,92
19.	Transport illicite de devise	Altération du code des devises	0,22	0,59	2,17
20.	Contrôle de marchandises	Exportation sans déclaration	0,27	0,56	1,07
21.	Contrôle de marchandises	Abus des exonérations	0,33	0,52	1,1
22.	Dédouanement de véhicules volés	Altération du numéro de série	0,245	0,59	2,35

- Analyse :

A l'issue des expérimentations, nous avons obtenu des règles d'associations qui ont été regroupées en trois grandes catégories de règles conformément aux valeurs seuils (Support, confiance) indiquées comme hypothèses. Et le lift comme indicateur de validité des règles.

- **Les règles de prévision**

Les règles n°2 (*Dédouanement marchandises → Fausse déclaration de valeur)*, n°3 (contrôle marchandises → Fausse déclaration d'origine) et n° 5 (Dédouanement marchandises → Fausse déclaration d'espèces) sont des règles de qualité qui rentrent dans la catégorie de la règle de prévisions.

- **Interprétation de la règle n°2** : *Dédouanement marchandises → Fausses déclarations de valeur*

La règle n°2 nous renseigne que 57% des infractions constatées dans les opérations de dédouanement de marchandises proviennent essentiellement des fausses déclarations de valeur. Elles sont largement répandues et trouvent leur intérêt dans le souci permanent des opérateurs de payer le moins de droits possibles par la diminution de la valeur de la marchandise dans la mesure où la taxation est ad valorem et se calcule par conséquent sur la base des prix des marchandises.

On note une fréquence élevée de cette règle représentant 39% des données des infractions du PVS.

- **Interprétation de la règle n°3** : *contrôles marchandises → Fausses déclarations d'origine*

Cette règle stipule que si une infraction concerne une opération de contrôle de marchandises, alors dans 59% des cas il s'agit d'une fausse déclaration d'origine. La fausse déclaration

d'origine ou de provenance a pour but d'éluder une mesure de prohibition, de contourner une restriction ou une mesure de contrôle.

La prévision de cette règle est vérifiée à 16% des cas des infractions du PVS.

- **Interprétation de la règle n°5** : Dédouanement marchandises → Fausses déclarations d'espèces tarifaire

Cette troisième règle de prévision nous informe que les fausses déclarations d'espèces constituent 55% des infractions dans une opération liée au dédouanement de marchandises. Les fraudes sur les espèces sont le fait des opérateurs qui exploitent les difficultés rencontrées lors des contrôles pour les rendre encore plus complexes avec l'importation dans un même conteneur plusieurs articles ou produits.

Il s'agit de la fausse dénomination quant à la nature et aux conditions d'utilisation des marchandises ; les marchandises ont été déclarées totalement ou partiellement sous une rubrique autre que celle qui leur est assignée par le tarif. Cette règle constitue 15% des cas d'infractions du PVS. L'origine des marchandises, l'espèces tarifaire, et la valeur des marchandises sont des notions essentielles pour le traitement douanier à réservé à une marchandise tant à l'importation qu'à l'exportation. Les infractions liées à ces trois rubriques ont des conséquences économiques et fiscales :

- Perte de recettes douanières
- Évasion fiscale
- Statistiques erronées du commerce extérieur
- Concurrence déloyale

Ces trois règles fréquentes représentent 70% des infractions du PVS *(SYDAM World).* Cela dénote la forte propension des opérateurs à agir sur ces rubriques dans l'optique de minimiser les droits douaniers à acquitter à l'administration douanière.

Cette tendance significative de ces trois infractions dans le PVS indique par ailleurs un effort substantiel de la lutte contre la fraude par les douanes ivoiriennes en vue de juguler son impact négatif sur les recettes douanières.

Ces connaissances relevées sont prépondérantes à la mise en œuvre de politiques de prévision et d'anticipation efficaces pour la gestion de ces infractions :

- Contrôle anticipé de l'évaluation, de la classification des marchandises et des origines des marchandises.
- Renforcement des capacités en matière d'évaluation
- Sensibilisation des opérateurs économiques au civisme fiscal
- Conception des programmes de renforcement de capacité en matière de classification tarifaire et de règle d'origine

▪ **Les règles de ciblage** :

Les règles n°1 (*Contrôle de change → fuite de capitaux*) et n°19 (*Transport illicite de devise → Altération du code des devises*) sont des règles qui permettent de cibler une infraction spécifique, précisément ici le transport illicite de devises et justifie le fait que 61% des risques de fuite de capitaux sont essentiellement liés aux opérations de contrôle de change. Cette règle constitue 6% des cas d'infractions du PVS.

Les pays membres de la zone franc (UEMOA) doivent respecter une règlementation commune des opérations financières qu'ils réalisent avec des pays non-membres aussi bien à l'importation qu'à l'exportation.

En effet, le contrôle des changes vise à empêcher les achats excessifs de monnaies étrangères qui peuvent contribuer à la dépréciation de la monnaie d'un pays ou d'une zone franc. L'objectif est aussi de favoriser les importations considérées comme les plus utiles et les plus urgentes tout en protégeant les entreprises du pays. Le contrôle des changes est donc un

instrument de lutte contre la fuite des capitaux et spéculation, consistant plus particulièrement en des mesures prises par un gouvernement pour règlementer l'achat et la vente des monnaies étrangères par ses ressortissants.

Il ressort ainsi un suivi rigoureux des contrôles de change au niveau des douanes ivoiriennes. Pour tenir compte de l'impact négatif des fuites des capitaux sur les balances de paiement, il requiert une politique permanente de recherche et de répression aux fuites de capitaux.

- **Les règles neutres** :

La règle d'associations : *Dédouanement de marchandises* → *détournement*, est une règle neutre qui ne fournit pas de précision détaillée sur l'infraction constatée, en lien avec la valeur de la confiance générée qui est inférieure à la valeur de la confiance seuil. Cette règle n'entrevoit pas de mise en place d'actions pertinentes de gestion du risque.

3.9 Comparaison entre le modèle économétrique et l'approche de la Fouille de Données

La comparaison du modèle économétrique et l'approche de la méthode de fouille de données se fera à l'aide de quatre paramètres : *La sélection des données, les critères de sélection, le résultat et le type de modèle.*

Tableau 3.10: Tableau de comparaison entre le modèle économétrique et l'approche de fouille de données

Libellé	Modèle Économétrique (Régression linéaire multiple)	Fouille de Données (Recherche de règles associatives)
Données	Sélection de données de la base des opérations douanières (échantillonnage)	Données des Infraction issues du PVS
Critères	Établis par un Comité de sélectivité [Walsh and Widdowson, 2008]	Support, Confiance et le Lift
Résultats	Affectation de scores sur la fraude constatée	Extraction de connaissance à partir de règles d'association
Type de modèle	Modèle statique (Analyse de données)	Modèle dynamique (Algorithme de fouille de données)

La méthode économétrique bien que permettant l'analyse du risque douanier présente cependant des limites pour l'analyse des grandes masses de données (plusieurs variables), et dans son approche méthodologique (sondage et test statistique sur des données expérimentales).

Quant à l'approche de la Fouille de données (règles d'association) ; elle permet l'analyse de plusieurs variables numériques ou non sans hypothèses prédéfinies ainsi que l'exploration automatique de grandes quantités de données.

3.10 Conclusion

L'Extraction de Connaissances à partir des Données est de nos jours l'un des moyens les plus utilisés pour apprendre des grands volumes de données. Dans cette première contribution, nous avons présenté une approche originale de découverte de connaissances appliquée à des données relatives aux infractions douanières du Procès-verbal Simplifié 'SYDAM World) des douanes de la République de Côte d'Ivoire.

Le résultat obtenu à partir de l'algorithme Apriori de la méthode des règles d'associations, est un ensemble de règles de connaissances de prévision et de ciblage des situations à risque. Pour tester et valider notre approche, nous nous sommes basés sur un critère de sélection lié à la fréquence d'apparition des motifs (infractions). L'emprunt de ce critère de sélection a permis de montrer la pertinence de cette approche ; en faisant ressortir des règles d'associations permettant d'avoir la cartographie des risques douaniers à l'importation et à l'exportation.

Cette cartographie constitue ainsi un outil indispensable de gestion stratégique de la lutte contre les infractions dans le cadre des opérations de dédouanement.

Dans le chapitre suivant, nous présentons la seconde contribution de la thèse qui est une analyse prédictive des cas de comportements à risques des opérateurs économiques, par une extension des données quantitative aux données symboliques.

CHAPITRE 4 : Analyse Prédictive des Comportements à Risques.

Résumé du chapitre *: Ce quatrième chapitre est notre seconde contribution qui traite de l'exploration de la structure symbolique des données avec une extension de l'algorithme Apriori. Cette approche a permis de déterminer de nouvelles règles d'associations au niveau des opérateurs afin d'étudier leur comportement par rapport à la fraude tout en considérant un nouvel indicateur : « Confiance Diagramme (CD) ».*

Sommaire :

		Pages
4.1	Introduction	160
4.2	Positionnement du problème	160
4.3	Notion de la fouille de données symboliques : Concept de base	163
4.4	Méthode Apriori Étendu	166
4.5	Étapes algorithmique de l'approche Apriori Étendu	169
4.6	Application et Résultats	170
4.7	Comparaison de l'algorithme Apriori et de l'algorithme Apriori étendu	173
4.8	Conclusion	174

4.1 Introduction

Dans ce chapitre, nous proposons une méthodologie pour l'extraction de connaissances sous la forme de concept décrivant des comportements à risques. En effet, plutôt que d'extraire des règles d'association au niveau des opérations de douane entre le type et la nature des infractions douanières comme précédemment décrit au chapitre 3, nous explorons la structure symbolique des données dans l'idée sous-jacente d'extraire de nouvelles règles d'associations au niveau des opérateurs afin d'étudier leur comportement dans le processus de fraude. Cette méthodologie est composée d'un ensemble d'indicateurs que nous définissons en spécifiant le seuil minimum fixé pour les expérimentations à faire. La problématique de notre travail consistera à savoir si les méthodes ensemblistes (règles d'associations, analyse formelle de concepts) issues de la fouille de donnés peuvent nous permettre de prédire un comportement à risques des opérateurs économiques lors des opérations de dédouanement. La suite de la contribution s'articule autour du plan suivant :

- Positionnements du problème
- Définition de la notion de fouille des données symboliques
- Approche de résolution dénommée Apriori par diagramme ou Algorithme Apriori Etendu
- Applications et résultats

4.2 Positionnement du problème

Dans le chapitre précèdent, nous avons vu que les règles d'associations ont été créées pour extraire de la connaissance à partir de données et sont, généralement de la forme suivante : ''*Si <antécédent>, Alors <conséquent> FinSi*''. Dans la littérature, nous observons que la définition des règles d'associations varie selon les trois principaux courants :

- (1) Les règles d'associations avec la statistique décisionnelle [Gras, 1979] ;
- (2) les règles d'associations avec une représentation ordonnée de concepts informatives [Guigues et Duquenne, 1986] ;
- (3) les règles d'associations avec analyse des données transactionnelles dans les grandes Bases de Données [Agrawal et Srikant, 1993].

Dans cette thèse, l'approche de règles d'associations est basée sur la troisième (3ème) définition, c'est à dire les règles d'association qui associent l'analyse des données liées aux bases de données transactionnelles. Un exemple classique de cette règle d'association est l'approche dite du panier de la ménagère par Agrawal et Srikant dans (Agrawal et Srikant 1994), donne une vue sur un ensemble d'achats effectués dans un supermarché. Ainsi, en considérant deux articles X et Y, la règle d'associations du type $X \rightarrow Y$ signifie que : *si l'article X est présent dans le panier de la ménagère alors il y a aussi l'article* Y. Cette analyse avait pour but de dégager des relations intéressantes afin de pouvoir bâtir une stratégie commerciale conséquente pour les décideurs :

- *« Soit, faut-il mettre les deux produits sur la même rangée ou faut-il les séparer pour obliger le consommateur à passer plus de temps dans le supermarché ?*
- *Soit faut-il mettre près d'eux les produits les plus générateurs de profits ou les produits dont la date de péremption est proche ? »*
- *Etc, …*

Dans la suite, l'on constate que plusieurs études ont été menées afin d'optimiser la base des règles d'association en exploitant la structure complexe des données dans les bases de données. Ainsi, nous citons les travaux de *Guigues* et *Duquenne* [Guigues et Duquenne, 1986] qui ont défini une base minimale pour les règles d'association exactes avec une probabilité de la confiance fixée à 1. D'autres part, la fréquence d'apparition des articles dans le panier de

la ménagère, a permis à [Wang et al. 2000] et [Cai et al. 1998] de découvrir des règles pondérées. [Srikant et al., 1997] et [Han et Fu, 1995] exploitent quant eux les relations de taxonomie dans les données. La création de nombreux indicateurs de qualité a permis sa généralisation à d'autres types de données, notamment dans l'exploitation des ensembles flous [Kuok et al., 1998], et l'utilisation des données quantitatives [Srikant et Agrawal, 1996] et [Miller et Yang, 1997]. Dans les travaux précédemment cités, l'extraction de règles d'associations est basée sur deux principaux indicateurs : Le support et la confiance pour des valeurs supérieures à des seuils minimaux ; Dès lors, de nombreux indicateurs, outre le support et la confiance, ont été proposés afin d'évaluer la qualité des règles obtenues : *confiance centrée*, *conviction*, *gain d'entropie*, *taux informationnels, lift, Piatetsky-Shapiro, Laplace*.... pour la plupart, ces indicateurs sont corrélés au support et à la confiance [Bayardo et Agrawal, 1999] et [Blanchard J., 2004]. A l'inverse, certains auteurs ont travaillé à proposer des algorithmes plus performants sur la complexité des données pour éviter une explosion du temps d'extraction des règles à cause de la capacité des espaces de stockage des données [Pasquier, 2000]. Dans notre première contribution, nous avons proposé un Algorithme Apriori qui a permis l'extraction de règles d'association mettant en relation la nature de l'infraction et l'opération de dédouanement. La seconde contribution s'inscrit dans le prolongement de ce travail. En effet, nous nous proposons d'étendre cette approche dans l'idée sous-jacente de mettre en évidence de nouvelles règles d'associations au niveau des opérateurs économiques en analysant leur comportement par rapport aux infractions constatées au cours des opérations de dédouanement. Ainsi les problématiques suivantes sont dégagées :

- *Quelle relation peut-on établir entre une fraude constatée et le comportement d'un operateur ?*

- *Peut-on établir un lien entre deux ou plusieurs types d'infractions pour un même opérateur économique ?*

La résolution de ces différents problématiques implique notamment une redéfinition des indicateurs tels : **le support** et **la confiance** afin de tirer parti de la structure ou notion symbolique des données.

4.3 Notion de la fouille de données symbolique : Concept de base

Depuis [Agrawal et al. 1993], la recherche d'algorithmes capables d'extraire des règles d'associations dans de grandes bases de données a été un thème très étudié. La découverte de règles d'association entre différents produits présents dans le panier de la ménagère a été un exemple d'application particulièrement exploité.

Tableau 4.1 : Exemple d'une matrice avec 10 cas de transactions classiques.

Transaction	Libellé opérateur	Items
t_1	1	$a_1, a_2, a_5,$
t_2	1	a_2, a_4
t_3	1	a_2, a_3
t_4	2	a_1, a_2, a_4
t_5	2	a_1, a_2, a_3
t_6	3	a_2, a_3, a_5
t_7	3	a_1, a_3
t_8	3	a_1, a_2, a_3, a_5
t_9	4	a_1, a_2, a_3
t_{10}	4	a_2, a_3

Nous étendons ces règles au niveau concept pour élucider le comportement des opérateurs économiques (*Importateurs, exportateur, commissionnaires en douane agrées*) par rapports

aux différentes infractions constatées sur les différentes opérations de dédouanement répertoriées.

Dans le tableau 4.1, nous présentons un exemple classique de la matrice des transactions.

Table 4.1 : Exemple d'une matrice avec 10 cas de transactions classiques, en considérant quatre (4) types d'opérateurs économiques identifier par les Id (Identifiants) :1, 2, 3 et 4).

Pour une meilleure lecture et compréhension du tableau, le lecteur est invité à se référer au chapitre 3 du mémoire de la thèse.

4.3.1 Description de la notion d'objet symbolique

Un objet symbolique (OS) modélise des concepts. L'idée principal de l'analyse des données symboliques est de passer de l'étude des individus à l'étude des concepts décrits par des variables intervalles, pondérées, diagrammes et pour lesquelles les opérateurs numériques standards ($\times$, $+$, $-$) ne peuvent être appliqués de façon directe [Bock et Diday 2000]. Ainsi, un concept est généralement défini par un ensemble de propriétés appelé intension et un ensemble d'individus satisfaisant ces propriétés appelé extension [Bock et Diday 2000].

4.3.2 Définition de la notion de données symboliques

Définition 4.1 (Donnée symbolique) **:** Soit Ω, l'ensemble des individus, et D l'ensemble des descriptions d'individus ou de classe d'individus. Un OS est un triplet $s = (a, R, d)$ où$d \in D$ est une description, R est une relation entre ''d'' et ''a'' de $\Omega \to L$: C'est une fonction de reconnaissance entre les individus et leurs descriptions.

Deux types d'OS sont répertoriés pour deux ensembles L différents : Les OS du type booléens et les OD des types modaux.

- Les OS booléens sont tels que : $[y(w)Rd] \in L = \{\text{vrai}, \text{faux}\}$.

Par exemple $a(w) =$ Nature_Infraction $(w) \subseteq [a_1, a_2] \wedge$ [Typologie_Infraction$(w) \subseteq [b_1, b_2]]$= (vrai $\vee$ faux) = Vrai où $w \in \Omega$ et b_1, b_2 sont des types d'infractions. Nature_*Infraction et Typologie_Infraction* sont deux variables qui décrivent w.

- Les OS Modaux sont tel que : $[y(w)Rd] \in L = [0,1]$. Dans la suite de cet article, nous utilisons que les OS Booléens pour analyser les comportements à risque liés à l'activité douanière.

Définition 4.2 : Une assertion est un OS défini par$[d'Rd] = \wedge_{i=1,p} [d'_i R_i d_i], p \geq 1$.

L'extension d'un OS est donnée par $Ext(s) = \{ w \in \Omega / a(w) = vrai\}$

Remarque : Nous précisons que l'extension donnée à la **définition 2** ne concerne que le cas booléen

4.3.3 Présentation d'une matrice symbolique

Se référant à une matrice classique des transactions, il s'agira plutôt d'avoir un diagramme dans chaque case, c'est-à-dire, des valeurs multiples pondérées telles que la somme des poids (p_i) soit égale à un $(\sum p_i = 1)$ au lieu d'une valeur unique par case dans notre matrice de données (Voir tableau 4.2) ou bien un ensemble d'items par transaction comme dans le cas classique.

Dans l'approche Apriori classique, les unités statistiques sont des transactions ; en revanche avec cette nouvelle approche : ‘‘Apriori Diagramme”, ce sont des concepts que nous étudions, c'est à dire le comportement des opérateurs par rapport aux infractions douanières plutôt que des opérations douanières.

Tableau 4.2 : Matrice des données symboliques composée d'une valeur diagramme

N°	Concept = Opérateur	A= items
1.	1	$\frac{1}{7}a_1, \frac{3}{7}a_2, \frac{1}{7}a_3, \frac{1}{7}a_4, \frac{1}{7}a_5$
2.	2	$\frac{1}{3}a_1, \frac{1}{3}a_2, \frac{1}{6}a_3, \frac{1}{6}a_4$
3.	3	$\frac{2}{9}a_1, \frac{2}{9}a_2, \frac{1}{3}a_3, \frac{2}{9}a_5$
4.	4	$\frac{1}{5}a_1, \frac{2}{5}a_2, \frac{2}{5}a_3$

On remarquera que, pour chaque opérateur, la somme des poids (p_i) soit égale à un $(\sum p_i = 1)$ (À observer dans la dernière colonne de la tableau4.2)

4.4 Méthode de l'algorithme Apriori Étendu

4.4.1 Principe de la méthode

Pour appliquer l'approche Apriori étendu, nous "discrétisons" les fréquences de chaque catégorie des diagrammes en deux principales étapes sous la forme de cet algorithme :

Début

Étape 1 : Segmentation en intervalles des fréquences des diagrammes

Nous partitionnons en intervalles les fréquences$F_{X_{i,c}}$ pour chaque catégorie de chaque variable X_i. Ainsi, nous regardons les supports des intervalles de fréquences $0 < F_{X_{i,c}} \leq \frac{1}{h}$, $\frac{1}{h} < F_{X_{i,c}} \leq \frac{2}{h}$, $\frac{2}{h} < F_{X_{i,c}} \leq \frac{3}{h}, \ldots, \frac{h-1}{h} < F_{X_{i,c}} \leq 1$où h détermine la précision du découpage.

Répéter

Étape 2 : Union des intervalles fréquences

Nous faisons l'union 2 à 2 des intervalles de poids contigus ayant des supports strictement positifs $0 < F_{X_{i,c}} \leq \frac{2}{h},\ \frac{1}{h} < F_{X_{i,c}} \leq \frac{3}{h}, \dots, (h-2) < F_{X_{i,c}} \leq 1$.

Jusqu'à obtenir un unique intervalle $0 < F_{X_{i,c}} \leq 1$.

