Uses of Artificial Intelligence in the Brazilian Customs Fraud Detection System

Luciano A. Digiampietri*
Institute of Computing
Av. Albert Einstein, 1251
13084-971 Campinas, SP
(BRAZIL)

Norton Trevisan Roman*
Institute of Computing
Av. Albert Einstein, 1251
13084-971 Campinas, SP
(BRAZIL)

Luis A. A. Meira*
Institute of Computing
Av. Albert Einstein, 1251
13084-971 Campinas, SP
(BRAZIL)

Jorge Jambeiro Filho†
Brazil's Federal Revenue
Rodovia Santos Dummont,
Km 66
13055-900 Campinas, SP
(BRAZIL)

Cristiano D. Ferreira*
Institute of Computing
Av. Albert Einstein, 1251
13084-971 Campinas, SP
(BRAZIL)

Andreia A. Kondo*
Institute of Computing
Av. Albert Einstein, 1251
13084-971 Campinas, SP
(BRAZIL)

ABSTRACT

There is an increasing concern about the control of customs operations. While globalization incentives the opening of the market, increasing amounts of imports and exports have been used to conceal several illicit activities, such as, tax evasion, smuggling, money laundry, and drug traffic. This fact makes it paramount for governments to find automatic or semi-automatic solutions to guide the customs' activities in order to minimize the number of manual inspections of goods. In this context, this paper presents an overview of some approaches developed in the HARPIA project that is a partnership between universities and the Brazilian Federal Revenue for the development of computational intelligence solutions to the management of customs risk.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; J.1 [**Administrative Data Processing**]: Government

Keywords

E-government, fraud detection, outlier detection

1. INTRODUCTION

Imports and exports are fundamental aspects of the global economy. Goods are typically taxed proportionally to their value, varying according to the type of product. Each product is classified following a specific classification system. For Mercosul [12], this classification system is called NCM (Mercosul Common Nomenclature), which is similar to the "Harmonized Commodity Description and Coding System" used by World Customs Organization (WCO) [15]. This classification system has approximately ten thousand different codes that describe products categories (instead of specific products). Often, it is not trivial to assign a category code to a product due to the great number of categories and the fact that descriptions of some categories are abstract. Moreover, many importers assign an incorrect category to products in order to pay a smaller tax. However, product misclassification is only one of several frauds related to customs operations [14]. We highlight other kinds of fraud: overvaluation, undervaluation, smuggling and drug traffic. All these kinds of fraud can be used to support terrorists, drug traffickers and organized crime in general.

Each country is responsible to inspect the customs operations in order to identify frauds and punish the transgressors. Given the limited amount of available resources, it became impossible to inspect all the customs operations and identify all frauds. The goal of this paper is to describe part of an ongoing project, called HARPIA[3]. This project is a partnership between Brazilian universities and the Brazilian Federal Revenue for detecting several types of fraud through the application of artificial intelligence. In this paper we describe two aspects of this project: (i) an outlier based detection system that helps customs officers to identify suspicious customs operations; and (ii) a product and foreign exporter information system that aims to help the importers in the registration and classification of their products and corresponding exporters.

The rest of this paper is organized as follows. Section 2 presents the related work. Section 3 describes our approach to the problems of identifying suspicious customs operations and registering goods and exporters. Section 4 presents the conclusions and future steps.

2. RELATED WORK

Detecting fraud using normal audit procedures is an expensive and a laborious task. There are few customs officers that have the necessary expertise and hundreds (or sometimes, thousands or even millions) of operations that must be verified. This brings up a new challenge: how to construct computational solutions to automatically or semi-automatically identify suspicious operations. Data mining and statistical approaches are being applied to try to identify these fraudulent operations.

[1] {luciano.digiampietri,nortontr,augustomeira, crferreira,andreia.kondo}@gmail.com
[2] jorge.filho@jambeiro.com.br

[3] HARPIA: Risk Analysis and Applied Artificial Intelligence

The detection of suspicious activities is a problem in several domains, such as, credit card fraud, telecommunications fraud, terrorism detection, financial crime detection, and computer intrusion detection. Detecting fraud is essential as prevention mechanisms fail [17] and a good detection system must be self-adaptive to detect new fraudulent behaviors.

There are several approaches to deal with fraud detection. We highlight the use of neural networks [3, 5], bayesian networks [11], expert systems [2], rule based systems [1] and the detection of statistical outliers [6, 14, 16]. These approaches can be subdivided in two groups: supervised and unsupervised. In the supervised approaches there is a training set of operations that are labeled either as *fraudulent* or *normal*. These operations are used as input to some systems, such as neural network systems, that need labeled inputs to construct the model that will be used to detect frauds.

