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*], Daniel Bourget[4]

[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
[4]Institut Mines Telecom Atlantique, Brest, France
Email: *oasseu@yahoo.fr, *zeherobi@yahoo.fr, *broupacom@hotmail.fr

How to cite this paper: Zehero, B.B., Soro, E., Gondo, Y., Brou, P., Asseu, O. and Bourget, D. (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**).

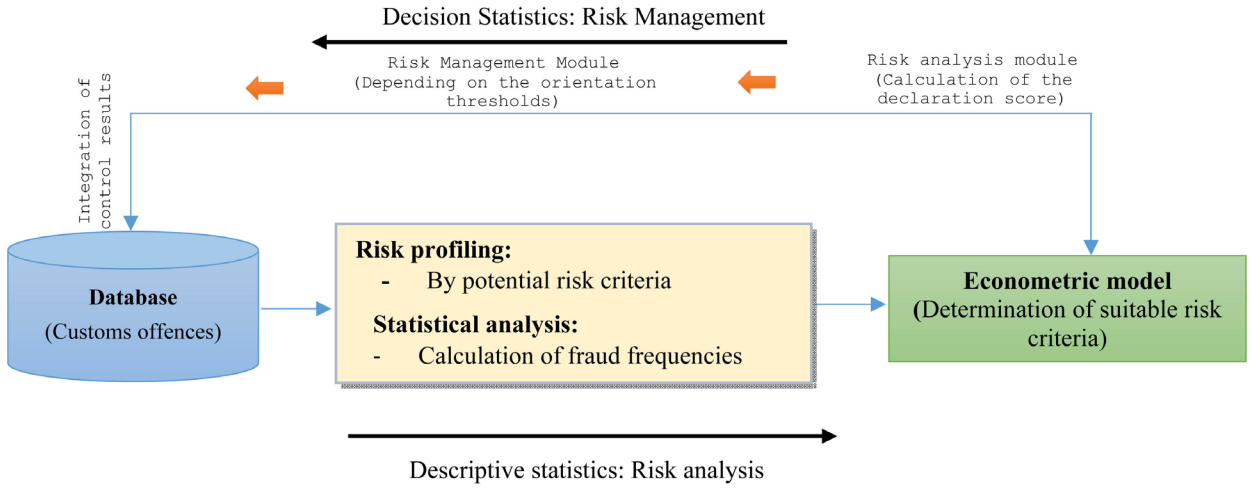

Figure 1. Risk analysis in a customs system using an econometric model.

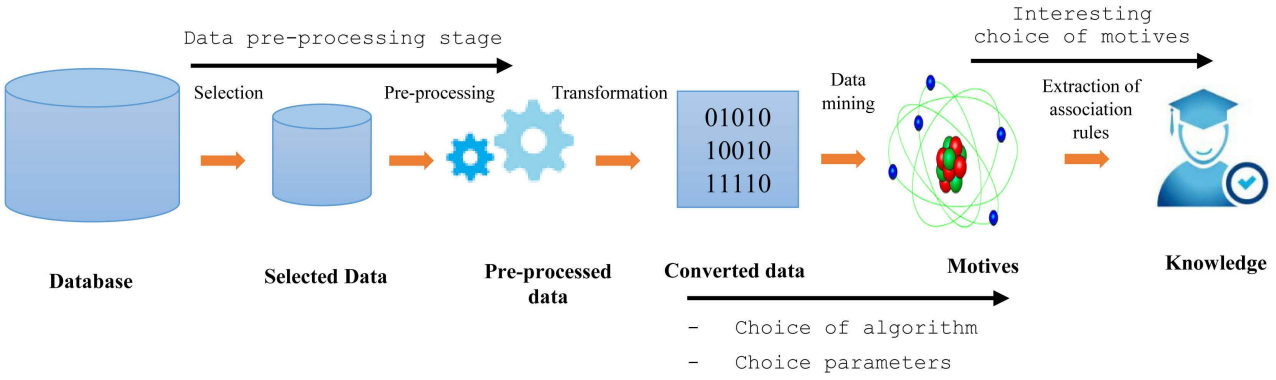

Figure 2. The steps for extract association rules.