Fin

A la suite, nous considérons ces différentes classifications d'intervalles. Ainsi, nous travaillons avec des objets symboliques (OS) booléens et non plus avec des ensembles d'items où les intervalles de fréquences sont les propriétés des OS qui ont donc pour intensions $a(w) = [\frac{a}{h} < F_{X_{i,c}}(\omega) \leq \frac{b}{h}](a = 0 \dots h-1, b = 1 \dots h, a < b)$. Finalement, un *k-OS* est une assertion booléenne définie à partir de *k* propriétés. Par exemple, en considérant ε et ε' deux catégories de deux variables diagrammes Aet A' avec $F_{X_\varepsilon}, F_{X'_{\varepsilon'}}$ leurs fréquences respectives alors $\left[\frac{1}{3} < F_{X_\varepsilon} \leq \frac{2}{3}\right] \wedge \left[0 < F_{X'_{\varepsilon'}} \leq \frac{2}{3}\right]$ sera un 2-OS.

Remarque : Ces *k*-OS ne seront pas totalement traités comme des catégories de l'algorithme classique. Il ne faut pas croiser des intervalles de même catégorie et nous devons à chaque fois utiliser les plus petits intervalles de fréquences possibles pour un même support.

4.4.2 Précision du découpage h

Le choix d'une valeur de h est fonction du nombre de modalités des variables à étudier et de son besoin de résultats plus ou moins précis. Bien évidemment, plus la précision est grande plus le nombre de *k*=1-OS sera grand par rapport au nombre d'items dans le cas classique. En contrepartie, la transformation de la matrice des données classiques en données symboliques aura réduit le nombre d'individus à étudier. Par exemple, en se référant au tableau 4.2, nous pouvons utiliser une précision de *h=3* (nombre minimum de catégories par

concept ou *h=5* (nombre maximum de catégories par concept) ou tout autre valeur nécessaire à la précision pour les besoins de l'étude à faire.

4.4.3 Définition des indicateurs : *Support, Confiance, Confiance Diagramme*

Soient Ω un échantillonnage (ensemble de concepts), A et B deux OS ayant pour extensions $a_x(\omega) = \wedge_{i,u} [\frac{a_{i;u}}{h} < F_{X_{i,u}}(\omega) \leq \frac{b_{i,u}}{h}$ et $a_y(\omega) = \wedge_{j,v} [\frac{c_{j,v}}{h} < F_{Y_{j,v}}(\omega) \leq \frac{d_{j,v}}{h}$

Avec $\forall, u, j, v, X_{i,u} \neq Y_{j,v}$ où $F_{X_{i,u}}(F_{Y_{j,v}})$ est la fréquence de la catégorie $u\ (v)$ de la variable diagramme $X_i(Y_j)$, $\frac{a_{i;u}}{h}, \frac{b_{i,u}}{h}$ ($\frac{c_{j,v}}{h}$, $\frac{d_{j,v}}{h}$) sont les bornes des intervalle de fréquences.

- Définition du support (Supp)

$$\text{Supp } (A \to B) = \frac{Card(ext(A \wedge B) = \{\omega \in \Omega / a_x(\omega) = vrai,\ a_y(\omega) = vrai}{Card(\omega)}$$

- **Définition de la Confiance (Conf.)**

$$\text{Conf.}(A \to B) = \frac{Card(ext(A \wedge B) = \{\omega \in \Omega /\ a_x(\omega) = vrai,\ a_y(\omega) = vrai}{Card(ext(A) = \{\omega \in \Omega /\ a_x(\omega) = vrai} = \frac{\textbf{Supp}\ (A \to B)}{\textbf{Supp}\ (A)}$$

- **Définition de la Confiance Diagramme (CD.)**

En adoptant l'approche par diagramme, il est intéressant de définir un nouvel indicateur de qualité (confiance diagramme ou CD) pénalisant les règles ayant les plus grands intervalles de fréquences et donc la plus grande imprécision en conclusion.

$\text{CD } (A \to B) = Conf(A \to B)/(1 + \frac{\sum_{j,v}(d_{j,v}, c_{j,v})}{n_v X h})$ où n_v est le nombre de propriétés en conclusion.

L'indicateur CD est tel que : $\frac{1}{2} Conf(A \to B) \leq CD(A \to B) \leq \frac{h}{h+1} Conf(A \to B)$.

Pour générer les règles d'associations symboliques, nous définissons un CD minimum (*MinCD)* pour la génération des règles.

4.5 Étapes algorithme de l'approche Apriori Étendu

Les grandes étapes de l'algorithme ''Apriori diagramme '' sont présentées dans cette section. En initialisation, il faut définir la valeur de la précision h selon les besoins de l'utilisateur et le support minimum.

Étape 1 : Discrétiser les fréquences de chaque catégorie

Étape 2 : Calculer les supports des intervalles de poids précédents avec un passage dans la matrice des données. Nous faisons alors l'union deux à deux des intervalles contigus de supports strictement positifs. Nous répétons les unions 2 à 2 de nos nouveaux intervalles jusqu'à obtenir un unique intervalle $0 < F_{X_{i,c}} \leq 1$. Les supports de ces intervalles sont calculés sans passage dans la matrice des données car si $A \; et \; A'$ sont des intervalles contigus alors $\mathrm{Sup}(A \cup A') = \mathrm{Sup}(A) + \mathrm{Sup}\,(A')$. Nous ajoutons à $L_{k=1}$ les 1-OS de support supérieur au seuil *Minsup*.

Étape 3 : Faire tant que l'ensemble des *k-OS* (assertion définie avec la conjonction de *k* intervalles de fréquences) fréquents $L_k \neq \emptyset \; (k \geq 1)$:

a- Générer les *k+1*-OS candidats en calculant le produit cartésien entre les k-OS deL_k. Dans le cas des diagrammes, nous générons les *k+1*-OS entre intervalles de catégories différentes (et "non marqués" ne voir point (c)). Ainsi, l'ensemble des candidats C_{k+1} est généré. Du fait de la propriété 3, nous supprimons de C_{k+1} tout *k+1*-OS I tel qu'il existe un *k-OS* $J \subset I$ n'appartenant pas à L_k.

b- Pour tout $c \in C_{k+1}$, calculer le support avec un passage dans la matrice de données. Tout *k+1*-OS I $\in C_{k+1}$ fréquent est ajouté à L_{k+1}.

c- Marquer tout *k+1-OS* $I \in L_{k+1} / \; \exists J \in L_{k+1}$ avec $J \subset I$ et *sup(I) = sup(J).* Il s'agit de *k+1-OS* définis avec les mêmes catégories mais avec des intervalles de poids différents et nous conservons uniquement les plus petits intervalles pour un même support. Nous

les marquons au lieu de les supprimer car ces *k+1*-OS ne sont pas utilisés pour la génération de *k+2*- OS mais ils sont utilisés pour la génération de règles.

d- Générer les règles avec un CD supérieur *MinCD* (Minimum de Confiance Diagramme).

4.6 Application et Résultats

Nous précisons que les conditions expérimentales décrites au chapitre troisième du mémoire de thèse, les sections 3.6 et 3.7 sont les mêmes.

4.6.1 Résultats et Analyses

Dans la première contribution de cette thèse, nous avons mis en évidences trois catégories de règles d'association :

- *La règle de prévision* dont l'appétence reçoit 57% des risques permet à l'administration douanières de mettre en place des stratégies d'anticipation à l'effet de juguler les fraudes ;
- *La règle de ciblage* est une règle qui permet de cibler aux administration douanières un risque spécifique de fraude ;
- *La règle neutre* : cette règle ne présente pas d'intérêt, car l'information extraite n'est pas pertinente.

Ces règles d'associations bien qu'elles fournissent des renseignements importants entre une operation de dédouanement et le type d'infraction ne renseignent pas sur les comportements à risques des opérateurs par rapport aux activités de dédouanement.

Pour expliquer et prédire le comportement à risque des opérateurs économiques liés à l'activité de contrôle douanier, nous proposons une analyse symbolique par diagramme des données des infractions douanières issues du Procès-Verbal Simplifié (SYDAM World)

Dans le cadre de l'analyse des données symboliques, les infractions saillantes recensés à partir de la méthode des règles associatives au chapitre 3 sont : (a) *fausse déclaration de valeur*, (b) *fausse déclaration d'espèce*, et (c) *fausse déclaration d'origine*.

Tableau 4.3 : Matrice des opérations de dédouanement et des infractions

Transaction	Opérateur	A= Infractions
1	OP 001	a,b
2	OP 001	a,b,c
3	OP 001	a
4	OP 002	b,c
5	OP 002	c
6	OP 003	a
7	OP 003	a,b,c
8	OP 003	a,b,c
9	OP 003	a
10	OP 004	a,b
11	OP 004	a

Tableau 4.4 : Matrice des données symboliques

Concepts = Opérateurs	A = Infractions	Concepts = Opérateurs	A= Infractions
001	$\frac{1}{2}b,\quad \frac{2}{3}c$	003	$\frac{1}{2}a,\quad \frac{1}{4}b,\quad \frac{1}{4}c$
002	$\frac{1}{3}a,\quad \frac{1}{3}b,\quad \frac{1}{6}c$	004	$\frac{2}{3}a,\quad \frac{1}{3}b$

NB : Nous recensons au plus trois (3) infractions par opérateur

Le paramétrage de la précision dans l'algorithme est $h = 3$ (Correspondant au plus grand nombre d'infractions recensées par opérateur) ; Dans les expérimentations les paramétrages des indicateurs de mesure sont :

- Support Minimum (*MinSupp*) est fixé à 70 %,;
- Minimum du Confiance Diagramme (*MinCD)* à 55%.

Tableau 4.5 : Règles d'associations symboliques

N°	Règles d'associations	Supp	Conf	CD
1	$\frac{1}{3} < F_a \leq \frac{2}{3} \rightarrow 0 < F_b \leq \frac{1}{3}$	0,70	1	0,70
2	$0 < F_b \leq \frac{1}{3} \rightarrow \frac{1}{3} < F_a \leq \frac{2}{3}$	0,70	1	0,70
3	$0 < F_c \leq \frac{1}{3} \rightarrow 0 < F_b \leq \frac{2}{3}$	0,70	1	0,60
4	$0 < F_b \leq \frac{2}{3} \rightarrow \frac{1}{3} < F_a \leq \frac{2}{3}$	0,70	0,75	0,55
5	$0 < F_b \leq \frac{2}{3} \rightarrow 0 < F_c \leq \frac{1}{3}$	0,70	0,75	0,55

Analyses et interprétations des résultats (Tableau 4.5)

Dans le tableau 4.5 :

- la première règle stipule que les opérateurs qui fraudent sur la valeur sont les mêmes qui font la fausse déclaration sur les espèces ; bien qu'ils fraudent le plus la valeur que sur les espèces.
- La règle n°2 met en exergue les profils de fraude des opérateurs, aussi bien sur les déclarations d'espèces, que sur les déclarations d'origine des marchandises. Les deux infractions commises par les opérateurs impactent négativement le revenu douanier.
- La règle 3 indique que l'opérateur à une forte propension à falsifier les valeurs des marchandises et à commettre des glissements tarifaires (Fausses déclarations

d'espèces). La valeur est un élément prépondérant dans la détermination des droits de taxes car de façon générale, le calcul des droits au cordon douanier est ad valorem. La déclaration des valeurs à la baisse est de nature à minimiser les recettes douanières. Quant aux fausses déclarations d'espèces, elles sont également de nature à rabaisser les droits de porte car les opérateurs concernés déclarent les marchandises à faible taxation.

L'approche de l'analyse des données par diagramme à partir de l'Algorithme Apriori Etendu a permis de mettre en lumière la propension à commettre plusieurs infractions par un même opérateur économique au cordon douanier.

L'analyse des risques douaniers sur la base de l'approche par diagramme offre aux administrations douanières un outil d'analyse de profils des opérateurs dans le cadre de la procédure des Opérateurs Economiques Agréés (OEA).

4.7 Comparaison de l'algorithme Apriori et de l'algorithme Apriori étendu

La comparaison de l'algorithmes Apriori et de l'algorithmes Apriori Etendu se fera à l'aide de deux paramètres : *Les indicateurs et le résultat*

Tableau 4.6: Tableau de comparaison entre les algorithmes Apriori et la méthode par diagramme

Libelle	Approche Apriori (*Règle d'associations*)	Approche par diagramme (*Approche Apriori Etendu*)
Données	Base de Données : Ensemble de données issues des opérations douanières	
Critères	Support, Confiance et le Lift	Support, Confiance et Confiance Diagramme
Résultats	Toutes les règles d'association ne sont pas exploitables	Règles d'association symboliques pour des informations prédictives

Le tableau de comparaison 4.6 montre que si **la méthode des règles associatives** permet d'obtenir des informations de prévision et de ciblage en établissant une relation entre une opération de dédouanement et une infraction ; ne précise pas clairement le rôle de

l'opérateur économique dans l'infraction ; tandis que le passage des données quantitatives aux données qualitatives avec **l'approche par diagramme** établit nettement le lien entre l'opérateur économique et une ou plusieurs infractions avérés.

4.8 Conclusion

L'approche par diagramme permet d'établir une corrélation entre le comportement des opérateurs économiques et les infractions constatées. En effet, en général, les systèmes de dédouanement des marchandises sont déclaratifs car les opérateurs économiques, personnes physiques ou morales, importateurs ou exportateurs sont astreints à faire des déclarations de marchandises afin de déterminer des droits et taxes exigibles. Dans un tel système de contraintes déclaratives, la tentation à la fraude est grande, c'est pourquoi, la lutte contre la fraude douanière à partir de l'analyse symbolique est essentielle et constitue une recherche de solution et d'anticipation afin de devancer la fraude au lieu de la combattre à posteriori. Ainsi, l'application de l'extension de l'algorithme Apriori en considérant des variables symboliques par diagramme a permis de mettre en évidence de nouvelles règles d'associations étendues aux comportements à risques des opérateurs dans le processus de dédouanement. Dans une analyse prédictive, on pourrait prévenir le comportement à risques de certains opérateurs liés aux activités douanières à chaque nouvelle infraction. Cependant, en observant le paramètre du critère de sélection basé sur la notion de la fréquence des motifs et l'indicateur confiance diagramme, le temps de calcul mis par l'algorithme pour effectuer des passages dans la matrice (*Calcul du support et détermination des ensembles fréquents*) des données augmentent de façon exponentielle au fur et mesure que le volume des données devient croissant. Ce qui induit une difficulté en temps d'exécution de l'algorithme en vue d'extraire des règles. Il se pose dès lors une problématique d'optimisation de l'algorithme qui ouvre à d'autres options de recherches.

CONCLUSION GENERALE ET RECOMMANDATIONS

1. Bilan des contributions

L'extraction de connaissances à partir de données est le processus de découvertes de connaissances utiles à partir d'un jeu de données. Ce processus se décompose en plusieurs étapes mais nous avons mis l'emphase dans ce projet de thèse sur l'étape qui consiste à extraire les connaissances : la fouille de données. L'approche d'une méthode de fouilles de données pour l'analyse des comportements à risques liés à l'activité douanière est le thème sous lequel a été élaboré ce projet de thèse de doctorat. Les termes « *Approche d'une méthode* » et « *analyse des comportements à risques* » sont très importants puisqu'ils stipulent l'application d'une approche de fouille de données pour extraire des connaissances sur les comportements à risques des acteurs en lien avec la nature des fraudes. Et vu qu'il n'existe pas de publication sur le datamining utilisé dans le domaine de l'analyse des risques liés à l'activité douanière, nous étions face à un défi car la plupart des solutions proposées pour les administrations douanières sont issues d'études statistiques et économétriques. Au terme de ces trois années de recherches, nous avons abordé la problématique en deux grandes étapes expliquées dans les contributions suivantes :

- La première contribution d'ordre théorique et applicative prend la forme d'une adaptation de l'algorithme Apriori afin d'exploiter dans un contexte précis un entrepôt de données de fraudes issues de transactions douanières. Le paramètre principal considéré est la fréquence des motifs (*fraude ou facteur de risques de fraudes*). Ce qui a permis d'obtenir des règles d'associations dites valides en considérant les indicateurs suivants : *support*, *confiance* et *Lift* à partir d'un seuil minimum. Cette contribution a permis de mettre en évidence trois grandes règles d'associations issues de données binaires : les règles de prévision, les règles de ciblage et les règles neutres qui ont

permis d'établir une corrélation entre une opération de dédouanement et la nature des fraudes ;

- La deuxième contribution d'ordre méthodologique est une suite de la première. En effet, la première contribution, bien qu'elle fournisse des renseignements sur les différents cas d'infractions lors des opérations de dédouanement ne donnent pas d'informations relatives au comportement à risques des opérateurs par rapport à l'activité douanière. Cette contribution modélise des concepts en explorant avec la structure symbolique des données par une extension de l'algorithme Apriori. Dans cette approche, ce sont les concepts qui sont étudiés, c'est à dire le comportement des opérateurs par rapport aux infractions douanières. Ainsi, nous avons déterminé de nouvelles règles d'associations au niveau des opérateurs afin d'étudier leur comportement par rapport à la fraude tout en considérant un nouvel indicateur :« Confiance Diagramme (CD) ». Cette contribution permet d'analyser, le comportement à risques des opérateurs en relation avec la nature des fraudes. Cette contribution aide à la prise de décision dans le cadre de la procédure des Opérateurs Economiques Agréés (OEA).
- Les règles élicitées à partir de l'algorithme apriori et son extension sont des règles dynamiques en fonction la mise à jour des jeux de données des infractions douanières.

2. Intérêts des travaux pour les administrations douanières

La fouille de données appliquée aux infractions douanières à travers les deux contributions présente des intérêts majeurs pour les administrations douanières à savoir :

- Etablir la cartographie des risques de fraude en douane ;
- Analyser et prédire les comportements à risques des opérateurs ;

- Optimiser le contrôle douanier (mieux contrôler en contrôlant moins) ;
- Optimiser la collecte des finances publiques au cordon douanier ;
- Déterminer le profil des opérateurs dans le cadre des Opérateurs Economiques Agréées (OEA) ;
- Fiabiliser les statistiques du commerce international.

3. Limites des travaux effectués

Ces diverses contributions sont toutefois soumises à certaines limites et insuffisances. En dépit des résultats satisfaisants obtenus, les performances de notre approche sont tributaires des choix méthodologiques que nous avons effectués, et les premières limitations proviennent d'une part de la recherche exhaustive des règles d'associations. Parmi les règles obtenues, certaines sont dites valides (positives) et d'autres négatives. Ce qui engendre des coûts computationnels. D'autre part, la connaissance induite par ces règles dites négatives est inaccessible puisque seules les règles valides sont extraites. Cependant, la prise en compte des règles négatives peut s'avérer utile puisqu'elles permettraient d'étendre les connaissances et renfermer des informations non accessibles avec les règles valides.

4. Travaux futurs : Perspectives

Cette section, avec laquelle nous allons conclure ce manuscrit de thèse, présente d'éventuelles pistes de recherches. Les voies de recherches résultent, pour la plupart, des limitations présentées dans la section précédente. Les perspectives de ce travail sont à la fois nombreuses et prometteuses.

- Il serait intéressant d'adapter nos améliorations sur l'algorithme FP-Growth [Han et al. 2000]. Même si Apriori est l'un des algorithmes les plus connus pour l'extraction de règles d'associations, FP-Growth est réputé pour être plus rapide. Cet algorithme utilise une structure compacte appelée FP-Tree qui permet d'extraire les motifs

fréquents sans générer de candidats. De plus, alors qu'Apriori interroge la base pour chaque niveau de motifs de candidats générés, FP-Growth nécessite seulement deux passages, ce qui accélère encore les traitements.

- Au niveau des méthodes, il serait utile de poursuivre la recherche avec la technique de la fouille de données supervisée. Ce qui permettrait de développer des algorithmes d'apprentissage automatisé ("machine Learning" en anglais). En effet avec l'apprentissage supervisée, les résultats escomptés par le biais des algorithmes dédiés sont connus et les algorithmes apprennent grâce à un jeu de données de formation qui contient toutes réponses correctes. Aux fins de cette formation, les déclarations douanières sont utilisées pour détecter les données aberrantes.
- Concevoir et mettre en œuvre une politique efficace de renseignements douaniers en vue d'enrichir les données relatives aux comportements à risques liés à l'activité douanière à l'effet d'une gestion optimale des risques douaniers au travers des outils technologiques de gestion et d'analyse des risques.

REFERENCES BIBLIOGRAPHIQUES

[Aaker et al.,2007] Aaker, V.Kumar, & S.Day, G. (2007). Marketing Research. NY: John Wiley and Sons.

[Abbas, 2012] Abbas, H. (2012). Expansion de la représentation succincte des générateurs minimaux. (Mémoire de maîtrise). Université du Québec à Montréal.

[Afonso et al., 2004] F. Afonso, L. Billard et E. Diday. Régression Linéaire Symbolique avec Variables Taxonomiques. Actes des 4èmes journées d'Extraction et de Gestion des Connaissances, EGC'04, Clermont-Ferrand, Cépadues, 2004.

[Agrawal et al., 1993] R. Agrawal, T. Imielinski and A. N. Swami. Mining association rules between sets of items in large databases. In Proceedings of ACM International Conference on Management of Data (SIGMOD), pages 207–216, May 1993. (Cited on pages 2, 42 and 44.)

[Agrawal et Srikant, 1994a] Agrawal R., Srikant R., « Fast Algorithms forMining Association Rules in Large Databases », Proc. of the 20th Int'l Conf. on Very Large Data Bases (VLDB), juin 1994, p. 478-499, version etendue : IBM Research Report RJ 9839.

[Agrawal et Srikant, 1994b] Agrawal, R. & Srikant, R. (1994b). Fast algorithms for mining association rules. In J.B. Bocca, M. Jarke & C. Zaniolo, eds., Proc. 20th Int. Conf. Very Large Data Bases, VLDB, 487-499, Morgan Kaufmann. 62

[Agrawal et Srikant, 1994c] Agrawal, R. & Srikant, R. (1994c). Fast algorithms for mining association rules in large databases. In VLDB '94 : Proceedings of the 20th International Conference on Very Large Data Bases, 487-499, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 25

[Agrawal et Srikant, 1995] Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. Proc. of Int. Conf. on Data Engineering, pp. 3-14.