The use of supervised learning by Brazilian customs to select goods for human verification was originally described in [4]. Alternative strategies have been employed in [7] without benefits, but a novel approach, described in [8], achieved significant improvements in some performance measures. The unsupervised approaches do not need labeled inputs, as they use a set of rules to classify an operation as a fraud or compare each one with the previous operations to identify those that might be considered suspicious (outliers).

Rule based systems are unsupervised approaches that use a set of rules to classify the operations as *fraudulent* or *normal*, or to assign a value to each operation corresponding to the chance an operation has to be a fraud. The rules are typically constructed following the advises of experts. These systems have the advantage of being unsupervised and taking account of the experts' knowledge to construct the rules that evaluate each operation. One of the disadvantages of these systems is the fact that the rules frequently need to be updated to deal with new fraudulent behaviors. Otherwise, the rules will eventually become obsolete.

The identification of frauds using outlier detection (e.g. [14, 16]) is an unsupervised approach that identifies suspicious operations comparing each operation with the previous ones. One advantage of this approach is the capability to adapt (and identify) new behaviors while new operations are stored in the system. Another advantage is the clear statistical meaning that is assign to each suspicious operation. For example, the system can calculate that one operation is four standard deviations away from its expected value and that this happens only once in one thousand operations. This operation is an outlier and deserves to be inspected (as it is a suspicious operation).

One important prerequisite of outlier detection systems for fraud detection is that the majority of the operations stored in the system must be normal (not fraudulent). Moreover it is important to emphasize that being an outlier does not mean to be a fraud. Besides this assumption, it is also important to ensure that the importers, exporters and products are correctly registered and classified.

Every day, hundreds or even thousands of import declarations are written. Since there is no global database of foreign companies and products, each importer must re-type the name, description and classification of the products and the name of the company (exporter) that sold them. This process is susceptible to several kinds of errors. We highlight (i) the misclassification of products (because it is a laborious work to assign one of the ten thousand categories to each product), and (ii) the registration of companies or products with mistakes such as misspelling. To avoid these two problems a common approach is the development of spell verification systems and/or approximate search engines that try to identify what the user is trying to type.

The first approach that was used in the HARPIA project to avoid redundancy in the Brazilian's foreign companies database was based on a modified edit-distance algorithm [13]. This solution extended the edit distance proposed by Levenshtein [10]. The main idea of the modified algorithm is to break the strings into words, compare and compute the distance between them, and search for the minimum cost "path" that links them together. See [13] for details about this approach.

The edit-distance based approach presented good initial results but it was not robust enough to deal with all problems in the products and foreign exporter database. Section 3.2 presents a more complex approach using Markov Chain and n-grams [9].

3. OUR APPROACH

Our approach to identifying possible frauds is based on the interaction between the customs officer and the decision support system we developed [14]. This system, called Carancho, highlights suspicious operations through outlier detection. It assumes that the majority of the international commerce operations are correct, *i.e.*, they are in accordance to the law and the products are correctly classified.

Due to the great amount of products and companies (exporters, importers, transporters, etc) it is very difficult to ensure that the products and the companies are correctly classified, avoiding misclassification or multiple registration of the same company. To solve this problem, we are developing a Product and Foreign Exporter Information System (PFEIS) that uses features from orthographic verification to suggest possible duplicities (*i.e.* when the user tries to register an already registered company or product) and to help on their classification.

Although both systems may seem only loosely related, they actually draw on a bigger picture, as shown in Figure 1, which presents some of the main modules that build up the artificial intelligence part of the HARPIA architecture. This figure also illustrates the strategies followed by the HARPIA project to tackle the problem of customs fraud detection. These strategies, in turn, concentrate mainly on (i) building a reliable database of products and foreign exporter (PFEIS), (ii) trying to identify suspicious operations before (Carancho) and after (ANACOM) clearance, and (iii) controlling for small imports coming to the country through the express mailing service. Thus, as it can be seem, both PFEIS and Carancho are linked together by the former building the dataset needed in the later. This paper describes only the

Carancho and the Product and Foreign Exporter Information System modules.

3.1 THE CARANCHO SYSTEM

Instead of trying to formulate an exhaustive set of rules to cover the broadest number of frauds possible, the approach we followed relies upon the graphical visualization of historical import/export data (see Carancho [14]). In a nutshell, it takes the historical record of import operations as a start point and presents it to the user. The user then can check whether some specific transaction can be considered an outlier according to a number of predefined dimensions.