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$\in A$**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

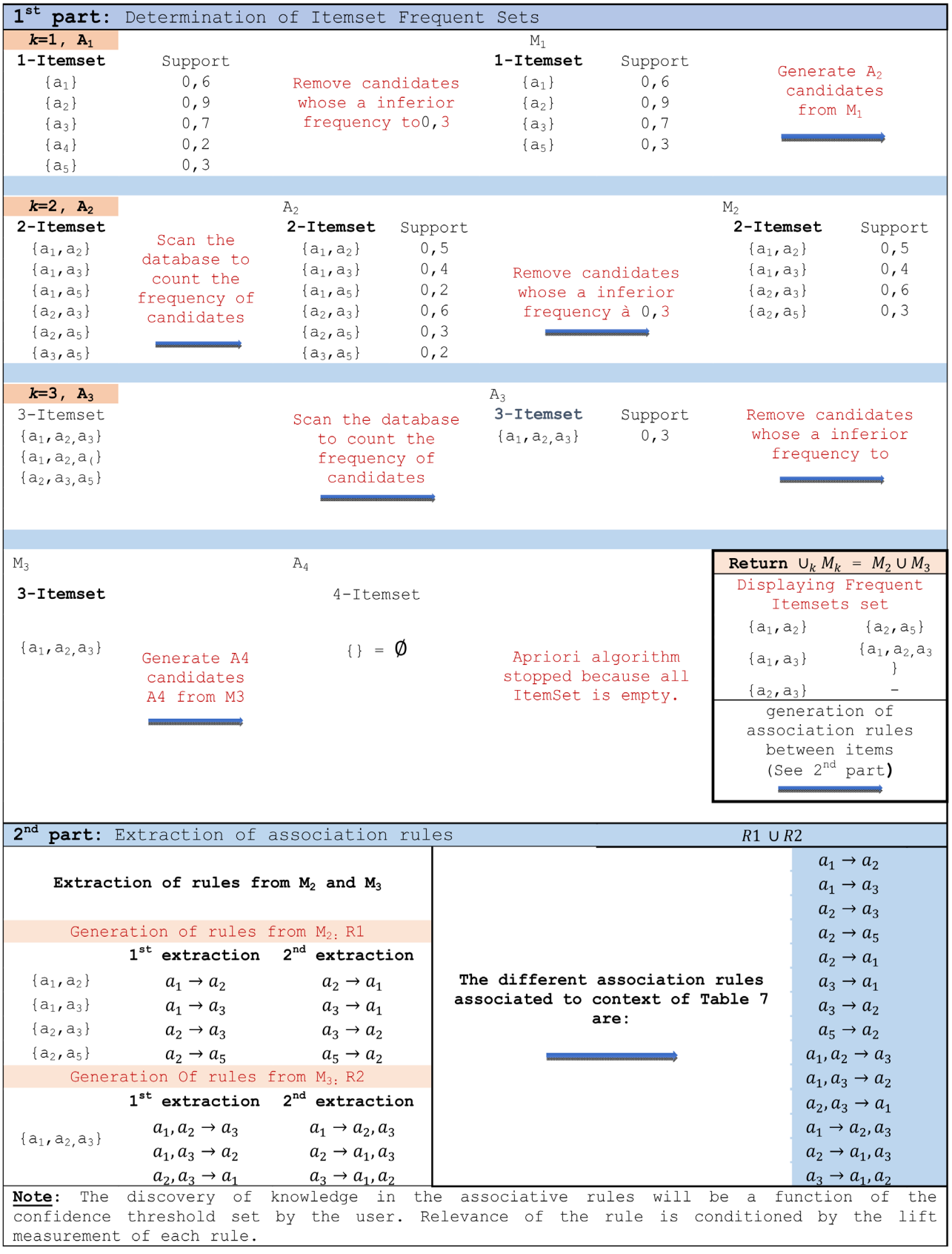

Figure 6. Illustration of the Apriori algorithm from the context described in **Table 3**.

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.