[Agrawal et Srikant, 1996] Agrawal, R. Srikant, H, (1994) Fast algorithms for mining association rules in large databases, Research Report RJ 9839, IBM Almaden Research Center, San Jose, California, June 1996.

[Atluri et al., 2009] Atluri, G., Gupta, R., Fang, G., Pandey, G., Steinbach, M., and Kumar, V. (2009). Association analysis techniques for bioinformatics problems. In Bioinformatics

[Aumann et Lindell, 2003] Aumann, Y., & Lindell, Y. (2003). A statistical theory for quantitative association rules.

[Ayadi et al., 2009] Ayadi, W., Elloumi, M., and Hao, J.-K. (2009). A biclustering algorithm based on a bicluster enumeration tree: application to dna microarray data. BioData mining, 2(1) :9.

[Ayres et al., 2002] Ayres, J., Flannick, J., Gehrke, J., &Yiu, T. (2002). Sequential pattern mining using a Barbut M., Monjardet B., Ordre et Classification : Algèbre et Combinatoire. Volume I & II. Classique Hachette, Paris, 1970.

[Bartunov et al., 2012] Bartunov, S., A. Korshunov, S. Park, W. Ryu, et H. Lee (2012). Joint Link-attribute User Identity Resolution in Online Social Networks. In SNA-KDD Workshop.

[Bastide, 2000] Bastide Y., « Data mining : algorithmes par niveaux, techniques d'implantation et applications », thèse d'université, université Clermont-Ferrand II, décembre 2000.

[Bayardo et Agrawal, 1999] Bayardo Jr R. J., Agrawal R., "Mining the most interesting rules", Proc. of the Fifth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, 1999, p. 145-154.

[Berlin, 2002] Berlin, Heidelberg: Springer Berlin Heidelberg. "Association Rule Mining : Models and Algorithms, 2002.

[Berry et Linoff, 1997] Berry, M. J., & Linoff, G. (1997). Data Mining: For Marketing, Sales, and Customer Support. John Wiley & sons, Inc.

[Bilal, 2013] Bilal, I. Méthodologie d'extraction de connaissances spatio-temporelles par fouille de données pour l'analyse de comportements à risques : application à la surveillance maritime. Architecture, aménagement de l'espace. (2013), Thèse, Ecole Nationale Supérieure des Mines de Paris, 2013.

[Blanchard J., 2004] Blanchard J., Guillet F., Gras R., Briand H. (2004). Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel TIC", revue RNTI, Extraction et Gestion des Connaissances, Vol.1, 2004, p. 287-298, Cépadues, Paris.

[Bock et Diday 2000] H-H. Bock et E. Diday. Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data, Springer Verlag, Heidelberg, 2000.

[Bosc, et al., 2001] Bosc, P., Pivert, O., Dubois, D., & Prade, H. (2001). On fuzzy association rules based on fuzzy cardinalities. Proc. of the 10th IEEE Int. Conf. on Fuzzy Systems, pp. 461-464

[Botta et al., 2002] Botta, M., Boulicaut J.-F., Masson C., MeoR.(2002). A Comparison between Query Languages for the Extraction of Association Rules. DaWaK 2002, p. 1-10

[Breiman et al., 1984] Breiman, L., Freidman, J. H., Olshen, R. A., Stone, C. J., Classification and Regression Trees. Wadsworth, 1984. candidate generation. In ACM SIGMOD Record, volume 29, pages 1–12.

[Brinet al., 1997] S. Brin, R. Motwani, J. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In Proceedings ACM SIGMOD, USA, pages 255-264, 1997.

[Bruno, 2006] Bruno, A., Catherine, M., & Martin, T. (2006). Mining Public Transport User Behavior From Smart Card Data. 12 th IFAC Symposium on Information Control Problems in Manufacturing, 14, pp. 193-203.

[Buccafurri et al., 2012] Buccafurri, F., G. Lax, A. Nocera, et D. Ursino (2012). Discovering Links among Social Networks. In Machine Learning and Knowledge Discovery in Databases, Volume 7524 of Lecture Notes in Computer Science, pp. 467–482. Springer Berlin Heidelberg.

[Brin et al., 1997] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: generalizing association rules to correlations. In Proceedings of the 1997 ACM SIGMOD international conference on Management of data, pages 265-276, May 1997.

[Cadot et Napoli, 2004] Cadot, M., Napoli, A. (2004) RA et codage flou des données. SFC'04. (Bordeaux). p.130-133.Forgy E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications, Biometrics, vol. 21, n° 3, p. 768.

[Cai et al., 1998] Cai C.H., Fu A.W.C., Cheng C.H. et Kwong W.W. (1998). Mining association rules with weighted items, Proc. of the 1998 Int'l Database Engineering and Applications Symposium (IDEAS'98), pp 68-77.

[Cao et al., 2005] H. Cao, N. Mamoulis et D.W. Cheung: Mining frequent spatio-temporal sequential patterns. IEEE International Conference on Data Mining ICDM, pages 82–89, 2005. ISSN 1550-4786.

[Cao et al., 2007] H. Cao, N. Mamoulis et D. W. Cheung : Discovery of Periodic Patterns in Spatiotemporal Sequences. IEEE Transactions on Knowledge and Data Engineering TKDE, 19(4):453–467, 2007. ISSN 1041-4347.

[Castleberry, 2001] Castleberry, S. B. (2001). Using Secondary Data in Marketing Research: A Project That Melds Web and Off-Web Sources. Journal of Marketing Education, 23, 195-203.

[Ceglar 2006] A. Ceglar and J.F. Roddick. Association mining. ACM Computing Surveys, vol. 38, 2006.

[Chiu et al., 2004] Chiu, D.-Y., Wu, Y.-H., & Chen, A. (2004). An efficient algorithm for mining frequent sequences by a new strategy without support counting. Proc. of the 20th Int. Conf. on Data Engineering, pp, 375-386.0.

[Christian Bohm et Claudia, 2010] Christian Bohm, A.O., Claudia Plan, SkyDist: Data Mining on Skyline Objects. Springer, 2010.

[Cook et Holder, 2000] Cook, D. J. et Holder, L. B. (2000). Graph-based data mining. IEEE Intelligent Systems, 15(2).

[Cortis et al., 2012] Cortis, K., S. Scerri, I. Rivera, et S. Handschuh (2012). Discovering Semantic Equivalence of People Behind Online Profiles. In In Proceedings of the Resource Discovery (RED) Workshop, ser. ESWC.

[Croft, 1994] Croft, M. J. (1994). Market Segmentation: A Step-by-step Guide to Profitable New Business. Routledge.

[Davey et Priestley] Davey B.A., Priestley H.A., Introduction to Lattices and Order. Cambridge University Press, 4th Edition, 1994.

[David M. Fram, 2008] David M. Fram, J.S.A., MD, PhD, William DuMouchel PhD, Empirical Bayesian Data Mining for Discovering Patterns in Post-Marketing Drug Safety.2008

[DeSarbo, 1984] DeSarbo, W., Carrol, J., & Clark, L. (1984). Synthesized Clustering: A method for Amalgamating Alternative Clustering Bases with Differential Weighting of Variables. Psychometrika , 49, 57-78.

[Devroye et al. 1996] Luc Devroye, László Györfi, and Gábor Lugosi. A probabilistic theory of pattern recognition, volume 31 of Applications of Mathematics (New York). Springer-Verlag, New York, 1996

[Dolnicar et Leisch, 2004] Dolnicar, S., &Leisch, F. (2004). Segmenting Markets by Bagged Clustering. Australasian Marketing Journal , 12, 51-65.

[Dubois et Prade, 1992] Dubois, D., & Prade, H. (1992). Gradual inference rules in approximate reasoning. Proc. Of the IEEE Int. Conf. on Fuzzy Systems, pp. 103-122).

[Džeroski, 1996] Džeroski, S. (1996). Inductive logic programming and knowledge discovery in databases. In Advances in knowledge discovery and data mining, pages 117–152. American Association for Artificial Intelligence, Menlo Park, CA, USA.

[Džeroski, 2003] Džeroski, S. (2003). Multi-relational data mining : an introduction. ACM SIGKDD Explorations Newsletter, 5(1):16.

[Fayyad et al., 1996] Fayyad, U. , Piatetsky-Shapiro, G. et Smyth, P. (1996). From data mining to knowledge discovery in databases. Al Magazine, 17(3), 37-54.

[Florennce, 1990] Florence, M. (1990). International Trade and the Furniture Industry. Alabama International Trade Center , 18-32.

[Forest, 2006] Forest, D. 2006. Application de techniques de forage de textes de nature prédictive et exploratoire à des fins de gestion et d'analyse thématique de documents textuels non structurés. Thèse de doctorat, Montréal, Université du Québec à Montréal.

[Forgy, 1965] R. Forgy, Cluster Analysis of Multivariate Data : Efficiency versus Interpretability of Classification, Biometrics (1965), no 21, 768–769.

[Frawley et al., 1991] Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. (1991).

[Fukuda et al., 1996a] Fukuda, T., Morimoto, Y., Morishita, S., & Tokuyama, T. (1996a). Data mining using towdimensional optimized association rules : Scheme, algorithm and visualisation. Proc. of the ACM Int. Conf. on SIGMOD, pp. 12-23.

[Fukuda et al., 1996b] Fukuda, T., Morimoto, Y., Morishita, S., & Tokuyama, T. (1996). Mining optimized association rules for numeric attributes. Proc. of the ACM Int. Conf. on Sigact-Sigmod-Sigart, pp. 12-23.

[Ganter and Wille, 1999] Ganter, B. and Wille, R. (1999). Formal Concept Analysis. Springer, mathematical foundations edition.

[Geourjon et al., 2010] Geourjon A.-M., B. Laporte et G. Rota-graziosi (2010), « Comment moderniser l'analyse du risque et la sélectivité des contrôles douaniers dans les pays en développement ? », OMD Actualités, n° 62, pp. 29-31.

[Geourjon et al., 2012] Geourjon A.-M., B. Laporte, O. Coundul et M. Gadiaga (2012), “Inspecting less to inspect better: The use of data mining for risk management by customs administrations”, in T. Cantens, R. Ireland, and G. Raballand, Reform by Numbers: Measurement applied to Customs and Tax Administrations in Developing Countries, Development Series, World Bank, Washington, DC.

[Geourjon et Laporte, 2005] Geourjon A.-M. et B. Laporte (2005), “Risk Management For Targeting Customs Controls in Developing Countries: a Risky Venture For Revenue Performance?”, Public Administration and Development, 25, pp. 105-113.

[Giannotti et al., 2011] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo et R. Trasarti : Unveiling the complexity of human mobility by querying and mining massive trajectory data. The Very Large Data Bases Journal (VLDB), 20(5):695–719, 2011. ISSN 1066-8888.

[Giannotti et Pedreschi, 2008] F. Giannotti et D. Pedreschi, éditeurs. Mobility, Data Mining and Privacy - Geographic Knowledge Discovery. Springer, 2008. ISBN 978-3-540-75176-2.

[Gourjeon et al, 2010] Geourjon A.-M., b. Laporte et G. Rota-Graziosi (2010), « Comment moderniser l’analyse du risque et la sélectivité des contrôles douaniers dans les pays en développement ? », OMD Actualités, n° 62, pp. 29-31.

[Gras, 1979] Gras R., (1979) Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques, Thèse, Rennes I, 1979.

[Green, 1997] Green, P. (1977). A new approach to market segmentation. Business Horizons , 20, 61-73.

[Greeno et al., 1973] Greeno, D., Summers, N., & Kernan, J. (1973). PersonalityandImplicit Behavior Patterns. Journal of Marketing Reseach , 10, 63-69.

[Guigues et Duquenne, 1986] Guigues J.L., Duquenne V. (1986) Familles minimales d'implications informatives résultant d'un tableau de données binaires, Math. Sci. Hum. n°95, pp. 5-18

[Guillaume, 2000] Guillaume S. (2000) Traitement des données volumineuses, mesures et algorithmes d'extraction de RA et règles ordinales, Thèse Nantes, 2000.

[Guillet, 2004] Guillet F. (2004) Mesure de qualité des connaissances en ECD, Cours donné lors des journées de la conférence EGC 2004, Clermont-ferrand, 20 janvier 2004.

[Han et 2011] J. Han, M. Kamber, and J. Pei. Data mining : concepts and techniques. Morgan Kaufmann Pub, 2011.

[Han et al., 2000] Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without

[Han et FU, 1995] Han J. et Fu Y. (1995). Discovery of multiple-level association rules from large databases, Proc. of the 21th Int'l Conf. on Very Large Data Bases, 1995.

[Han et Kamber, 2012]. Han, J. et Kamber, M. (201 2). Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann Publishers.

[Han, 2000] Han J., Towards on-line analytical mining in large databases. pp 97-107, SIGMOD Record, ACM Press, N.1 Vol.27, 2000.

[Harrisson, 2007] Harrisson M. (2007), Challenges for customs, Customs and Supply Chain Security-"The Demise of Risk Management?" Annual Conference on APEC centers, 18-20 avril, Melbourne, Australia.

[Hartigan et Wong, 1979] J. A. Hartigan et M. A. Wong, Algorithm AS 136 : a k-means clustering algorithm, Applied Statistics 28 (1979), 100–108.

[Hastie et al., 2001] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning. Springer Series in Statistics. Springer-Verlag, New York, 2001. Data mining, inference, and prediction.

[Houtsma and Swami] M. Houtsma and A. Swami. Set-oriented mining for association rules in relational databases. In P. S. Yu and A. L. P. Chen, editors, Proceedings of the 11th International Conference on Data Engineering, pages 25-34, Los Alamitos, CA, USA, mar 1995. IEEE Computer Society Press.

[Huang, 1998] Zhexue Huang, Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values, Data Min. Knowl. Discov. 2 (1998), No 3, 283–304

[Hüllermeier, 2001] Hüllermeier, E. (2001). Implication-based fuzzy association rules. Principles of Data Mining and Knowledge Discovery, pp. 241-252. Journal of Intelligent Information Systems, 20, 255-283.

[IJCAI, 1989] IJCAI : International Joint Conference on Artificial Intelligence. (1989).Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (I) IJCAI-89 Contents Vol1, https://www.ijcai.org/proceedings/1989-1

[Jain et al., 2013] Jain, P., P. Kumaraguru, et A. Joshi (2013). i Seek 'fb.me' : Identifying Users Across Multiple Online Social Networks. In www (Companion Volume), pp. 1259–1268.

[Jayanta, 2004] Jayanta Basak, A.S.a.M.S.S., Weather Data Mining Using Independent Component Analysis. Journal of Machine Learning Research 2004.

[Jin and Han, 2017] Jin X., Han J. (2017) K-Medoids Clustering. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA

[Kantardzic, 2003] M. Kantardzic. Data mining: concepts, models, methods, and algorithms. Wiley- Interscience, 2003.

[Kaynak et Hassan, 1994] Kaynak, E., & Hassan, S. (1994). Globalization of Consumer Markets: Structures and Strategies. Haworth Press. Knowledge discovery in databases: An overview. pages 1–27.

[Kaytoue et al., 2011] Kaytoue, M., Kuznetsov, S. O., and Napoli, A. (2011a). Biclustering numerical data in formal concept analysis. In Formal Concept Analysis, pages 135-150. Springer.

[Laporte, 2011] Laporte B. (2011), “Risk Management Systems: Using Data Mining in Developing Countries’ Customs Administrations”, World Customs Journal, vol. 5, 1, pp. 17-27.

[Lavrač et Džeroski, 1994] Lavrač, N. et Džeroski, S. (1994). Inductive Logic Programming : Techniques and Applications. Ellis Horwood.

[Le Floc’h, et al., 2003] Le Floc’h, A., Fisette, C., Missaoui, R., Valtchev, P., and Godin, R. JEN : un algorithme efficace de construction de générateurs pour l’identification des règles d’association. Revue ECA X, Y (2003), 1–Z.

[Li et al., 2011] Z. Li, J. Han, M. Ji, L.A. Tang, Y. Yu, B. Ding, J.G. Lee et R. Kays: Movemine : Mining moving object data for discovery of animal movement patterns. ACM Transactions on Intelligent Systems and Technology (TIST), 2(4):37 :1–37 :32, 2011.

[Luxenburger, 1991] Luxenburger M. (1991). "Implications partielles dans un contexte", Mathématiques informatique et sciences humaines, année 29, 113, 1991, p. 5-18.

[Mamoulis et al., 2004] N. Mamoulis, H. Cao, G. Kollios, M. Hadjieleftheriou, Y. Tao et D.W. Cheung : Mining, indexing and querying historical spatiotemporal data. ACM SIGKDD Conference on Know-ledge Discovery and DataMining, page 236, 2004.

[Marascu et Masseglia, 2006] A. Marascu et F. Masseglia: Mining sequential patterns from data streams : a centroid approach. Journal of Intelligent Information Systems, 27(3):291–307, 2006.

[Martin Ester et al., 2004] Martin Ester, R.G., Wen Jin, Zengjian Hu, A Microeconomic Data Mining Problem: Customer-Oriented Catalog Segmentation. KDD'04.

[Masseglia et al., 1998] Masseglia, F., Cathala, F., & Poncelet, P. (1998). The PSP approach for mining sequential patterns. Principles of Data Mining and Knowledge Discovery, pp. 176-184.

[Masseglia et al., 2008] F. Masseglia, P. Poncelet, M. Teisseire et A. Marascu : Web usage mining : extracting unexpected periods fromweb logs. DataMining and Knowledge Discovery (DMKD), 16(1):39– 65, 2008.

[Mata et al., 2002] Mata, J., Alvarez, J. L., & Riquelme, J. C. (2002). An evolutionary algorithm to discover numeric association rules. Proc. of the ACM Int. Conf. on Symposium on Applied computing, pp. 590-594.

[Maulika et Bandyopadhyay, 2000] Maulika, U., & Bandyopadhyay, S. (2000). Genetic algorithmbased clustering technique. Pattern Recognition, 33, 1455-1465.

[Mc Queen, 1967] Mc Queen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley symposium on mathematical statistics and probability, pp. 281-297.

[Miller et Yang, 1997] Miller, R. J., & Yang, Y. (1997). Association rules over interval data. Proc. of the Int. Conf. on Management of Data, pp. 452-461.

[Missaoui and Kwuida, 2011] Missaoui, R., and Kwuida, L. Mining Triadic Association Rules from Ternary Relations. In 9th International Conference ICFCA (May 2011), pp. 204–218.

[Mokeddem, 2016] Mokeddem, S.,.(2016), Fouille de données pour l'analyse des traces patients doctorat 3ème Cycle-L.M.D, Université d'Oran 1 Ahmed Benbella.

[Murillo, 2007] Murillo, L. M. (2007). Manufacturers-Retailers: The New Actor in the U.S. Furniture Industry. Characteristics and Implications for the Chinese Industry. Proceedings of world academy of science, Engineering and technology

[Narayanan et Shmatikov, 2009] Narayanan, A. et V. Shmatikov (2009). De-anonymizing Social Networks. In 30th IEEE Symposium on Security and Privacy, pp. 173–187. IEEE.

[Naulaerts et al., 2015] Naulaerts, S., Meysman, P., Bittremieux, W., Vu, T., Berghe,W. V., Goethals, B., and Laukens, K. (2015). A primer to frequent itemset mining for bioinformatics. Briefings in Bioinformatics, 16(2) :216–231.

[Nehmé et al., 2005] Nehmé, K., Valtchev, P., Rouane, M. H., and Godin, R. On Computing the Minimal Generator Family for Concept Lattices and Icebergs. Springer Berlin Heidelberg, 2005, pp. 192–207.

[Niharika et al., 2012] Niharika, S., Latha, V. S., and Lavanya, D. (2012). A survey on text categorization. International Journal of Computer Trends and Technology volume 3 Issue1-2012.

[Nortet et al., 1995] Nortet, C., Salleb, A., Turmeaux, T., &Vrain, C. (2006). Data Mining quantitative association.

[OMD, 2002] OMD, (2002). Convention internationale pour la simplification et la normalisation des procédures douanières (selon révision) (Convention de Kyoto révisée), Bruxelles, 2002.

[Park et al., 1995] J.S. Park, M-S. Chen, and P.S. Yu. An effective hash based algorithm for mining association rules. In Michael J. Carey and Donovan A. Schneider, editors, Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, pages 175-186, San Jose, California, 22-25 1995.

[Pasquier, 2000] Pasquier N., « Data Mining : algorithmes d'extraction et de réduction des règles d'association dans les bases de données », thèse d'université, université Blaise Pascal, Clermont-Ferrand, janvier 2000.

[Patze, 1995] Patze, G. L. (1995). Using Secondary Data in Marketing Research. Quorum Books.

[Pearson, 1904] Pearson, K. (1904). On the Theory of Contingency and its Relation to Association and Normal Correlation. Draper's Co. Res. Mem. Biometric Ser, 1, 1-35.

[Pei et al., 2000] J. Pei, J. Han, B. Mortazavi-Asl et H. Zhu : Mining access patterns efficiently from web logs. In Knowledge Discovery and DataMining, Current Issues and New Applications, 4th Pacific-Asia Conference, PADKK'00, pages 396–407. Springer, 2000.

[Pei et al., 2004] Pei, J., Han, J., Member, S., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., & Hsu, M. (2004). Mining sequential patterns by Pattern-Growth : The prexspan approach. IEEE Transactions on Knowledge and Data Engineering, 16, 1424-1

[Pennerath, 2009] Pennerath, F. (2009). Methodes d'extraction de connaissances à partir de données modélisables par des graphes. Application à des problèmes de synthèse organique. PhD thesis, Université de Nancy.

[Perera et al., 2009] D. Perera, J. Kay, I. Koprinska, K. Yacef et O.R. Zaïane : Clustering and sequential pattern mining of online collaborative learning data. IEEE Transactions on Knowledge and Data Engineering, 21:759–772, 2009. ISSN 1041-4347.

[Piatetsky-Shapiro, 1991] Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In Knowledge Discovery in Databases, pages 229–248.

[Plasse et al., 2008] Plasse Marie, Ndeye Niang, Saporta Gilbert, Villeminot Alexandre, Leblon Laurent. "Méthodes de classification pour l'extraction de règles", CNAM Laboratoire CÉDRIC, 2008

[Punj, 1983] Punj, G., & Stewart, D. W. (1983). Cluster Analysis in Marketing Research: Review and Suggestions for Application. Journal of Marketing Research, 20 (2), 134-148.

[Quinlan, 1986] Quinlan J. R., Induction of Decision Trees. Machine Learning, 1 :81–106, 1986. rules in a atherosclerosis dataset. Proc. of the Int. Conf. on PKDD, pp. 495-506.

[Raad et al., 2010] Raad, E., R. Chbeir, et A. Dipanda (2010). User Profile Matching in Social Networks. In Network-Based Information Systems (NBiS), 2010 13th International Conference on, pp. 297–304. IEEE.

[Rabatel et al., 2010] J. Rabatel, S. Bringay et P. Poncelet : Aide à la décision pour lamaintenance ferroviaire préventive. In Extraction et Gestion des Connaissances, EGC'10, pages 363–368. CépaduèsÉditions, 2010.

[Ruggieri et al., 2010] Ruggieri, S., D. Pedreschi, and F. Turini, Data mining for discrimination discovery. ACM Transactions on Knowledge Discovery from Data, 2010. 4(2): p. 1-40.

[Salle et al., 2009] P. Salle, S. Bringay et M. Teisseire : Mining discriminant sequential patterns for aging brain.In Artificial Intelligence in Medicine, volume 5651 de Lecture Notes in Computer Science, pages 365–369. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-02975-2.

[Salleb-Aouissi et al., 2013] Salleb-Aouissi, A., Vrain, C., Nortet, C., Kong, X., Rathod, V., &Cassard, D. (2013). Quantminer for mining quantitative association rules. Journal of Machine Learning Research, 14, 3153-3157.

[Salvador et Chan, 2004] Salvador, S., & Chan, P. (2004). Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms. Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence, (pp. 576-584).

[Savasere et al., 1995] A. Savasere, E. Omiecinski, and S.B. Navathe. An efficient algorithm for mining association rules in large databases. In VLDB '95: Proceedings of the 21th International Conference on Very Large Data Bases, pages 432- 444, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[Shouning et Peihua, 2005] Shouning Qu, CD., Peihua Liu, Research and Implementation on Data mining Applied to Learning Guidance System 2005.

[Simon, 2000] Simon A., Outils classificatoires par objets pour l'extraction de connaissances dans les bases de données. Thèse de doctorat de l'université Henri Poincaré Nancy 1, Nancy, 2000.

[Singh, 1990] Singh, J. (1990). A typology of consumer response styles. Journal of Retailing, 66 (1), 57-99.

[Smith, 1956] Smith, W. (1956). Product differentiation and Market Segmentation as Alternative Marketing Strategies. Journal of marketing, 3-8.

[Srikant et Agrawal, 1997] Srikant R., Vu Q. et Agrawal R. (1997). Mining association rules with item constraints, Proc. of the 3rd Int'l Conf. on Knowledge Discovery in Databases and Data Mining, 1997.

[Sung, 2004] Sung, H. (2004). Classification of Adventure Travelers:Behavior, Decision Making,and Target Markets. Journal of Travel Research, 42, 343-356.

[Stewart and White, 1991] B.S. Stewart and C.C. White. Multiobjective. Journal of the ACM, 38(4) :775-814, 1991.

[Tan et al., 2006] Tan, P. -N., Steinbach, M., Kumar, V. et al. (2006). Introduction ta data mining, volume 1. Pearson Addi on Wesley Boston.

[Toivonen, 1996] H. Toivonen. Sampling large databases for association rules. In T. M. Vijayaraman, Alejandro P. Buchmann, C. Mohan, and Nandlal L. Barda, editors, In Proc. 1996 Int. Conf. Very Large Data Bases, pages 134-145. Morgan Kaufman, 09 1996.

[Umesh, 1987] Umesh, U. (1987). Transferability of preference models across segments and geographic areas. Journal of Marketing, 51, 59-70.

[Usama et Padhraic, 1996] Usama Fayyad, G.P.-S., and Padhraic Smyth, From Data Mining to Knowledge Discovery in Databases. AI Magazine, 1996. 17(3): p. 37-54.

[Vladimir N, 98] Vladimir N. Vapnik. Statistical learning theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.

[Vladimir, 82] Vladimir Vapnik. Estimation of dependences based on empirical data. Springer Series in Statistics. Springer-Verlag, New York, 1982. Translated from the Russian by Samuel Kotz.

[Walsh, 2003] Walsh J. T. (2003), « Vérification et audit a posteriori » in M. Keen (éd.), Moder- niser la douane : défis et stratégies de réforme de l'administration douanière, Washington, D.C., FMI.

[Wang et al. 2004] K. Wang, Y. Xu et J.X. Yu : Scalable sequential pattern mining for biological sequences. In CIKM '04 : Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management, pages 178–187, New York, NY, USA, 2004b. ACM. ISBN 1-58113-874-1.

[Wang et al., 2000] Wang W., Yang J. et Yu P. (2000). Efficient mining of weighted association rules (WAR), Proc. of the sixth ACM SIGKDD int'l conf. on Knowledge discovery and data mining, pp 270-274.

[Wang et al., 2012] Wang, Xin, Liu; Xiaodong, Pedrycz, Witold and Zhu, Xiaolei, Hu, Guangfei. "Mining axiomatic fuzzy set association rules for classification problems. ", European Journal of Operational Research, 2012.

[Washio et Motoda, 2003] Washio, T. et Motoda, H. (2003). State of the art of graph-based data mining. ACM SIGKDD Explorations Newsletter, 5(1):59.

[Webb, 2001] Webb, G. I. (2001). Discovering associations with numeric variables. Proc. of the ACM Int. Conf. on SIGKDD, pp. 383-388.

[Wedel et Kamakura, 1998] Wedel, M., & Kamakura, W. A. (1998). Market Segmentation: Conceptual and Methodological Foundation. Massachusetts: Kluwer Acadamic Publishers.

[Widdowson, 2005] Widdowson D. (2005), « Managing Risk in the Customs Context » in L. De Wulf and J. B. Sokol (eds), Customs Modernization Handbook, Washington D.C., World Bank.

[Wille, 1982] Wille, R. (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. Ordered sets, pages 445–470.

[Wind, 1978] Wind, Y. (1978). Issues and Advances in Segmentation Research. Journal of Marketing Research, 15 (3), 317-337.

[Zak, 1999] M.J. Zaki. Parallel and distributed association mining: A survey. IEEE Concurrency, 7(4) :14-25, / 1999.

[Zaki et Hsiao, 2005] Zaki, M. J., & Hsiao, C.-J. (2005). Efficient algorithms for mining closed itemsets and theirlattice structure. IEEE Transactions on Knowledge and Data Engineering, 17, 462-478.

[Zaki, 2001] Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. Machine Learning Journal, special issue on Unsupervised Learning, 42, 31-60.

[Zehero Bi et al., 2018] Zehero Bi Bolou, Etienne Soro, Yake Gondo, Pacôme Brou, Olivier Asseu. (2018), Elicitation of Association Rules from Information on Customs Offences on the Basis of Frequent Motives, Engineering, Vol.10 No9, September 2018.

ANNEXE A : Procès-Verbal Simplifié (PVS)

MINISTERE AUPRES DU PREMIER MINISTRE
CHARGE DE L'ECONOMIE ET DES FINANCES

REPUBLIQUE DE COTE D'IVOIRE
Union - Discipline - Travail

Direction Générale des Douanes

CIRCULAIRE N°16 15 /MPMEF/DGD/DU 21 JUIN 2013

(DIFFUSION GENERALE)

OBJET : Informations sur la mise en place du Procès-verbal Simplifié

Réf : - Note de service n° 85 du 11 juillet 2011

J'ai l'honneur de faire connaitre à l'ensemble du service et des usagers que, dans le cadre de l'adoption d'une approche des contrôles douaniers basés sur l'analyse du risque, il a été mis en place une base de données des infractions douanières à partir du Procès-verbal Simplifié (PVS). Le PVS permet :

- l'information des procès-verbaux (PV) ;
- la simplification de la rédaction des PV ;
- la centralisation de tous les procès verbaux ;
- les statistiques des procès verbaux ;
- la gestion des amendes ;
- l'évaluation des agents et des services en matière de contentieux ;
- la sélectivité des circuits des déclarations ;
- l'analyse de risque sur la base de l'historique des déclarations.

En conséquence, tout opérateur économique ou commissionnaire en douane agréé peut, désormais, accéder à la plate forme du PVS sur l'intranet de la Douane à l'adresse **app.douanes.ci** pour consulter les Procès-verbaux (PV) dressés à son encontre par les services des douanes.

LE DIRECTEUR GENERAL DES DOUANES

DIRECTION GENERALE DES DOUANES
Le Directeur Général

Col-Maj. Issa COULIBALY

Ampliations :
- MPMEF/Cab
- DG Economie
- FEDERMAR
- GEPEX
- CGECI
- Conseil du Café-Cacao
- Chbre Cce & Industrie
- PAA
- PASP
- OIC
- Synd. des Trans. s/c BOLLORE
- Synd. Nat. Des Transitaires
- Toutes Directions Douanes

ANNEXE B: Dictionnaire des Données et table du PVS (SYDAM World)

SW_PVS_ACTE_MATERIEL				
Description colonne	**Table**	**Champ BD**	**Type**	**Taille**
	SW_PVS_ACTE_MATERIEL	INSTANCE_ID	NUMBER	10
Numéro acte matériel saisi	SW_PVS_ACTE_MATERIEL	ACM_RNK	NUMBER	10
Code exportateur	SW_PVS_ACTE_MATERIEL	ACM_COD	VARCHAR2	10
Nom exportateur	SW_PVS_ACTE_MATERIEL	ACM_DSC	VARCHAR2	200

SW_PVS_DOCUMENT_REQUIS				
Description colonne	**Table**	**Champ BD**	**Type**	**Taille**
	SW_PVS_DOCUMENT_REQUIS	INSTANCE_ID	NUMBER	10
Numéro acte matériel saisi	SW_PVS_DOCUMENT_REQUIS	DOR_RNK	NUMBER	10
Code exportateur	SW_PVS_DOCUMENT_REQUIS	DOR_TYP_COD	VARCHAR2	10
Nom exportateur	SW_PVS_DOCUMENT_REQUIS	DOR_TYP_DSC	VARCHAR2	200
Date restitution	SW_PVS_DOCUMENT_REQUIS	DOR_RES_DAT	DATE	

SW_PVS_EVALUATION				
Description colonne	**Table**	**Champ dans la BD**	**Type**	**Taille**
	SW_PVS_EVALUATION	INSTANCE_ID	NUMBER	10
Numéro Évaluation saisi	SW_PVS_EVALUATION	EVAL_RNK	NUMBER	10
Matricule Agent Évaluation	SW_PVS_EVALUATION	EVAL_AGT_MAT	VARCHAR2	10
Nom Agent Evaluation	SW_PVS_EVALUATION	EVAL_AGT_NOM	VARCHAR2	20
Prénom Agent Evaluation	SW_PVS_EVALUATION	EVAL_AGT_PREN	VARCHAR2	20
Photo Agent Evaluation	SW_PVS_EVALUATION	EVAL_AGT_PHO	VARCHAR2	100
Motif Evaluation	SW_PVS_EVALUATION	EVAL_MOTIF	VARCHAR2	500
Montant Evaluation	SW_PVS_EVALUATION	EVAL_MNT	NUMBER	10

SW_PVS_MARCH				
Description colonne	**Table**	**Champ dans la BD**	**Type**	**Taille**
	SW_PVS_MARCH	INSTANCE_ID	NUMBER	10
Numéro marchandise saisie	SW_PVS_MARCH	ITM_RNK	NUMBER	10
Marchandise Prohibé Statut	SW_PVS_MARCH	ITM_MSE_PRO_FLG	VARCHAR2	50
Code pays origine	SW_PVS_MARCH	ITM_MSE_PAY_ORG	VARCHAR2	10
Code pays provenance	SW_PVS_MARCH	ITM_MSE_PAY_PRV	VARCHAR2	10

Unité	SW_PVS_MARCH	ITM_MSE_UNI	VARCHAR2	10
Quantité	SW_PVS_MARCH	ITM_MSE_QTE	NUMBER	30
Numéro réquisition	SW_PVS_MARCH	ITM_MSE_REQ_NUM	VARCHAR2	10
Date Réquisition	SW_PVS_MARCH	ITM_MSE_REQ_DAT	DATE	
Date Saisie	SW_PVS_MARCH	ITM_MSE_SAI_DAT	DATE	
Code SH marchandise déclarée	SW_PVS_MARCH	ITM_MSE_DEC_COD	VARCHAR2	10
Description SH marchandise déclarée sur 8	SW_PVS_MARCH	ITM_MSE_DEC_DSC	VARCHAR2	200
Marchandise Commerciale déclarée	SW_PVS_MARCH	ITM_MSE_DEC_CLE	VARCHAR2	100
Code SH Marchandise Reconnue	SW_PVS_MARCH	ITM_MSE_REC_COD	VARCHAR2	10
Description SH Marchandise reconnue	SW_PVS_MARCH	ITM_MSE_REC_DSC	VARCHAR2	200
Marchandise Commerciale Reconnue	SW_PVS_MARCH	ITM_MSE_REC_CLE	VARCHAR2	100
Code Titre Infraction	SW_PVS_MARCH	ITM_MSE_INF_GEN_TIT_COD	VARCHAR2	10
Description Titre Infraction	SW_PVS_MARCH	ITM_MSE_INF_GEN_TIT_DSC	VARCHAR2	200
Code Nature Infraction	SW_PVS_MARCH	ITM_MSE_INF_GEN_NAT_COD	VARCHAR2	10
Description Nature Infraction	SW_PVS_MARCH	ITM_MSE_INF_GEN_NAT_DSC	VARCHAR2	200
Code Infraction	SW_PVS_MARCH	ITM_MSE_INF_GEN_COD	VARCHAR2	10
Description Infraction	SW_PVS_MARCH	ITM_MSE_INF_GEN_DSC	VARCHAR2	200
Droit et Taxe Déclarée	SW_PVS_MARCH	ITM_MSE_DT_DEC	NUMBER	15
Droit et Taxe Reconnue	SW_PVS_MARCH	ITM_MSE_DT_REC	NUMBER	15
Droit Compromis	SW_PVS_MARCH	ITM_MSE_DC	NUMBER	15
Valeur FOB déclarée	SW_PVS_MARCH	ITM_MSE_FOB_DEC	NUMBER	15
Valeur FOB Reconnue	SW_PVS_MARCH	ITM_MSE_FOB_REC	NUMBER	15
Différence de Valeur FOB	SW_PVS_MARCH	ITM_MSE_FOB_DIFF	NUMBER	15
Valeur CAF déclarée	SW_PVS_MARCH	ITM_MSE_CAF_DEC	NUMBER	15
Valeur CAF Reconnue	SW_PVS_MARCH	ITM_MSE_CAF_REC	NUMBER	15
Différence de Valeur CAF	SW_PVS_MARCH	ITM_MSE_CAF_DIF	NUMBER	15
Valeur Marché Intérieur	SW_PVS_MARCH	ITM_MSE_VMI	NUMBER	12

SW_PVS_PREVENU

Description colonne	Table	Champ dans la BD	Type	Taille
	SW_PVS_PREVENU	INSTANCE_ID	NUMBER	10
Numéro Prévenu	SW_PVS_PREVENU	PREV_RNK	NUMBER	10
Nom Prévenu	SW_PVS_PREVENU	PREV_NOM	VARCHAR2	20
Prénoms Prévenu	SW_PVS_PREVENU	PREV_PREN	VARCHAR2	50
Sexe Prévenu	SW_PVS_PREVENU	PREV_SEX	VARCHAR2	10
Code Prévenu Nationalité	SW_PVS_PREVENU	PREV_NAT_COD	VARCHAR2	20
Description Prévenu Nationalité	SW_PVS_PREVENU	PREV_NAT_DSC	VARCHAR2	200
Prévenu Filiation	SW_PVS_PREVENU	PREV_FIL	VARCHAR2	10
Date Naissance Prévenu	SW_PVS_PREVENU	PREV_DAT_NAIS	DATE	
Lieu Naissance Prévenu	SW_PVS_PREVENU	PREV_LIEU_NAIS	VARCHAR2	10
Profession Prévenu	SW_PVS_PREVENU	PREV_PROF	VARCHAR2	10
Libellé narratif	SW_PVS_PREVENU	PREV_NAR	VARCHAR2	4000
Libellé date	SW_PVS_PREVENU	PREV_DAT_NAR	DATE	
Numéro Personne Morale	SW_PVS_PREVENU	PREV_IMP_COD	VARCHAR2	10
Libellé Personne Morale	SW_PVS_PREVENU	PREV_IMP_NAM	VARCHAR2	50
Numéro déclarant	SW_PVS_PREVENU	PREV_DEC_COD	VARCHAR2	10
Libellé déclarant	SW_PVS_PREVENU	PREV_DEC_NAM	VARCHAR2	50

SW_PVS_RECOUVRE				
Description colonne	**Table**	**Champ dans la BD**	**Type**	**Taille**
	SW_PVS_RECOUVRE	INSTANCE_ID	NUMBER	10
Numéro Recouvrement	SW_PVS_RECOUVRE	REC_RNK	NUMBER	10
Code Bureau Recouvrement	SW_PVS_RECOUVRE	REC_CUO_COD	VARCHAR2	10
Libellé Bureau Recouvrement	SW_PVS_RECOUVRE	REC_CUO_NAM	VARCHAR2	50
Numéro de Série Recouvrement	SW_PVS_RECOUVRE	REC_SER	VARCHAR2	10
Numéro Incrémental Recouvrement	SW_PVS_RECOUVRE	REC_NBR	NUMBER	10
Date Recouvrement	SW_PVS_RECOUVRE	REC_DAT	DATE	
Montant Recouvrement	SW_PVS_RECOUVRE	REC_AMT	NUMBER	10
Lieu de Naissance Prévenu	SW_PVS_RECOUVRE	PREV_LIEU_NAIS	VARCHAR2	10
Profession Prévenu	SW_PVS_RECOUVRE	PREV_PROF	VARCHAR2	10
Narratif Prévenu	SW_PVS_RECOUVRE	PREV_NAR	VARCHAR2	4000
Date Narratif Prévenu	SW_PVS_RECOUVRE	PREV_DAT_NAR	DATE	
Numéro Prévenu Morale	SW_PVS_RECOUVRE	PREV_IMP_COD	VARCHAR2	10

Libellé Prévenu Morale	SW_PVS_RECOUVRE	PREV_IMP_NAM	VARCHAR2	50
Numéro Prévenu déclaré	SW_PVS_RECOUVRE	PREV_DEC_COD	VARCHAR2	10
Libellé Prévenu déclaré	SW_PVS_RECOUVRE	PREV_DEC_NAM	VARCHAR2	50

SW_PVS_SAISISSANT

Description colonne	Table	Champ dans la BD	Type	Taille
	SW_PVS_SAISISSANT	INSTANCE_ID	NUMBER	10
Numéro Saisissant	SW_PVS_SAISISSANT	SAI_RNK	NUMBER	10
Numéro Matricule Agent Saisissant	SW_PVS_SAISISSANT	SAI_AGT_MAT	VARCHAR2	10
Nom Agent Saisissant	SW_PVS_SAISISSANT	SAI_AGT_NOM	VARCHAR2	10
Prénoms Agent Saisissant	SW_PVS_SAISISSANT	SAI_AGT_PREN	VARCHAR2	50

SW_PVS_SERIAL_REG_TAB

Description colonne	Table	Champ dans la BD	Type	Taille
	SW_PVS_SERIAL_REG_TAB	INSTANCE_ID	NUMBER	10
Année	SW_PVS_SERIAL_REG_TAB	YER	NUMBER	10
Code Site	SW_PVS_SERIAL_REG_TAB	SIT_COD	VARCHAR2	5
Libellé Site	SW_PVS_SERIAL_REG_TAB	SIT_NAM	VARCHAR2	35
	SW_PVS_SERIAL_REG_TAB	CHR	VARCHAR2	1
Service Initial	SW_PVS_SERIAL_REG_TAB	SER_INI	NUMBER	10
Service Actuel	SW_PVS_SERIAL_REG_TAB	SER_CUR	NUMBER	10

SW_PVS_TAB

Description colonne	Table	Champ dans la BD	Type	Taille
	SW_PVS_TAB	INSTANCE_ID	NUMBER	10
Code Type PVS	SW_PVS_TAB	PVS_TYP_COD	VARCHAR2	5
Description du Type PVS	SW_PVS_TAB	PVS_TYP_DSC	VARCHAR2	200
Numéro AC	SW_PVS_TAB	PVS_AC_NUM	VARCHAR2	10
Code Type AC	SW_PVS_TAB	PVS_AC_TYP_COD	VARCHAR2	5
Description Type AC	SW_PVS_TAB	PVS_AC_TYP_DSC	VARCHAR2	200
Date AC	SW_PVS_TAB	PVS_AC_DAT	DATE	

Code mode de contrôle	SW_PVS_TAB	PVS_MOD_CTL_COD	VARCHAR2	10
Description mode de contrôle	SW_PVS_TAB	PVS_MOD_CTL_DSC	VARCHAR2	200
Date début contrôle	SW_PVS_TAB	PVS_CTL_DEB	DATE	
Date fin contrôle	SW_PVS_TAB	PVS_CTL_FIN	DATE	
Heure début de contrôle	SW_PVS_TAB	PVS_CTL_HEU_DEB	NUMBER	10
Heure fin de contrôle	SW_PVS_TAB	PVS_CTL_HEU_FIN	NUMBER	10
Description mode opératoire	SW_PVS_TAB	PVS_CTL_NAR	VARCHAR2	4000
Cachette	SW_PVS_TAB	PVS_CTL_CAC	VARCHAR2	50
Itinéraire	SW_PVS_TAB	PVS_CTL_ITI	VARCHAR2	50
Code Control Niveau1	SW_PVS_TAB	PVS_CTL_NIV1_COD	VARCHAR2	20
Description Control Niveau1	SW_PVS_TAB	PVS_CTL_NIV1_DSC	VARCHAR2	200
Code Control Niveau2	SW_PVS_TAB	PVS_CTL_NIV2_COD	VARCHAR2	20
Description Control Niveau2	SW_PVS_TAB	PVS_CTL_NIV2_DSC	VARCHAR2	200
Code Control Niveau3	SW_PVS_TAB	PVS_CTL_NIV3_COD	VARCHAR2	20
Description Control Niveau3	SW_PVS_TAB	PVS_CTL_NIV3_DSC	VARCHAR2	200
Code Control Niveau4	SW_PVS_TAB	PVS_CTL_NIV4_COD	VARCHAR2	20
Description Control Niveau4	SW_PVS_TAB	PVS_CTL_NIV4_DSC	VARCHAR2	200
Code Control Niveau5	SW_PVS_TAB	PVS_CTL_NIV5_COD	VARCHAR2	20
Description Control Niveau5	SW_PVS_TAB	PVS_CTL_NIV5_DSC	VARCHAR2	200
Code titre infraction	SW_PVS_TAB	PVS_INF_GEN_TIT_COD	VARCHAR2	10
Description titre infraction	SW_PVS_TAB	PVS_INF_GEN_TIT_DSC	VARCHAR2	200
Code nature infraction	SW_PVS_TAB	PVS_INF_GEN_NAT_COD	VARCHAR2	10
Description nature infraction	SW_PVS_TAB	PVS_INF_GEN_NAT_DSC	VARCHAR2	200
Code infraction	SW_PVS_TAB	PVS_INF_GEN_COD	VARCHAR2	10
Description infraction	SW_PVS_TAB	PVS_INF_GEN_DSC	VARCHAR2	200
Lieu de Saisie	SW_PVS_TAB	PVS_SAI_LIEU	VARCHAR2	100
Numéro Ordre de service	SW_PVS_TAB	PVS_ORD_NUM	VARCHAR2	10
Date Ordre de service	SW_PVS_TAB	PVS_ORD_DAT	DATE	
Narratif Indicateur	SW_PVS_TAB	PVS_NAR_IND	VARCHAR2	4000
Date narratif indicateur	SW_PVS_TAB	PVS_NAR_DTE	DATE	
Narratif Indicateur	SW_PVS_TAB	PVS_NAR_IND	VARCHAR2	5

Description type de visite	SW_PVS_TAB	PVS_TYP_VIS_DSC	VARCHAR2	200
Description Marchandise prohibe	SW_PVS_TAB	PVS_MSE_PRO_DSC	VARCHAR2	50
Code Marchandise prohibe Niveau1	SW_PVS_TAB	PVS_PRO_NIV1_COD	VARCHAR2	20
Description Marchandise prohibe Niveau1	SW_PVS_TAB	PVS_PRO_NIV1_DSC	VARCHAR2	200
Code Marchandise prohibe Niveau2	SW_PVS_TAB	PVS_PRO_NIV2_COD	VARCHAR2	20
Description Marchandise prohibe Niveau2	SW_PVS_TAB	PVS_PRO_NIV2_DSC	VARCHAR2	200
Code Marchandise prohibe Niveau3	SW_PVS_TAB	PVS_PRO_NIV3_COD	VARCHAR2	20
Description Marchandise prohibe Niveau3	SW_PVS_TAB	PVS_PRO_NIV3_DSC	VARCHAR2	200
Code Marchandise prohibe Niveau4	SW_PVS_TAB	PVS_PRO_NIV4_COD	VARCHAR2	20
Description Marchandise prohibe Niveau4	SW_PVS_TAB	PVS_PRO_NIV4_DSC	VARCHAR2	200
Code Marchandise prohibe Niveau5	SW_PVS_TAB	PVS_PRO_NIV5_COD	VARCHAR2	20
Nom du dépositaire	SW_PVS_TAB	PVS_DEP_NOM	VARCHAR2	20
Prénom du dépositaire	SW_PVS_TAB	PVS_DEP_PREN	VARCHAR2	50
Date de naissance du dépositaire	SW_PVS_TAB	PVS_DEP_NAIS_DAT	DATE	
Lieu de naissance du dépositaire	SW_PVS_TAB	PVS_DEP_NAIS_LIEU	VARCHAR2	10
Adresse du dépositaire	SW_PVS_TAB	PVS_DEP_ADR	VARCHAR2	10
Lieu d'habitation du dépositaire	SW_PVS_TAB	PVS_DEP_HAB	VARCHAR2	50
Filiation du dépositaire	SW_PVS_TAB	PVS_DEP_FIL	VARCHAR2	10
Nationalité du dépositaire	SW_PVS_TAB	PVS_DEP_NAT	VARCHAR2	10
Frais engagé	SW_PVS_TAB	PVS_FRAI_ENG	NUMBER	15
Droit déclaré	SW_PVS_TAB	VS_DT_DEC	NUMBER	15
Droit reconnu	SW_PVS_TAB	PVS_DT_REC	NUMBER	15
Droit Compromis	SW_PVS_TAB	PVS_DT_DC	NUMBER	15
Valeur FOB déclarée	SW_PVS_TAB	PVS_FOB_DEC	NUMBER	15
Valeur FOB reconnue	SW_PVS_TAB	PVS_FOB_REC	NUMBER	15
Différence de valeur FOB	SW_PVS_TAB	PVS_FOB_DIFF	NUMBER	15
Valeur CAF déclarée	SW_PVS_TAB	PVS_CAF_DEC	NUMBER	15
Valeur CAF reconnue	SW_PVS_TAB	PVS_CAF_REC	NUMBER	15

Différence de valeur CAF	SW_PVS_TAB	PVS_CAF_DIFF	NUMBER	15
Valeur marché intérieur	SW_PVS_TAB	PVS_VMI	NUMBER	15
Amende légale	SW_PVS_TAB	PVS_AMD_LEG	NUMBER	15
Amende réelle	SW_PVS_TAB	PVS_AMD_REL	NUMBER	15
Nombre Total de marchandise	SW_PVS_TAB	NBR_ITM_TOT	NUMBER	15
Nombre impression	SW_PVS_TAB	PVS_IMP_NBR	NUMBER	15
Date début de la saisie	SW_PVS_TAB	PVS_SYS_DEB	DATE	
Date fin de la saisie	SW_PVS_TAB	PVS_SYS_FIN	DATE	
Heure début de la saisie	SW_PVS_TAB	PVS_SYS_HDE	NUMBER	10
Heure fin de la saisie	SW_PVS_TAB	PVS_SYS_HFI	NUMBER	10
Année	SW_PVS_TAB	PVS_YER	NUMBER	4
Code Bureau	SW_PVS_TAB	PVS_CUO_COD	VARCHAR2	5
Nom Bureau	SW_PVS_TAB	PVS_CUO_NAM	VARCHAR2	50
Numéro Incrémental	SW_PVS_TAB	PVS_INC	NUMBER	10
Numéro de Série	SW_PVS_TAB	PVS_SER	VARCHAR2	1
Date PVS	SW_PVS_TAB	PVS_DAT	DATE	
Dernier statut	SW_PVS_TAB	PVS_FLG	VARCHAR2	10
Numéro Matricule Agent	SW_PVS_TAB	PVS_AGT_NUM_MAT	VARCHAR2	30
Nom Agent	SW_PVS_TAB	PVS_AGT_NOM	VARCHAR2	50
Prénom Agent	SW_PVS_TAB	PVS_AGT_PREN	VARCHAR2	200
Cellulaire1 Agent	SW_PVS_TAB	PVS_AGT_CEL1	VARCHAR2	5
Cellulaire2 Agent	SW_PVS_TAB	PVS_AGT_CEL2	VARCHAR2	5
Mail Agent	SW_PVS_TAB	PVS_AGT_MAIL	VARCHAR2	10
Photo Agent	SW_PVS_TAB	PVS_AGT_PHO	VARCHAR2	100
Direction Code	SW_PVS_TAB	PVS_DIR_COD	VARCHAR2	60
Sous-Direction Code	SW_PVS_TAB	PVS_SDIR_COD	VARCHAR2	60
Code mode transport	SW_PVS_TAB	PVS_TRN_MOD_COD	VARCHAR2	10
Description mode transport	SW_PVS_TAB	PVS_TRN_MOD_DSC	VARCHAR2	200
Amende	SW_PVS_TAB	PVS_TRN_MOD_AMD	VARCHAR2	10
Confiscation	SW_PVS_TAB	PVS_TRN_MOD_CONF	VARCHAR2	50
Titre de transport	SW_PVS_TAB	PVS_TRN_MOD_TIT	VARCHAR2	50
Marque véhicule	SW_PVS_TAB	PVS_VEH_MRQ	VARCHAR2	50
Poids total en charge véhicule	SW_PVS_TAB	PVS_VEH_PTC	NUMBER	10
Immatriculation véhicule	SW_PVS_TAB	PVS_VEH_IMM	VARCHAR2	50

Lieu de chargement	SW_PVS_TAB	PVS_VEH_CHR	VARCHAR2	50
Code Mode Niveau1	SW_PVS_TAB	PVS_MOD_NIV1_COD	VARCHAR2	20
Description Mode Niveau1	SW_PVS_TAB	PVS_MOD_NIV1_DSC	VARCHAR2	200
Code Mode Niveau2	SW_PVS_TAB	PVS_MOD_NIV2_COD	VARCHAR2	20
Description Mode Niveau2	SW_PVS_TAB	PVS_MOD_NIV2_DSC	VARCHAR2	200
Code Mode Niveau3	SW_PVS_TAB	PVS_MOD_NIV3_COD	VARCHAR2	20
Description Mode Niveau3	SW_PVS_TAB	PVS_MOD_NIV3_DSC	VARCHAR2	200
Code Mode Niveau4	SW_PVS_TAB	PVS_MOD_NIV4_COD	VARCHAR2	20
Description Mode Niveau4	SW_PVS_TAB	PVS_MOD_NIV4_DSC	VARCHAR2	200
Code Mode Niveau5	SW_PVS_TAB	PVS_MOD_NIV5_COD	VARCHAR2	20
Nom chauffeur véhicule	SW_PVS_TAB	PVS_VEH_CHA_NOM	VARCHAR2	20
Prénom chauffeur véhicule	SW_PVS_TAB	PVS_VEH_CHA_PREN	VARCHAR2	50
Date de naissance chauffeur véhicule	SW_PVS_TAB	PVS_VEH_CHA_NAIS_DAT	DATE	
Lieu de naissance chauffeur véhicule	SW_PVS_TAB	PVS_VEH_CHA_NAIS_LIEU	VARCHAR2	50
Filiation chauffeur véhicule	SW_PVS_TAB	PVS_VEH_CHA_FIL	VARCHAR2	10
Code nationalité chauffeur véhicule	SW_PVS_TAB	PVS_VEH_CHA_NAT_COD	VARCHAR2	2
Libelle nationalité chauffeur véhicule	SW_PVS_TAB	PVS_VEH_CHA_NAT_DSC	VARCHAR2	200
Caution main levée du véhicule	SW_PVS_TAB	PVS_VEH_CAU	NUMBER	15
Permis du chauffeur véhicule	SW_PVS_TAB	PVS_VEH_CHA_PER	VARCHAR2	50
Nombre total de prévenu	SW_PVS_TAB	PVS_NBR_TOT_PREV	VARCHAR2	50
Itinéraire	SW_PVS_TAB	PVS_ITI	VARCHAR2	500
Flag CEN	SW_PVS_TAB	PVS_CEN	VARCHAR2	1
Identifiant	SW_PVS_TAB	PVS_ORD_IDE	VARCHAR2	10
Adresse service	SW_PVS_TAB	PVS_ORD_ADR	VARCHAR2	10
Dernier statut Reconnu	SW_PVS_TAB	PVS_FLG_REC	VARCHAR2	1
Code des douanes nationales	SW_PVS_TAB	PVS_BAS_NAT	VARCHAR2	100
Code des douanes communautaires	SW_PVS_TAB	PVS_BAS_COM	VARCHAR2	100
Autres	SW_PVS_TAB	PVS_BAS_AUT	VARCHAR2	100
Code pénalité	SW_PVS_TAB	PVS_PEN_COD	VARCHAR2	50

Description pénalité	SW_PVS_TAB	PVS_PEN_DSC	VARCHAR2	100

ANNEXE C: Publication Scientifique #1

"*Elicitation of Association Rules from Information on Customs Offences on the Basis of Frequent Motives*"

ABSTRACT :

The fight against fraud and trafficking is a fundamental mission of customs. The conditions for carrying out this mission depend both on the evolution of economic issues and on the behavior of the actors in charge of its implementation. As part of the customs clearance process, customs are nowadays confronted with an increasing volume of goods in connection with the development of international trade. Automated risk management is therefore required to limit intrusive control. In this article, we propose an unsupervised classification method to extract knowledge rules from a database of customs offences in order to identify abnormal behavior resulting from customs control. The idea is to apply the Apriori principle based on frequent grounds on a database relating to customs offences in customs procedures to uncover potential rules of association between a customs operation and an offence for the purpose of extracting knowledge governing the occurrence of fraud. This mass of often heterogeneous and complex data thus generates new needs that knowledge extraction methods must be able to meet. The assessment of infringements inevitably requires a proper identification of the risks. It is an original approach based on data mining or data mining to build association rules in two steps: first, search for frequent patterns (support >= minimum support) then from the frequent patterns, produce association rules (Trust >= Minimum Trust). The simulations carried out highlighted three main association rules: forecasting rules, targeting rules and neutral rules with the introduction of a third indicator of rule relevance which is the Lift measure. Confidence in the first two rules has been set at least 50%.

Key word:
Data Mining, Customs Offences, Unsupervised Method, Principle of Apriori, Frequent Motive, Rule of Association, Extraction of Knowledge.

Scientific Journal: *Engineering, September 14, 2018*

Zehero, B.B., *Soro,E., Gondo, Y., Brou, P. and Asseu, O. (2018) Elicitation of Association Rules from Information on Customs Offences on the Basis of Frequent Motives. Engineering, 10, 588-605.*
https://doi.org/10.4236/eng.2018.109043

Scientific Research Publishing

Engineering, 2018, 10, 588-605
http://www.scirp.org/journal/eng
ISSN Online: 1947-394X
ISSN Print: 1947-3931

Elicitation of Association Rules from Information on Customs Offences on the Basis of Frequent Motives

Bi Bolou Zehero[1*], Etienne Soro[2,3], Yake Gondo[3], Pacôme Brou[2*], Olivier Asseu[1,2*]

[1]Institut National Polytechnique—Houphouët Boigny, Yamoussoukro, Côte d'Ivoire
[2]Ecole Supérieure Africaine des TIC-ESATIC, Abidjan-Treichville, Côte d'Ivoire
[3]Université Felix Houphouët Boigny, Abidjan-Cocody, Côte d'Ivoire
Email: *oasseu@yahoo.fr, *zeherobi@yahoo.fr, *broupacom@hotmail.fr

How to cite this paper: Zehero, B.B., Soro, E., Gondo, Y., Brou, P. and Asseu, O. (2018) Elicitation of Association Rules from Information on Customs Offences on the Basis of Frequent Motives. *Engineering*, **10**, 588-605.
https://doi.org/10.4236/eng.2018.109043

Received: August 16, 2018
Accepted: September 11, 2018
Published: September 14, 2018

Abstract

The fight against fraud and trafficking is a fundamental mission of customs. The conditions for carrying out this mission depend both on the evolution of economic issues and on the behaviour of the actors in charge of its implementation. As part of the customs clearance process, customs are nowadays confronted with an increasing volume of goods in connection with the development of international trade. Automated risk management is therefore required to limit intrusive control. In this article, we propose an unsupervised classification method to extract knowledge rules from a database of customs offences in order to identify abnormal behaviour resulting from customs control. The idea is to apply the Apriori principle on the basis of frequent grounds on a database relating to customs offences in customs procedures to uncover potential rules of association between a customs operation and an offence for the purpose of extracting knowledge governing the occurrence of fraud. This mass of often heterogeneous and complex data thus generates new needs that knowledge extraction methods must be able to meet. The assessment of infringements inevitably requires a proper identification of the risks. It is an original approach based on data mining or data mining to build association rules in two steps: first, search for frequent patterns (support >= minimum support) then from the frequent patterns, produce association rules (Trust >= Minimum Trust). The simulations carried out highlighted three main association rules: forecasting rules, targeting rules and neutral rules with the introduction of a third indicator of rule relevance which is the Lift measure. Confidence in the first two rules has been set at least 50%.

Keywords

Data Mining, Customs Offences, Unsupervised Method, Principle of Apriori,

Frequent Motive, Rule of Association, Extraction of Knowledge

1. Introduction

The mobilization of customs revenue in developing countries constitutes both in terms of the balance of public finances and in terms of poverty reduction. Due to the context of reduced customs revenue base resulting from economic integration, free movement, tariff dismantling processes, economic partnership agreements and large-scale fraud, customs in the context of revenue mobilization need to use robust risk analysis and management methods for effective customs control. Whether it seems a long time ago, the management of procedures in the customs administration relied essentially on manual counting in order to detect offences due to fraud. Given the exponential volume of global trade, the most modern customs administrations rely on the technological development of digital data collection devices to store very large amounts of data for fraud risk analysis. This system (risk analysis) is then frequently used for research, evaluation and planning for other purposes in terms of analysis and forecasting of infringements in customs administrations. According to Harrison, it is an effective means of combating intrusive controls that meet the requirements of private operators to secure their transactions [1]; however, it is based solely on information provided during controls to combat bad practices [2]. Indeed, customs clearance does not mean the payment of duties and taxes, but rather the completion of all customs formalities for the assignment of a customs procedure to said goods, even in the absence of payment of customs duty. Thus, adapting to each context, risk analysis requires a specific approach every time [3]. Moreover, it is a risky adventure for the revenues, because this method neglected the importance of the moral risk, the administration not having control on the behavior of its agents [4].

Given the large number of customs transactions and the multiplicity of risks, risk analysis is not sufficiently adapted to help it identify customs offences and must evolve to meet these new challenges. Among the works in the literature dealing with these questions known as the system of surveillance of customs offences, a first attempt has been to propose an econometric approach capable of targeting customs declarations that present a real risk of fraud. This model developed by Laporte makes it possible to determine the relevant risk criteria to explain fraud on the basis of historical analysis and to calculate the probability of fraud for any new declaration [5].

$$\Pr\left(fraud_i = 1\right) = \alpha + \beta_1 fq\_crit_i + \beta_2 fq\_crit2_i + \cdots + \beta_N fq\_crit\ N_i + \varepsilon_i$$

With: Pr: probability; $fraud_i$: binary variable 0-1 for operation i (1 if fraud is detected and 0 otherwise); fq_i frequency of fraud for each risk criterion associated with the transaction i, ε_i: random deviation and parameters to be estimated and *Crit = criterion*.

The shortcoming of this model is that it does not take into account the nature of the offence. To solve this problem, he proposes two other models based on a linear probability model: PROBIT ou LOGIT more appropriate for estimating a model whose explained variable is binary in theory but the predicted value cannot be interpreted as a probability of fraud because it does not belong to the [0.1] interval. We can also cite other proposed methods such as the scoring technique to have a more structured approach by effectively assessing the risk and orienting the declarations in the different control circuits of the customs administrations of Developing Countries. Geourjon *et al.* have shown in a research article the relevance of this technique based on an experiment conducted in Senegal. They highlight that the relatively simple scoring technique allows developing countries' customs to assess risk in order to limit controls effectively, and that their development contributes to the modernisation of administrations [6]. Another study conducted by Grigoriou advocates the advantages of the scoring technique to organize controls while ensuring compliance with technical, sanitary and phytosanitary standards [7].

We note that the various methods identified show progress in terms of facilitation in the control process. However, too many issues remain unresolved open as to their uniformity in the different customs administrations. The work of Geourjon *et al.* has shown that each administration has adopted a specific approach to its context and needs [6].

Furthermore, as the analysis and management are mainly based on the use of data in the declarations of the various control circuits, we propose an integrated approach that exploits what already exists in terms of data mining. The idea is to explore historical data and exploit the usual relationships between these data in order to establish rules of association, and subsequently acquire knowledge that led to customs offences. This knowledge will be used for the automatic identification of offences linked to customs activities on the basis of facts (*customs clearance procedure*, *customs investigation*, *control materialisation*, *etc.*).

For example, if we search Google for the word "Fraud", we get 60,000,000 responses directing us to sites containing this word. Suppose we are fast enough to consult a page every three seconds, it will take us a little more than 1000 years to visit them all. This task is not feasible. We therefore need a means not only to store and search for information, but also to analyze and interpret it to help decision-making. Here we can see the importance of setting up an Intelligent Decision Support System to identify fraud and the discovery a priori of situations of infractions. It's in this very specific context that this research work is situated.

2. Learning Problem: Rules of Association

Long before the current development in the field of information and communication technologies, the problem of learning from our data has always been an issue. The development of both information storage and processing technologies has made the task of extracting knowledge more difficult [8], Indeed, we are witnessing not only an exponential growth in the volume of information stored

within our organizations, but also an increasing complexity of this data [9]. Data mining is defined as the non-trivial process of extracting implicit, new and potentially useful information from large volumes of data [10]. It proposes to use a set of techniques and algorithms that aim to discover grounds and knowledge from large amounts of data [11].

Data mining is the key step in the knowledge discovery process. Although this stage is only one part of the general process for knowledge discovery, it has generated the most work in the literature. The techniques and methods used to guide the process and achieve efficient knowledge extraction within data warehouses have been grouped under the name Knowledge Extraction from Data. Association rule extraction is an integral part of a data knowledge extraction process. It's an unsupervised data mining problem that allows from the data of a set frequently appearing in a database to extract knowledge rules.

3. Related Works

Trade facilitation accentuated by globalization has led to rapid growth in the size of databases available in customs administrations. Even if in a recent past, audit work on risk analysis has enabled the modernization of the customs administration's information system in developing countries [6], It must be acknowledged that this method based on descriptive statistic only made it possible to discover statistical irregularities in fraud situations over a given period, the results obtained only defined the probability that any new declaration would present an irregularity (see **Figure 1**).

A new methodological approach is then necessary! It's data mining. Indeed, it is a question of discovering rules of expertise to help in the detection of the notion of risk in a customs system which has become essential because of the volume of data due to the numerous customs operation. Thus, the analysis of its databases has become essential to help the decision-making process against customs offences (see **Figure 2**).

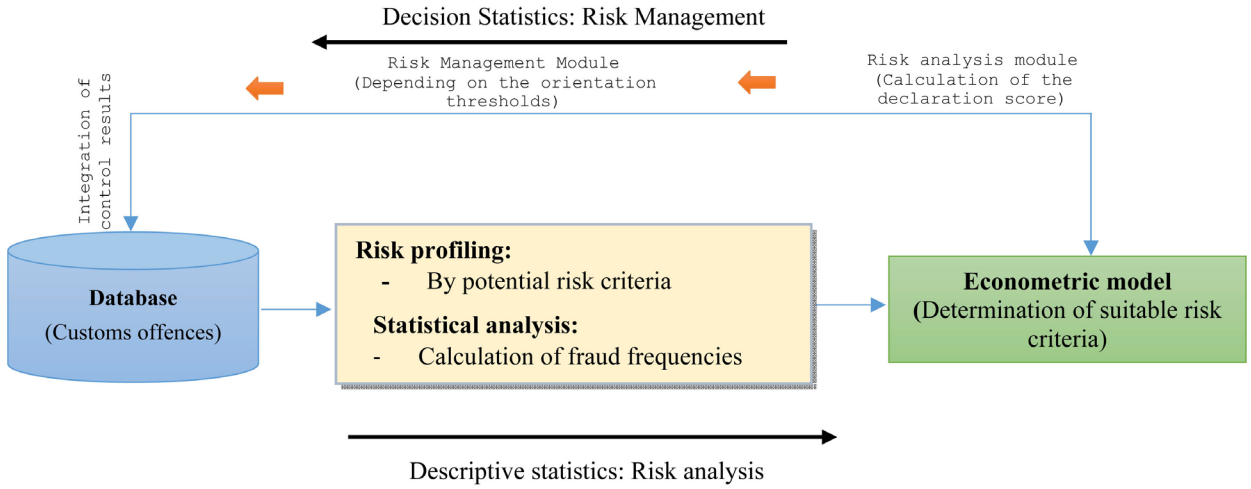

Figure 1. Risk analysis in a customs system using an econometric model.

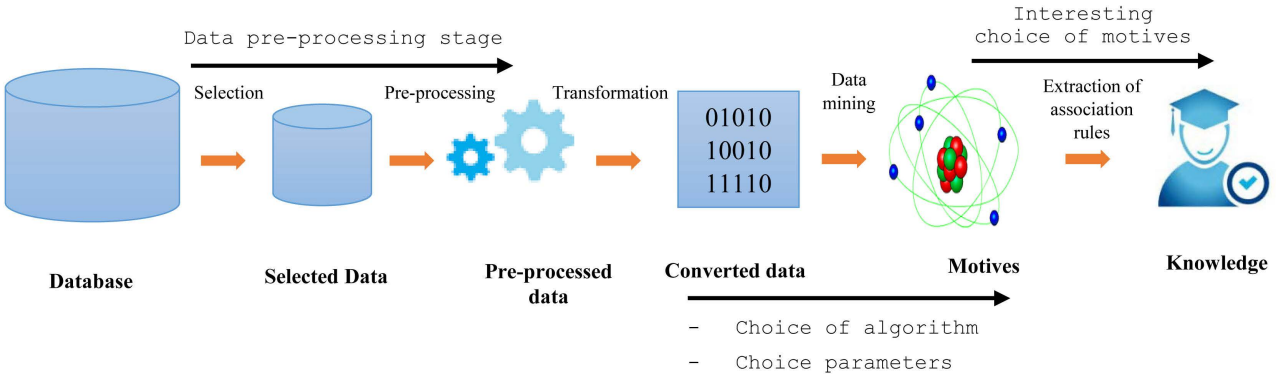

Figure 2. The steps for extract association rules.

Two aspects are noted to motivate this action:

- Extract a general rule from observed data (frequent patterns).
- To discover new knowledge after analyzing this data.

The emergence of new mobile technologies (Cloud Computing) has led to the collection of large amounts of data. The discovery of patterns in data is one of the issues in data mining. Thus, searching frequent motives was proposed to facilitate the extraction of association rules [12]. This approach thus gives a better abstraction of the trajectories and reduces the size of the data for analysis. Cao *et al.* have studied periodic pattern extraction from climate databases, the objects studied, for example storms, have the particularity of following approximately the same route at regular intervals of time. That is to say very frequently, there are seasonal rains at the beginning of the summer [13] [14].

In another idea, in order to extract knowledge rules in a database related to bus trajectories, Fisher *et al.* have highlighted motives which a priori are groups of objects sharing the same type of movement (direction, speed). Each sequence corresponding to the movements of a bus in a region [15]. In the same vein, they develop approximate calculation algorithms to extract identified space-time motives to predict climatic conditions in a given region. An example of patterns extracted by this type of approach is a large number of clouds announcing that rain moved northeast of Montpellier this morning. Recently, Hai *et al.* have proposed a "Framework" using a unifying approach to extract and manage multiple types of patterns representing trajectories (convoys, swarms, etc.) [16] [17]. The extraction of knowledge rules from frequent motives has been widely studied in the literature. Works presented in this document is not exhaustive. It is in this context of study that the work of this article is situated where we apply this knowledge base to a database relating to customs offences.

4. Methodology Approach

Our work concerns the extraction of frequent patterns (attributes) from a database. The generic approach proposed is based on an unsupervised iterative process that will extract frequent motives from a database of customs offences one after the other thus allowing step-by-step exploration of the data. The idea is

to discover associative rules adapted to the customs context to identify and solve problems related to fraud and customs offences. This approach will work on the basis of searching for intrinsic structures, relationships, or affinities in the input data set. In other words, it is about finding trends and correlations that summarize the relationships between data [9] [18]. The objective is to discover association rules to help detect risk situations (fraud, offences). The iterative process is repeated at the user's request. The extraction of a new data will take into account the previously extracted data.

We break the process down into four steps:

1) Stage 1: Identify the different types of reference offences.

2) Stage 2: Create the data structure for a sequential representation of frauds.

3) Stage 3: Find all "patterns" or frequent itemsets, which appear in the database with a frequency greater than or equal to a user-defined threshold, called *Minsup*.

4) Stage 4: Generate the set of associative rules, from these frequent patterns, having a confidence measure greater than or equal to a threshold defined by the user, called Minconf and choose motives representative to establish rules of knowledge.

A rule in this article is defined as the component unit of knowledge. It's of the form $X \to Y$, such that: X is called antecedent of the rule and Y is called ***consequence.*** Thus $X \bigcap Y = \varnothing$.

Creation Corpus

To perform data mining on the basis of frequent motives, we worked on a formal database containing information exclusively on customs operations from 2016 to May 2018 in Côte d'Ivoire (Risk, Intelligence and Value Analysis Directorate; Customs Directorate General).

This information concerns 6854 offences resulting either from customs clearance operations, internal customs investigations, goods controls or exchange controls.

The data selected in this database describe the frauds (nature of the risks, type of offences), and the context of the control carried out (method of operation, customs clearance; value, etc.). This selection of data will constitute the exploration context on which the extraction of association rules will focus in order to highlight the relationships between the different situation factors. The selection of attributes will optimize the number of variables to consider, the number of rules generated and thus facilitate the interpretation of results.

5. Mathematical Tenet: Basic Notion

The association rules have been used successfully in many areas: household basket management, commercial planning assistance, diagnostic assistance and medical research, image analysis and spatial data, organization and access to websites... As part of our work, it is adapted in a customs context to prevent risks of fraud.

The extraction of association rules will consist in extracting rules based on two main parameters: the support and confidence whose minimum thresholds are defined by the user. It is an iterative and interactive process, generally consisting of four steps for most approaches using the frequent motives search technique. These steps are:

1) Data preparation;
2) The search for frequent motives;
3) The generation of association rules;
4) Results interpretation: Discovery of knowledge.

5.1. Data Preparation

Search Context: This phase consists of selecting data useful (attributes and objects) from the database for extracting association rules and transforming these data into an extraction context.

The search for frequent patterns makes the hypothesis of a database describing a set of objects $O=\{o_1,o_2,\cdots,o_N\}$ (*Transactions*), by a finite set of attributes $A=\{a_1,a_2,\cdots,a_n\}$, called also Item. To identify and select an item, we consider a relationship $\mathcal{R}$ of the type 0-1 (Boolean) between an object O and an item a rated $O\mathcal{R}a\in\{0,1\}$. We'll call the Database the triplet $\mathcal{B}=(O,A,\mathcal{R})$.

Definition 1. *Item and Itemset*

1) An item is an occurrence of an object in the database
2) An Itemset is a set of items

In the context of this article, Transactions are represented by customs operations. Items are offences relating to fraud.

Thus if an infringement has been detected on a customs operation, the relation $\mathcal{R}$ takes the value of 1 otherwise 0. Therefore, the Database is modeled by a Boolean matrix where the rows and columns correspond respectively to the objects and attributes specifically offences (see Table 1).

Table 1. Example of a binary database.

$O\mathcal{R}A$	a_1	a_2	a_3	a_4	a_5
1	1	1	0	0	1
2	0	1	0	1	0
3	0	1	1	0	0
4	1	1	0	1	0
5	1	1	1	0	0
6	0	1	1	0	1
7	1	0	1	0	0
8	1	1	0	1	1
9	1	1	1	0	0
10	0	1	1	0	0

Table 1 is a context representing 10 customs operations (The rows) and 5 types of offences (columns) rated {a_1, a_2, a_3, a_4 and a_5}. Intercession in the table is the $\mathcal{R}$ relationship between a customs operation and an offence.

The Interpretation of Table 1 is: a_1, a_2 and a_4 offences are associated with customs operations N˚4.

5.2. Search for Frequent Motives

The method of searching for frequent motives is based on the formal notion of motive. This phase consists of extracting of context all sets of binary attributes $m \subseteq A$, called itemsets, that are frequent in context $\mathcal{B}$.

The set of frequent itemsets will be noted *M*. Frequent itemset search problem is exponentially complexity in size *n* of all items since the potential number of frequent 2^n.

An itemset is a subset of*A*itemset describes an object ***o*** when $\forall aM, o\mathcal{R}a$ and we note $o\mathcal{R}M$. description of an object $o \in O$ is the attribute $d(o) = \{a \in \mathcal{A} / o\mathcal{R}a\}$. Anitemset of size *k* in noted *k-itemset.*

To find the 2^n sets of itemsets that appear frequently, we introduce the notions of Galois connection and support of an itemset.

Definition 2. *Galois connection* [19]

Galois connections are a fundamental object in ordered set theory. In this article, the Galois correspondence associated with the Database is the pair of functions (f, g) defined by:

$$\begin{cases} f: 2^n \to 2^0 \\ m \to f(m) = \{o \in O / o \text{ contain } m\} \\ g: 2^n \to 2^0 \\ o \to g(o) = \{a \in \mathcal{A} / \forall o \in O, (o, a) \in \mathcal{R}\} \end{cases}$$

g is called dual of *f* and *f* is called dual of *g*. It's sometimes said that $f(m)$ eis the image of motives*m*

Definition 3. *Itemsets and Support*

An important notion for a set of item is its support which refers to the proportion of the objects in the database that contain it (Number of transactions observed). The support of an itemset is defined by:

$$\text{Support}: 2^n \to [0,1]$$

$$m \to \text{Support}(m) = |f(m)| / |O|$$

This definition is relative to the size of the database, the support of a set is always less than or equal to the support of its subsets or, considering a set of *X* item, items support $\varphi_s(X)$ associated with the itemset is:

$$\varphi_s(X) = \text{Card}(\{o \mid X \subseteq o, o \in O\})$$

Property 1. *Support for subsets*

Let be two sets X and Y. If $X \subseteq Y$ *for* itemsets *X, Y then Support* (X) ≥ *Support* (*Y*) because *all transactions in the Database that support Y also neces-*

sarily support X.

Definition 4. *Frequent Motives*

Let $\varphi_s \in [0,1]$, called the minimum support (*Minsup*).

A pattern m is said frequent if $\text{Support}(m) \geq \varphi_s = Minsup$.

Definition 5. *Confidence of a rule*

The confidence of a ruler is a measure of precision. Confidence in a rule $r: X \to Y$ is defined as follows:

$$\begin{aligned} Conf(r) &= \text{p}(Y \subseteq O \mid X \subseteq O) \\ &= \text{p}(Y \subseteq O \wedge X \subseteq O) / \text{p}(X \subseteq O) \\ &= \text{Support}(X \cup Y) / \text{Support}(X) \end{aligned}$$

Note: To reveal the relevance of a rule we use two concepts which are support and confidence. In order to be retained, each rule must have superior support to *Minsup* and superior confidence to *Minconf*. These two values are defined empirically by the system user.

5.3. Rules of Association

For this section, we refer to [20].

Definition 6. *Rules of association*

An association rule is a rule of implication between two sets to which are associated the supporting measure, which defines the scope of the rule, and the confidence measure, which defines the precision of the rule in the context of extraction. Support and confidence indicate the usefulness and relevance of the rule.

An association rule r is an implication of the form $X \to Y$ between two sets of items.

An association rule r is an implication of the form $X \to Y$ between two sets of items X and Y, $X \bigcap Y = \varnothing$, such as:

$$\text{Support}(r) = \text{Support}(X \cup Y) \mid N$$

$$\text{Confidence}(r) = \text{Support}(X \cup Y) \mid Support(X)$$

The notions of support and trust were identified in the first research studies of association rules conducted by Hajek, Havel and Chytil (1966) in the GUHA method [21].

Confidence is equal to a support ratio:

- A rule r is considered **valid** if Confidence(r) > Minconf
- A rule r is **total** if confidence(r) = 1 et **partial** otherwise

5.3.1. Extraction Method of Frequent Motives: Principle of Apriority

The reference algorithm based on this approach is the Apriori algorithm [22]. Like all association discovery algorithms, it works on transactional databases. The principle is based on a path by level of all the motifs. A set of rules (of candidates) is generated from this list. The candidates are tested on the database, in other words the instances of the generated rules and their occurrences are

searched, and the candidates not respecting Minsup and Minconf are removed. The algorithm repeats this process by increasing each time the size of the candidates of a unit as long as relevant rules are discovered. At the end, the discovered sets of rules are merged. The generation of candidates is done in two stages: **Joint** and **pruning**. The join consists of a crossing of a set of rules to $(k-1)$ elements on itself which results in the generation of a set of candidates to k elements. As for pruning, it deletes candidates whose at least one of the sub-chains with $(k-1)$ elements is not present in the set of rules with $(k-1)$ elements. Itemset lattice allows to use this extraction algorithm more efficiently by admitting the following properties:

Property 2: Any subset of a frequent Itemset is frequent.

Property 3: All itemset subset infrequent is infrequent.

The notations are presented in Table 2 and the pseudo code in algorithm 1.

The generic scheme of the algorithm is summarized as in Figure 3.

Pseudo code is presented in Figure 4.

Input: *Database* (*Extraction Context*), *Minsup*, *MinConf*

Output: frequent Itemset set: $\cup_k M_k$

1. Initialize the set of size 1 candidates 1, k = 1
2. **While** Non-empty set of candidates **Do**
3. **Pruning Stage**
   1) Calculate candidate support
   2) Pruning of all candidates in comparison to Minsup
4. **Construction stage**
   1) Build the set of candidates to use in the next iteration
   2) Go to Stage 3
5. **End_While**
6. **Return**: frequents itemset set
7. **Extraction of association rules** m => (l – m)

Figure 3. General scheme of the algorithm a priori.

Algorithme 1: Pseudo code for search of frequent

InPut: Database: Corpus, *Minsup*: Entier

OutPut: Ifrequent temSet Set

BEGIN
$A_1 \leftarrow$ {Singletons}
$k \leftarrow 1$
While $A_k <> \emptyset$ **Do**
For chaque $m \in A_k$ **Do**
For chaque $o \in O$ **Do**
If $m \in O$ **Then**
$Supp\ (m) \leftarrow Supp(m) + 1$
$M_k \leftarrow \{m \in A_k / Supp\ (m) \geq MinSupp\}$
$k \leftarrow k + 1$
$A_k \leftarrow$ Algo_Apriori-Gen (M_{k-1})
End_If
End For
End _For
End _While
Return $\cup_k M_k$
END

Figure 4. Algorithm a priori.

Table 2. Notation used in the algorithm.

Notation for Algorithm 1	
k	Current iteration number
A_k	Subset of attributes
M_K	Frequent motives of size k
(m)	Motives
Supp (m)	Support of m

Algo_Apriori-Gen (M_{k-1}) is the function that generates the candidate itemset by performing two major operations:

- The generation of candidates
- Pruning candidates

The basic idea of this function is to extend each set of frequent patterns of depth $k-1$ by adding to them other frequent patterns. This quick procedure makes it possible to find all the sets of frequent patterns of size k, however, in order to avoid being compared with several identical sets, we add a pruning step (*classification of the motives in alphabetical order, then we compare the itemset different obtained*) (**Figure 5**).

5.3.2. Basis for the Rules of Association

The search problem an association rule can be formulated as follows:

Given a transaction set T, found all the association rules having a *support* ≥ *Minsup* and a *confidence* ≥ *Minconf* where Minsup and Minconf are respectively thresholds for support and confidence.

A rule of association is of the form: **Antecedent** → **Consequence (**Support, Confidence) with

Support and **confidence** are interest measures defined by user.

It is an implication between two itemsets to which are associated the support, which defines the scope of the rule, and the confidence, which defines the precision of the rule in the context of extraction. To elicit associative rules, we search for generalizations of database motives that frequently appear in order to find regularities in the database in the form of frequently associated elements.

A rule can have excellent support and confidence without being "interesting"; In this case, we need a criterion in order to limit the proliferation of rules (Because if there are m items, there's $\sum_{k=2}^{m}\binom{m}{k}\left(2^k-2\right)$ possible associative rules) it's in this perspective that we introduce a new parameter that is an indicator of the relevance of associative rules: The Lift which is a measure of the performance of the association rule by checking whether the results obtained are not a result of chance [23]. His interpretation is as follows:

- If the measurement is greater than 1, it indicates a positive correlation: the ruler is considered interesting. If the measurement is 1, its correlation is zero, the measurement in this case is useless and when its measurement is less than 1, the correlation is negative. Calculation of the lift is defined as follows: $\text{Lift} = \text{Conf}\left(X \to Y\right) \mid N$.

Algorithm 2. Pseudo-code for frequent itemset search

Entrée: *M*cardinal's frequent itemset k

Begin

$A \leftarrow \{a = m_1 \cup m_2 \text{Such as}(m_1, m_2) \in A \times A, \text{card}(a) = k + 1\}$

Fora∈ *A***Do**

Foreach$m \subset a$ **Do**

$\text{Card}(m) = k$

Forchaque $o \in O$**Do**

If$m \notin M$**Then**

$A \leftarrow A \backslash \{m\}$

End_If

End _ For

End _ For

End _ For

Return*A*

END

Figure 5. Generating frequent itemsets set.

Finally, to facilitate the exploitation of these discovery rules, we categorize them into three groups:

1) Forecast rule: These are useful rules containing quality information. The antecedent is known a priori contrary to its consequent. In this case the confidence of the rule is greater than 50%.

2) Targeting rule: These are general knowledge rules that identify the relationships between the different attributes (motives). The antecedent and consequence of the rule are known but not the implication relationship between the two parties.

3) Neutral rule: These rules do not provide new information

A rule denotes of the interaction between two events (customs clearance transaction and a customs clearance fraud risk) where their actions are generally dependent, which can lead to a risk of fraud.

5.3.3. Algorithm Illustration

We present a detailed example of the steps followed by the algorithm Apriori from the context presented in **Figure 6**.

(**Figure 6** is determined from the context of the matrix (**Table 1**) set out in Section 5.1 of this article.)

Table 3 is an association rule extraction context consisting of ten transactions, each identified by a number, and five items. For this example, the minimum support is set at 0.3; that is, a minimum count required for three operations performed. (Frequency is expressed as a percentage)

- **Interpretation of Figure 6**

Stage 1. 1^{st} *Scanning of the DB and calculation of the* 1*-itemset supports*

To $k = 1$, algorithm performs the first scan counting the support of each 1-itemset of the Database, thus, we form the set of candidates A_1 which makes it possible to generate M_1, the set of frequent 1-itemsets

Stage 2. 1^{st} *pruning in Database*

During this step, the algorithm performs the first pruning by comparing the

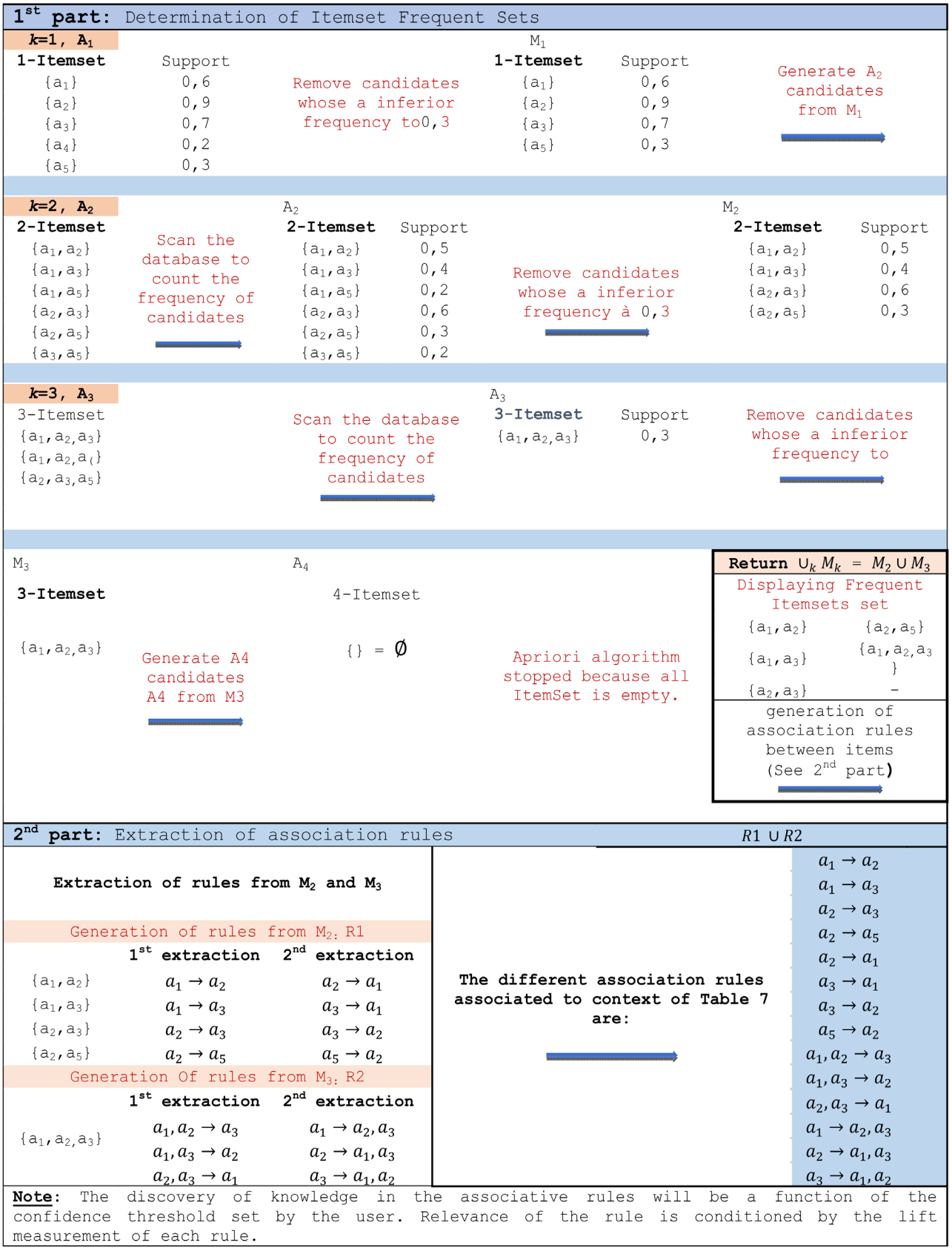

Figure 6. Illustration of the Apriori algorithm from the context described in **Table 3**.

Table 3. Example of a database of 10 operations.

N°	Items
1	a_1, a_2, a_5
2	a_2, a_4
3	a_2, a_3
4	a_1, a_2, a_4
5	a_1, a_2, a_3
6	a_2, a_3, a_5
7	a_1, a_3
8	a_1, a_2, a_3, a_5
9	a_1, a_2, a_3
10	a_2, a_3

frequency of each 1-itemset with the minimal support. All 1-itemset having their support ≥ Minsup defined by the system are kept to form

$$M_1 = \{\{a_1\},\{a_1\},\{a_1\},\{a_1\}\}$$

Stage 3. *The Junture*

The 1-itemsets of M_1 are used to generate candidate sets of A_2. The 1-itemsets of M_1 are used to generate candidate sets of A_2. This possible combination of de $n(n-1)|2$ where n is the number of Itemset is achieved by linking the k-Itemset of M_k between them. Applying this principle, the number of combinations to be formed to obtain A_2 is to six (6). The candidates obtained are:

$$A_2 = \{\{a_1,a_2\},\{a_1,a_3\},\{a_1,a_5\},\{a_2,a_3\},\{a_2,a_5\},\{a_3,a_5\}\}$$

Stage 4. 2nd *scanning of Database and calculation of supports to* 2-*itemsets*

The 2-itemsets of A_2 being generated, the algorithm performs another scan to determine the frequency of all A_2 candidates.

Stage 5. 2nd *pruning in Database*

The algorithm performs its second pruning by traversing A_2 in order to eliminate all Itemset whose support is lower than *Minsup*. The other 2-itemsets are kept to form $M_2 = \{\{a_1,a_5\},\{a_3,a_5\}\}$.

Stage 6. *Generation of candidates*

This is the generation of the candidates of the 3-itemsets, carried out by applying the principle of the join of step 3 as well as the properties 2 and 3 of section 5.2; at the end only the items and {a_1, a_2, a_3} is generated.

Stage 7. 3rd *Scan and frequency determination of* 3-*itemsets*

The third scan of the database is used to calculate the frequency of the items and {a_1, a_2, a_3} whose measurement is 0.3.

Stage 8. 3rd *pruning in database*

The algorithm compares the frequency of the items and {1, 2, 3} with the minimum frequency. Since {a_1, a_2, a_3} has the minimum frequency, it is kept and becomes the only item and M_3, the set of frequent 3-itemsets.

Stage 9. *Generation of candidates*

Since the M_3 primer set contains only one Itemset, {a_1, a_2, a_3}, no candidate 4-itemset can be generated. Therefore $A_4 = \varnothing$. The algorithm stops here.

Stage 10. *Set Itemset frequent*

The algorithm returns the sets of the different frequent *k*-Itemset (M_k):

$$\bigcup_k M_k = M_2 \cup M_3 = \{\{a_1, a_2\}, \{a_1, a_3\}, \{a_2, a_3\}, \{a_2, a_5\}, \{a_1, a_2, a_3\}\}$$

Thereafter, we can now establish the different associative rules.

Stage 11. Extraction of association rules

In this part, the algorithm will to extract all the association rules at each iteration *k*.

5.4. Experimental Validation: Material and Method

The objective of this section is to show the feasibility of the Apriori principle on the Database of Risk, Intelligence and Value Analysis Directorate in order to extract knowledge to prevent risks of fraud in customs operations. This database is composed of 6854 infringements over the period 2016 to May 2018 resulting from various customs operations.

The experiments were conducted on a computer platform Intel Core™ i7-3540M 3.00 GHz with 8 GB RAM on Linux operating system The Apriori algorithm has been implemented in the "Arules" package of the R software package. The Programing language is Python via the PyFIM library.

- **Computer coding**

- *To obtain rules with at least* 20% *support and more than* 60% *confidence, simply run the command:*

  rules < -apriori(Adult, parameter = list(support = 0.2, confidence = 0.6))

- *If you choose to focus on forecasting rules having the item "False_declaration of value" as a right member and sort by confidence:*

  Rules < -apriori (Adult, parameter=list(support = 0.2))

  rules.False_declaration of value<-subset(rules, subset = rhs %in% "False_declaration of value")

  rules. False_declaration of value <-sort(rules.False_declaration of value, by = "confidence")

  inspect(rules.False_declaration of value)

- *To specify properties of the searched rules, the* ***subset ()*** *function is used. Tests can also be combined in the subset() call with the interest measure Lift*

  subset = rhs %in% "False_declaration of value" & lift > 1.5.

5.5. Results and Interpretation

He analysis of the results revealed several interesting rules. Some of these are shown in **Table 4**.

We analyze and interpret some lines of the table:

- ***Forecast rule***: Operation 2 (*Supp.* = 0.35; *Conf.* = 0.57; *Lift* = 1.07)

Table 4. Implementation results.

N°	Customs operation category	Infringement-type	Supp.	Conf.	Lift	Rule-type
1	Exchange control	capital outflow	0.10	0.61	1.07	(b)
2	Clearance of goods	Misrepresentation of value	0.35	0.57	2.36	(a)
3	Goods control	Misrepresentation of origin	0.35	0.59	3.02	(a)
4	Clearance of goods	Embezzlement	0.14	0.41	1.5	(c)
5	Clearance of goods	Misreporting of currency	0.35	0.53	1.8	(a)

Clearence of goods → Misrepresentation of value. This rule is consistent because it informs us that 57% of the risks of fraud in goods customs clearance come from false declarations of value.

- **Targeting rule:** Operation 1 (*Supp.* = 0.1; *Conf.* = 0.61; *Lift* = 1.74)

Exchange control → Capital outflow, this rule gives us specific information, justifiable by the fact that 61% of the risks of capital flight are essentially linked to foreign exchange control operations.

- ***Neutrale rule***: Operation 4 (*Supp.* = 0.14; *Conf.* = 0.41; *Lift* = 1.5)

Clearence of goods → Embezzlement, this rule is of no interest because the information is not relevant because it has only one premise. The information it provides does not specify its nature of risk (diversions are indeed risks of fraud in a customs clearance operation).

6. Conclusion

Extraction of Knowledge from Data is nowadays one of the more and more used means to learn from our data. In this paper, we have presented an original approach to discovering knowledge applied to data relating to customs offences. The result obtained is a set of knowledge rules of forecasting and targeting certain risk situations. A selection criterion based on the frequency of reasons showed the effectiveness of this model in discovering associations rules aimed at preventing risks. However, control in the customs system depends both on administrative procedures and on the action of men in the control process; we propose, in future work, to develop an unsupervised clustering method adapted to the customs context allowing interpreting the results on different levels of granularity to facilitate the understanding of the model.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Harisson, M. (2007) Challenges for Customs, Customs and Supply Chain Security, The Demise of Risk Management? *Annual Conference on APEC Centers*, Melbourne, 18-20 April 2007.

[2] Truel, C. (2010) Guide rapide sur les risques douaniers. Séries de brefs guides sur les

risques. Gower Publishing Limited, Burlington & Union Road.

[3] Gates, S. (2006) Incorporating Strategic Risk into Enterprise Risk Management: A Survey of Current Corporate Practices. *Journal of Applied Corporate Finance*, **18**, 81-90. https://doi.org/10.1111/j.1745-6622.2006.00114.x

[4] Geourjon, A.M. and Laporte, B. (2004) L'analyse de risque pour cibler les contrôles douaniers dans les pays en développement: Une aventure risquée pour les recettes? *Politiques et Management Public*, **22**, 95-109. https://doi.org/10.3406/pomap.2004.2857

[5] Laporte, B. (2011) Risk Management Systems: Using Data mining in Developing Countries' Customs Administrations. *World Customs Journal*, **5**, 17-27.

[6] Geourjon, A.M., Laporte, B. Coundoul, O. and Gadiaga, M. (2012) Contrôler moins pour contrôler mieux: L'utilisation du data mining pour la gestion du risque en douane, CERDI, Etudes et Documents, E 2012.06.

[7] Grigoriou, C. (2012) How Can Risk Management Help Enforce Technical Measures? In: Cadot, O. and Malouche, M., Eds., *Non Tariff Measures*: *A Fresh Look at Trade Policy's New Frontier*, World Bank/CEPR, Washington DC, London.

[8] Kantardzic, M. (2003) Data Mining: Concepts, Models, Methods and Algorithms. Wiley-IEEE Press, Totowa, NJ.

[9] Tan, P.N., Steinbach, M. and Kumar, V. (2006) Introduction to Data Mining. Addison Wesley, Boston, MA.

[10] Fayyad, U.M. (1996) Data Mining and Knowledge Discovery: Making Sense Out of Data. *IEEE Intelligent Systems*, **11**, 20-25. https://doi.org/10.1109/64.539013

[11] Berry, M.J. and Linoff, G.S. (2011).Data Mining Techniques—For Marketing, Sales and Customer Support. 3rd Edition, Wiley Computer Publishing, New York.

[12] Agrawal, R., Tomasz, I. and Arun, S. (1993) Database Mining: A Performance Perspective. *IEEE Transactions on Knowledge and Data Engineering*, **5**, 914-925. https://doi.org/10.1109/69.250074

[13] Cao, N., Mamoulis, H. and Cheung, D.W. (2005) Mining Frequent Spatio-Temporal Sequential Patterns. *IEEE International Conference on Data Mining ICDM*, Houston, TX, 27-30 November 2005, 82-89.

[14] Cao, N., Mamoulis, H. and Cheung, D.W. (2007) Discovery of Periodic Patterns in Spatiotemporal Sequences. *IEEE Transactions on Knowledge and Data Engineering TKDE*, **19**, 453-467. https://doi.org/10.1109/TKDE.2007.1002

[15] Fisher, P., Laube, M.K. and Imfeld, S. (2005) Finding REMO—Detecting Relative Motion Patterns in Geospatial Lifelines. In: *Developments in Spatial Data Handling*, Springer Berlin Heidelberg, 201-215. https://doi.org/10.1007/b138045

[16] Hai, P.N., Poncelet, P. and Teisseire, M. (2012) GET MOVE: An Efficient and Unifying Spatio-Temporal Pattern Mining Algorithm for Moving Objects. 11th *International Conference on Advances in Intelligent Data Analysis*, Heidelberg, Berlin, 276-288.

[17] Hai, P.N., Ienco, D., Poncelet, P. and Teisseire, M. (2013) Mining Representative Movement Patterns through Compression. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, April 2013, 314-326. https://doi.org/10.1007/978-3-642-37453-1_26

[18] Hornick, M.F., Erik, M. and Venkayala, S. (2007) Java Data Mining: Strategy, Standard, and Practice: A Practical Guide for Architecture, Design, and Implementation. Morgan Kaufmann, Burlington.

[19] Birkhoff, G. (1967) Lattices Theory. 3rd Edition, American Mathematical Society,

New York.

[20] Borgelt, C. (2012) Frequents Item Set Mining. *Data Mining and Knowledge Discovery*, **2**, 437-456. https://doi.org/10.1002/widm.1074

[21] Hajek P., Havel, I. and Chytil, M. (1966) The GUHA Method of Automatic Hypotheses Determination. *Computing*, **1**, 293-308. https://doi.org/10.1007/BF02345483

[22] Agrawal, R. and Srikant, R. (1994) Fast Algorithms for Mining Association Rules. *Proceedings of the* 20*th VLDB Conference Santiago*, Chile, San Francisco, September 1994, 487-499.

[23] Le bras, Y., Lallich, S. and Lenca, P. (2011) Un cadre formel pour l'étude des mesures d'intérêt des règles d'association. *Journée d'animation du GRD I3 sur la fouille de données*, Lyon, September 2011.

ANNEXE D: Publication Scientifique #2

"*Predictive analysis of risk behaviors related to customs activity using a diagram algorithm*"

ABSTRACT:

Context and generality. The digitization of customs administrations in recent years and the volume of commercial exchanges have contributed to setting up databases of customs transactions from checkpoints. These databases are supplied by various customs clearance operations and transactions carried out. Objective and methodology. The aim is to help the custom administration to anticipate the risks of fraud, in this perspective, the article proposes an extension of the Apriori algorithm which allows the discovery of knowledge by proposing rules of association with level of concepts. Indeed, rather than extracting rules of association at the level of customs operations between the type and nature of customs offenses, we explore the symbolic structure of the data with the underlying idea of extracting new association rules at the level of operators in order to study their behavior in the fraud process. Result. Considering the "confidence diagram (CD)" indicator, whose minimum threshold value is set at 55%, we discover rules at the level of operators considered as concepts and whose extensions are their customs clearance operations.

Key word: *Diagram confidence association rules, symbolic data, extended Apriori algorithm, customs clearance operation.*

Scientific Journal: *Far East Journal of Mathematical Sciences, 2019, January.*

Zehero, B.B., *Brou Pacôme., Soro, Etienne, Asseu Olivier, Daniel Bourget. (2019). Far East Journal of Mathematical Sciences, vol 110, issue 1, pp: 217-232.*http://dx.doi.org/10.17654/MS110010217

Far East Journal of Mathematical Sciences (FJMS)

http://www.pphmj.com
http://dx.doi.org/10.17654/MS110010217
Volume 110, Number 1, 2019, Pages 217-232 **ISSN: 0972-0871**

PREDICTIVE ANALYSIS OF RISK BEHAVIORS RELATED TO CUSTOMS ACTIVITY USING A DIAGRAM ALGORITHM

Zehero Bi Bolou[1], Brou Pacôme[2], Soro Etienne[2], Asseu Olivier[1,2] and Daniel Bourget[3]

[1]Institut National Polytechnique - Houphouët Boigny (INPHB)
Yamoussoukro, Côte d'Ivoire

[2]Ecole Supérieure Africaine des Technologies de l'Information et de la Communication (ESATIC)
Abidjan, Côte d'Ivoire

[3]Institut Mine Télécom Atlantique (IMT Atlantique)
Brest, France

Abstract

Context and generality. The digitization of customs administrations in recent years and the volume of commercial exchanges have contributed to setting up databases of customs transactions from checkpoints. These databases are supplied by various customs clearance operations and transactions carried out.

Objective and methodology. The aim is to help the custom administration to anticipate the risks of fraud, in this perspective, the article proposes an extension of the apriori algorithm which allows the discovery of knowledge by proposing rules of association with level of concepts. Indeed, rather than extracting rules of association at the level of customs operations between the type and nature of customs

Received: July 15, 2018; Accepted: September 25, 2018
Keywords and phrases: diagram confidence association rules, symbolic data, extended apriori algorithm, customs clearance operation.

offenses, we explore the symbolic structure of the data with the underlying idea of extracting new association rules at the level of operators in order to study their behavior in the fraud process.

Result. Considering the “confidence diagram (CD)” indicator, whose minimum threshold value is set at 55%, we discover rules at the level of operators considered as concepts and whose extensions are their customs clearance operations.

1. Introduction and Motivation

Association rules have been created to extract knowledge from data, and are usually of the form “If <antecedent>, then <consequent>”. In the literature, its definition varies according to the three main currents: (1) The rules of association with decisional statistics (Gras [1]), (2) the rules of association with an ordered representation of informative concepts (Guigues and Duquenne [2]) and (3) association rules with analysis of transactional data in large databases (Agrawal et al. [3]). In this paper, we will present an approach to association rules based on the 3rd definition, that is, the one associated with the analysis of transactional data in large databases. A classic example of rules of association is the approach described homemaker’s basket by Agrawal and Srikant in [4], which gives an overview of a set of purchases made at the supermarket. Considering two articles *A* and *B*, the rule of type $A \rightarrow B$ means that “if article *A* is present in homemaker’s basket then there is also article *B*”. This analysis was intended to establish relationships interesting in order to build a strategy for decision-makers:

- « *Should we put the two products on the same row or should we separate them to force the consumer to spend more time in the store*?

- *Is it necessary to put near them the most profitable products or the products whose expiry date is close*? »

Subsequently, several studies have been conducted to optimize the basis of association rules by exploiting the complex structure of data in databases. Thus, we have: Guigues and Duquenne [2] who defined a minimal basis for

exact association rules with a confidence probability of 1. On the other hand, the frequency of appearance of items in homemaker's basket has allowed Cai et al. [5] and Wang et al. [6] to discover weighted rules. Srikant et al. [7] and Han and Fu [8] exploited the taxonomy relations in the data. Creation of numerous quality indices has allowed its generalization to other types of data, particularly in the exploitation of fuzzy sets (Kuok et al. [9]), and the use of quantitative data (Srikant and Agrawal [10] and Miller and Yang [11]). In the previously cited works, association rules extraction is based on two main indicators: support and trust for values above minimum thresholds. Therefore, many indicators, besides support and confidence, have been proposed to evaluate the quality of the rules obtained: *centered confidence, conviction, entropy gain, information rate, lift, Piatetsky-Shapiro, Laplace* ... for the most part, these indicators are correlated with support and trust (Bayardo and Agrawal [12] and Blanchard et al. [13]). Conversely, some authors have worked to propose more efficient algorithms on the complexity of data to avoid an explosion of rules extraction time because of the capacity of data storage spaces (Pasquier [14]). In Zehero et al. [15], we have proposed an apriori algorithm that extracts association rules between the nature and the type of offense identified in a customs operations database; this article is an extension of this work. Indeed, in this paper, we propose to extend the apriori algorithm in order to highlight new rules of association at the level of the operators to analyze their behavior with respect to the infringements observed during the operations of clearance. This includes new definitions for support and confidence to take advantage of the symbolic structure of the data.

2. Basic Concept

Search for algorithms capable of extracting association rules from large databases has been a much studied topic by Agrawal et al. [3]. The discovery of rules of association between different products in the basket of the housewife was an example of a particularly exploited application. In the case of our work, unlike homemaker's basket, this is a formal customs transaction

database, in particular customs clearance operations or infringements have been detected. Thus, discovery of the association rules aims to highlight the nature and type of infringement during customs clearance operations (see Zehero et al. [15]). In this article, we will extend these rules to the concept level to elucidate the behavior of the economic operators (importers, exporter) in relation to the various infringements found on the different customs clearance processes listed.

In Table 1, we present a classic example of the transaction matrix.

Table 1. Example of a matrix with 10 cases of classic transactions

Transaction	Wording	Items
t_1	1	a_1, a_2, a_5
t_2	1	a_2, a_4
t_3	1	a_2, a_3
t_4	2	a_1, a_2, a_4
t_5	2	a_1, a_2, a_3
t_6	3	a_2, a_3, a_5
t_7	3	a_1, a_3
t_8	3	a_1, a_2, a_3, a_5
t_9	4	a_1, a_2, a_3
t_{10}	4	a_2, a_3

2.1. Classic data mining: apriori algorithm and classic rules

Items $(a_i, i = 1, ..., 5)$ represent the different offenses. Subsets of items are called *itemsets*. A transaction is a customs clearance operation. Let $I = \{i_1, ..., i_n\}$ be the of *n* items and $T = \{t_1, ..., t_m\}$, $t_i \in P(I) \backslash \varnothing$ a set of *m* transactions. An association rule is a rule such that $A \to B$, $A \subset I$, $B \subset I$ (items subsets), $A \bigcap B = \varnothing$. In Agrawal and Srikant [4], the authors suggest the apriori algorithm. The idea is to generate the association rules with support (Sup) and confidence (Conf) higher than two minimum thresholds Minsup and Minconf, respectively, where:

$$\text{Sup}(A \to B) = \frac{\text{Card}(t \in T / A \cup B \subseteq t)}{\text{Card}(t)}, \quad \text{Conf}(A \to B) = \frac{\text{Sup}(A \to B)}{\text{Sup}(B)}.$$

In the remainder of this section, we present the properties on which the apriori algorithm is based, which allow the removal of non-frequent itemsets, as well as the generic scheme of the apriori algorithm. The lattice of the itemsets makes it possible to use this extraction algorithm more efficiently by admitting the following properties:

Property 1. Any subset of a frequent itemset is common.

Property 2. Any superset of a non-frequent itemset is uncommon.

Property 3. Any itemset included in a frequent itemset is itself frequent.

The generic scheme of the algorithm is thus summarized:

Table 2. General scheme of apriori algorithm

Input : *Database* (*Extraction Context*), *Minsup*, *Minconf*
Output : frequent Itemset set : $\bigcup_k M_k$
1. Initialize the set of size 1 candidates 1, $k = 1$
2. **While** Non-empty set of candidates **Do**
3. **Pruning Stage**
(1) Calculate candidate support
(2) Pruning of all candidates in comparison to Minsup
4. Construction stage
(1) Build the set of candidates to use in the next iteration
(2) Go to Stage 3
5. End_While
6. **Return :** frequents itemset set
7. **Extraction of association rules** $m => (1 - m)$

In Zehero et al. [15], we applied the classical apriori algorithm to the classical transaction matrix. For the execution of the algorithm, we invite the reader to consult for details.

2.2. Symbolic data mining: concept and assertions

2.2.1. Symbolic data: description and definition

Description of the notion of symbolic object. A symbolic object (OS) models concepts. The main idea of the analysis of symbolic data is to move from the study of individuals to the study of concepts described by variables interval, weighted, diagrams, etc., and for which the standard digital operators $(\times, +, -)$ cannot be applied directly. Thus, a concept is usually defined by a set of properties called *intension* and a set of individuals satisfying these properties called *extension* (Bock and Diday [16]).

Definition 1. Let Ω be all individuals, and D the set of individuals descriptions or class of individuals. OS is a triplet $s = (a, R, d)$, where $d \in D$ is a description, R is a relationship between "d" and "a" of $\Omega \to L$. It is a function of recognition between individuals and their descriptions. Two types of OS are listed for two different L sets: OS Boolean type and OD modal type.

- Boolean OS is such that: $[y(w)Rd] \in L = \{\text{True, false}\}$.

For example $a(w) = \text{Nature\_offence}\ (w) \subseteq [a_1, a_2] \wedge [\text{Typology\_Offence}\ (w) \subseteq [b_1, b_2]] = (\text{True} \vee \text{false}) = \text{true}$ where $w \in \Omega$ and b_1, b_2 are type offences.

Nature_offence and Typology_offence are two Typology_Offence w.

- Modal OS is such that: $[y(w)Rd] \in L = [0, 1]$.

In the remainder of this article, we use Boolean OS to analyze risk behaviors related to customs activity.

Definition 2. An assertion is an OS defined by $[d'Rd] = \wedge_{i=1,\, p}[d'_i R_i d_i]$, $p \geq 1$.

The extension of an OS is given by $Ext(s) = \{w \in \Omega / a(w) = True\}$.

NB. We specify that the extension given to Definition 2 concerns only the Boolean case.

2.2.2. Corpus and evaluation measures

To apply the algorithm to the case of diagram symbolic variables, we use a formal database of information on customs offenses from 2016 to May 2018 in Côte d'Ivoire (data from the simplified report database; General Directorate of Customs).

This information concerns the offenses arising from customs clearance operations, internal investigations by the customs administration, goods controls or exchange controls. Data are:

- Number of report simplified (Num_PVS) which represents the litigation number of the clearance operation;

- Wording of the offense;

- Operator code that identifies the operator;

- Office code that identifies the type of operator: importer or exporter;

- Number of the offense that identifies the offense;

- Fob value declared;

- Fob value recognized;

- The right of compromise.

Note. (1) If the Fob value declared by the operator < to the Fob value recognized by the customs services, then there is an offense or fraud found, in which case a compromised right is charged to the operator guilty of the fraud.

(2) According to the conceptual data model (CDM) of the formal database of customs operations, and the various keys Id, from Num_PVS, we can identify the nature of the offense and the type of the goods on which the fraud was identified as well as the type of operator.

2.2.3. Presentation of a symbolic matrix

By referring to a classic transaction matrix, it will be more a question of having a diagram in each box, that is, weighted multiple values such that the sum of the weights is equal to one instead of a single value per cell in our data matrix (see Table 1) or a set of items by transaction as in the classic case.

In classical apriori approach, statistical units are transactions; on the other hand, with this new approach: “apriori diagram”, it is concept that we study, that is to say the behavior of the operators with respect to the customs infractions rather than the customs operations.

To obtain the symbolic matrix of Table 1, we delete the first column of Table 1 and we describe Table 3: for each operator, this matrix aggregates the different items found in the form of a diagram constructed with the proportion of each offence in relation to the operator’s total items. Table 3 is a symbolic matrix where each line defines the “description” of an operator and each column is associated with a symbol variable.

Table 3. Matrix of symbolic data composed of a diagram value

No	Concept = Operator	A = items
1	1	$\frac{1}{7}a_1, \frac{3}{7}a_2, \frac{1}{7}a_3, \frac{1}{7}a_4, \frac{1}{7}a_5$
2	2	$\frac{1}{3}a_1, \frac{1}{3}a_2, \frac{1}{6}a_3, \frac{1}{6}a_4$
3	3	$\frac{2}{9}a_1, \frac{2}{9}a_2, \frac{1}{3}a_3, \frac{2}{9}a_5$
4	4	$\frac{1}{5}a_1, \frac{2}{5}a_2, \frac{2}{5}a_3$

3. Extended Apriori Method (Apriori Diagram)

3.1.1. Principle of the method

To apply the extended apriori approach, we “discretize” the frequencies of each category of diagrams into two main steps in the form of this algorithm:

START

Step 1: Segmentation into frequency intervals of diagrams

We partition intervals frequency $F_{X_{i,c}}$ for each category of each variable X_i. So, we look at the supports of the frequency intervals $0 < F_{X_{i,c}} \leq \frac{1}{h}, \frac{1}{h} < F_{X_{i,c}} \leq \frac{2}{h}, \frac{2}{h} < F_{X_{i,c}} \leq \frac{3}{h}, ..., \frac{h-1}{h} < F_{X_{i,c}} \leq 1,$ where *h* determines the precision of the cutting.

REPEAT

Step 2: Union of frequency intervals

We make the union 2 by 2 contiguous weight intervals with strictly positive supports $0 < F_{X_{i,c}} \leq \frac{2}{h}, \frac{1}{h} < F_{X_{i,c}} \leq \frac{3}{h}, ..., (h-2) < F_{X_{i,c}} \leq 1.$

UNTIL you get a single interval $0 < F_{X_{i,c}} \leq 1.$

END

As a result, we consider these different interval classifications. Thus, we work with Boolean symbolic object (OS) and no longer with sets of items where the frequency intervals are the properties of the OSs, which therefore have for intensions $a(w) = \left[\frac{a}{h} < F_{X_{i,c}}(\omega) \leq \frac{b}{h}\right] (a = 0 \cdots h-1,\ b = 1 \cdots h,$ $a < b)$. Finally, a*k*-OS is a Boolean assertion defined from *k* properties. For example, considering ε and ε' two categories of two variables diagrams *A* and A' with F_{X_ε}, $F_{X'_{\varepsilon'}}$ their respective frequencies $\left[\frac{1}{3} < F_{X_\varepsilon} \leq \frac{2}{3}\right] \wedge$ $\left[0 < F_{X'_{\varepsilon'}} \leq \frac{2}{3}\right]$ will be a 2-OS.

Note. These *k*-OS will not be fully treated as categories of the classical algorithm. It does not cross intervals of the same category and we must use the smallest possible frequency intervals for the same support each time.

3.1.2. Choosing the precision of the cutting h

User can choose a value of h according to the number of modalities of the variables to be studied and his need for more or less precise results. Of course, the higher the precision, the greater the number of $k = 1$-OS will be in relation to the number of items in the classical case. In return, the transformation of the matrix of classical data into symbolic data will have reduced the number of individuals (or concept) to be studied. For example, referring to Table 3, we can use an accuracy of $h = 3$ (minimum number of categories per concept or $h = 5$ (maximum number of categories per concept) or any other value necessary for precision for the needs of the study to be conducted.

3.1.3. Definition: support, confidence and confidence diagram

Let Ω be a sampling (set of concepts), A and B two OS having as intensions

$$a_x(\omega) = \wedge_{i,u}\left[\frac{a_{i,u}}{h} < F_{X_{i,u}}(\omega) \leq \frac{b_{i,u}}{h}\right]$$

and

$$a_y(\omega) = \wedge_{j,v}\left[\frac{c_{j,v}}{h} < F_{Y_{j,v}}(\omega) \leq \frac{d_{j,v}}{h}\right],$$

$\forall u,\ j,\ v,\ X_{i,u} \neq Y_{j,v},\ \ F_{X_{i,u}}(F_{Y_{j,v}})$ is the category frequency $u(v)$ of diagram variable $X_i(Y_j)$, $\frac{a_{i,u}}{h}$, $\frac{b_{i,u}}{h}\left(\frac{c_{j,v}}{h}, \frac{d_{j,v}}{h}\right)$ are the frequency interval boundaries.

- Definition of support (Supp)

$$\text{Supp}(A \to B) = \frac{\text{Card}(\text{ext}(A \wedge B)) = \{\omega \in \Omega / a_x(\omega) = true,\ a_y(\omega) = true\}}{\text{Card}(\omega)}.$$

- Definition of confidence (Conf)

$$\text{Conf}(A \to B) = \frac{\text{Card}(\text{ext}(A \wedge B)) = \{\omega \in \Omega / a_x(\omega) = true,\ a_y(\omega) = true\}}{\text{Card}(\text{ext}(A)) = \{\omega \in \Omega / a_x(\omega) = true\}}$$

$$= \frac{\text{Supp}(A \to B)}{\text{Supp}(A)}.$$

- Definition of confidence diagram (CD)

By adopting the diagram approach, it is interesting to define a new quality indicator (confidence diagram or CD) penalizing the rules having the largest frequency intervals and therefore the greatest inaccuracy in conclusion:

$$\text{CD}(A \to B) = \text{Conf}(A \to B) \Big/ \left(1 + \frac{\sum_{j,v} (d_{j,v},\ c_{j,v})}{n_v X h}\right),$$

where n_v is the number of properties in conclusion.

Indicator CD is such that:

$$\frac{1}{2}\text{Conf}(A \to B) \le \text{CD}(A \to B) \le \frac{h}{h+1}\text{Conf}(A \to B).$$

To generate the symbolic association rules, we define a minimum CD (MinCD).

- Steps of algorithm

In this section, we present the main steps of the algorithm “apriori diagram”.

In initialization, it is necessary to define the value of the precision *h* according to the needs of the user, the minimum support.

1. Discretize the frequencies of each category.
2. Calculate the supports of the previous weight intervals with a passage in the data matrix. We then make the union 2 to 2 contiguous

intervals of strictly positive supports. We repeat the unions 2 to 2 of our new intervals until we get a single interval $0 < F_{X_{i,c}} \leq 1$. The supports of these intervals are computed without passage in the matrix of the data because if A and A' are contiguous intervals, then $\text{Sup}(A \bigcup A') = \text{Sup}(A) + \text{Sup}(A')$. We add to $L_{k=1}$ the 1-OS of support above the Minsup threshold.

3. Do as long as all k-OS (assertion defined with the conjunction of k frequency intervals) frequent $L_k \neq \varnothing$ $(k \geq 1)$:

   a. Generate $k+1$-OS candidates by calculating the Cartesian product between to k-OS of L_k. In the case of diagrams, we generate $k+1$-OS between intervals of different categories (and "unmarked" see point c). Thus, all candidates C_{k+1} are generated. Due to the property 3, we remove from C_{k+1} all $k+1$-OS I as it exists a k-OS $J \subset I$ not belonging to L_k.

   b. For all $c \in C_{k+1}$, calculate the support with a passage in the data matrix. All $k+1$-OS $I \in C_{k+1}$ frequent is added L_{k+1}.

   c. Mark all $k+1$-OS $I \in L_{k+1} / \exists J \in L_{k+1}$ with $J \subset I$ and $\sup(I) = \sup(J)$. These are $k+1$-OS defined with the same categories but with different weight interval and we only keep the smallest intervals for the same support. We mark them instead of deleting them because these $k+1$-OSs are not used for the generation of $k+2$-OS but they are used for rule generation.

   d. Generate rules with a CD higher than MinCD (minimum confidence diagram).

4. Applications and Results: Classic Rules versus Symbolic Rules

By applying the classical algorithm in Zehero et al. [15], several interesting rules are highlighted. We expose here three rules:

- *Prediction rule* that states that 57% of the risks of fraud in goods clearance come from false declarations of value.

- *Targeting rule* which gives specific information, justifiable by the fact that 47% of the risks of capital flight are essentially related to exchange control operations.

- *Neutral rule*. This rule is not relevant because the information is irrelevant. Indeed, the information it gives does not specify its nature of the risk (indeed, all misappropriations are risks of fraud in a customs clearance operation).

However, even though this information provides information, the different cases of infringements during customs clearance operations do not give any information on the operators' behavior in relation to the risks related to the customs clearance activity.

To apply the apriori algorithm to the diagram data, we worked on 11 operations related to customs clearance activities for 4 different operators. The various offenses identified are: (a) false declaration of value, (b) false declaration of species, and (c) false declaration.

Table 4. Matrix for classic customs clearing operations

Transaction	Operator	*A* = offenses
1	OP 001	*a*, *b*
2	OP 001	*a*, *b*, *c*
3	OP 001	*a*
4	OP 002	*b*, *c*
5	OP 002	*c*
6	OP 003	*a*
7	OP 003	*a*, *b*, *c*
8	OP 003	*a*, *b*, *c*
9	OP 003	*a*
10	OP 004	*a*, *b*
11	OP 004	*a*

Table 5. Symbolic data matrix consisting of a single variable diagram

Concepts = Operators	A = offenses	Concepts = Operators	A = offenses
001	$\frac{1}{2}b, \frac{2}{3}c$	003	$\frac{1}{2}a, \frac{1}{4}b, \frac{1}{4}c$
002	$\frac{1}{3}a, \frac{1}{3}b, \frac{1}{6}c$	004	$\frac{2}{3}a, \frac{1}{3}b$

We record at most 3 cases of infringement by operators (see Table 1). The parameterization of the precision in the algorithm is $h = 3$ (corresponding to the largest number of infractions recorded per operator); the minimum support (Minsup) is set at 70%, the minimum of the confidence chart (MinDC) is set at 55%. It will be noted that the matrix of symbolic association rules (Table 6) with the diagram approach provides more information with details on operator behavior. For example, rule 1 gives more information than rule 4 (see Table 6). Indeed, we know that in addition to making false declarations of value on the goods, operators also defraud on the declaration of cash. Thus, we see that while the "degree of inclusion" of the offense "false statement of value" of the goods in the offense "false declaration of cash" in customs clearance operations is large, the symbolic analysis shows that indeed the operators who defraud by making false declaration of cash are also the same ones who make the false statements of value on the goods. By observing Table 6, the first rule states that the operators who fraud on the declaration of value are the same who make the declaration on the species although they defraud more on the declaration of false value than on the declaration of the species.

Table 6. Symbolic association rules

No	Association rules	Supp	Conf.	CD
1	$\frac{1}{3} < F_a \leq \frac{2}{3} \rightarrow 0 < F_b \leq \frac{1}{3}$	0,70	1	0,70
2	$0 < F_b \leq \frac{1}{3} \rightarrow \frac{1}{3} < F_a \leq \frac{2}{3}$	0,70	1	0,70
3	$0 < F_c \leq \frac{1}{3} \rightarrow 0 < F_b \leq \frac{2}{3}$	0,70	1	0,60
4	$0 < F_b \leq \frac{2}{3} \rightarrow \frac{1}{3} < F_a \leq \frac{2}{3}$	0,70	0,70	0,55
5	$0 < F_b \leq \frac{2}{3} \rightarrow 0 < F_c \leq \frac{1}{3}$	0,70	0,70	0,55

5. Conclusion

Predictive analysis by applying the apriori algorithm to the symbolic variables diagrams made it possible to discover new association rules in a customs database. These rules made it possible to highlight additional information on risky behaviors of certain operators related to customs activities. A selection criterion based on pattern frequency and indicator confidence diagram showed the effectiveness of this approach. However, it would also be interesting to extend this approach to other symbolic variables to discover richer information.

References

[1] R. Gras, Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques, Thèse, Rennes I, 1979.

[2] J. L. Guigues and V. Duquenne, Famille minimale d'implications informatives d'un tableau de données binaires, Mathématiques and Sciences Humaines, année 24(95) (1986), 5-18.

[3] R. Agrawal, T. Imielinski and A. N. Swami, Mining association rules between sets of items in large databases, Peter Buneman and Sushil Jajodia, eds., Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993, pp. 207-216.

[4] R. Agrawal and R. Srikant, Fast algorithms for mining association rules in large databases, Research Report RJ 9839, IBM Almaden Research Center, San Jose, California, 1994.

[5] C. H. Cai, A. W. C. Fu, C. H. Cheng and W. W. Kwong, Mining association rules with weighted items, Proc. of the 1998 Int'l Database Engineering and Applications Symposium (IDEAS'98), 1998, pp. 68-77.

[6] W. Wang, J. Yang and P. Yu, Efficient mining of weighted association rules (WAR), Proc. of the Sixth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, 2000, pp. 270-274.

[7] R. Srikant, Q. Vu and R. Agrawal, Mining association rules with item constraints, Proc. of the 3rd Int'l Conf. on Knowledge Discovery in Databases and Data Mining, 1997.

[8] J. Han and Y. Fu, Discovery of multiple-level association rules from large databases, Proc. of the 21th Int'l Conf. on Very Large Databases, 1995.

[9] C. M. Kuok, A. Fu and M. H. Wong, Mining fuzzy association rules in databases, ACM SIGMOD Record, Vol. 27, 1998, pp. 41-46.

[10] R. Srikant and R. Agrawal, Mining quantitative association rules in large relational tables, Proc. of the ACM-SIGMOD 1996 Conf. on Management of Data, 1996.

[11] R. J. Miller and Y. Yang, Association rules over interval data, Proc. of the 1997 ACM SIGMOD Int'l Conf. on Management of Data, 1997, pp. 452-461.

[12] R. J. Bayardo Jr and R. Agrawal, Mining the most interesting rules, Proc. of the Fifth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, 1999, pp. 145-154.

[13] J. Blanchard, F. Guillet, R. Gras and H. Briand, Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel TIC, revue RNTI, Extraction et Gestion des Connaissances, Cépadues, Paris, Vol. 1, 2004, pp. 287-298.

[14] N. Pasquier, Data Mining : Algorithmes d'Extraction et de Réduction des RA dans les Bases de Données, Thèse, Clermont-Ferrand II, 2000.

[15] Bi Bolou Zehero, Etienne Soro, Yake Gondo, Pacôme Brou and Olivier Asseu, Elicitation of association rules from information on customs offences on the basis of frequent motives, Engineering 10(9) (2018), 588-605.

[16] H.-H. Bock and E. Diday, Analysis of symbolic data, Exploratory Methods for Extracting Statistical Information from Complex Data, Springer-Verlag, Heidelberg, 2000.