As our main interest is detecting under and overvaluation, we have chosen a set of dimensions thought to be sensitive to such problems, according to the customs officers' practical experience and expertise. Then, in a sense, this approach combines the visual detection of outliers with the officers' empirical knowledge.

The main advantage in using this approach comes up more clearly when the trading system changes, as when the goods classification scheme changes, for instance. While these changes would demand the set of rules mapping conditions to consequences to be updated, some dimensions (like weight and price, for example) remain untouched, *i.e.*, they still can be used to characterize any import operation. That fact makes them a naturally long-lasting choice for detecting any abnormal behavior. The same way, changes in the importers' behavior, that otherwise would also demand updating the set of rules, are naturally captured by this approach, as it accounts for the whole amount of import operations that took place in some time range.

To verify the practical applicability of this idea, we have developed a computer system capable of analyzing the whole set of data and show it to the user in a way s/he can clearly spot any outliers (Figure 2). The rationale behind this approach is that it allows for an automatic outlier detection technique to be used alongside the user's decisions, either concurrently or giving them support.

Originally designed to deal with only three dimensions, the first version of this system shows the data distribution according to the predefined dimensions, along with the operation under evaluation (portrayed as a thin horizontal red line in Figure 3). However, such an approach presents a major limitation to the user, namely, it only allows for data to be analyzed in one single axis (as it is a histogram), thereby losing any information concerning the relation that different dimensions might hold with each other.

To deal with this shortcoming, and once more based on the customs officers' expertise, we have redesigned the way the system outputs the data. The new representation, as illustrated in Figure 4, deals with pairs of dimensions, allowing the user to determine any trend that might exist inside each pair. In this Figure, the four importers responsible for the highest amount of operations (numbered 0 to 3) are portrayed on different shapes and colors. A fifth shape (and corresponding color) is reserved for the rest of the data, *i.e.* the data coming from all the remaining importers.

As one may notice, this representation lacks information about the relative amount of import operations for a specific pair of dimensions, thereby lacking the very information needed to give the user some insight about the importance of a specific point (like an outlier, for example). Even worse, the overlapping points might generate some distortion in the coloring scheme, hiding some points out and perhaps rendering the whole visualization less reliable.

To avoid these drawbacks, the user can tick the "Densidade 2D" (2D Density) box, bringing the density of operations on to the picture. The system, in turn, colors each point according to the relative amount of imports it might contain, from yellow (the groups with fewer import operations) to red (the groups with the higher number of operations among the data). When the system colors some point, it does so using a Normal curve for intensity, *i.e.*, the color smoothes out as it moves away from the point, as illustrated in Figure 5.

Although this new representation seems to sort out most of the problems with the data visualization, it still suffers from a fundamental difficulty, namely, the considerably high degree of subjectivity brought to the system by the current goods classification scheme (*i.e.* the NCM). This subjectivity, which lets considerably different products be correctly classified in the same category, has the undesired property of grouping together very sparse data, thereby making it difficult for the user to determine what an outlier would look like, given such a dataset.

The solution we found to this problem was to develop a registration system to identify each foreign exporter and his/her corresponding exported goods. This system, described in the next section, should be able to evolve over time, naturally adapting to the new products brought forth by the market (and to new exporters coming into it), without any intervention from the customs office. Once it is accomplished, the system would give every exporter and product a unique identifier, allowing Carancho to group together only products that are really close to each other, thereby increasing the reliability of its output.

3.2 PRODUCT AND FOREIGN EXPORTER INFORMATION SYSTEM

It is a difficult task for the Brazilian Federal Revenue to create unique identifiers to companies situated out of Brazil. This requirement appears every time these companies buy or sell goods across our frontiers. Without a unique identifier, a foreign company can be repeatedly fraudulent without any special attention from the Federal Revenue and it can be treated as if it was a new enterprise at each transaction. To cope with this problem, we are developing a catalog to assign unique identifiers to each company. This catalog aims to minimize redundancy, by providing the importer with a search engine, so that s/he can search for previous registration of a company before registering it again.

The effort to keep foreign enterprises correctly registered can be naturally extended to products commercialized among them. The goods that enter or leave the country have similar demands for unique identifiers. These identifiers are desirable to facilitate automatic or semi-automatic fraud detection system (see Section 3.1). Inside the HARPIA project,

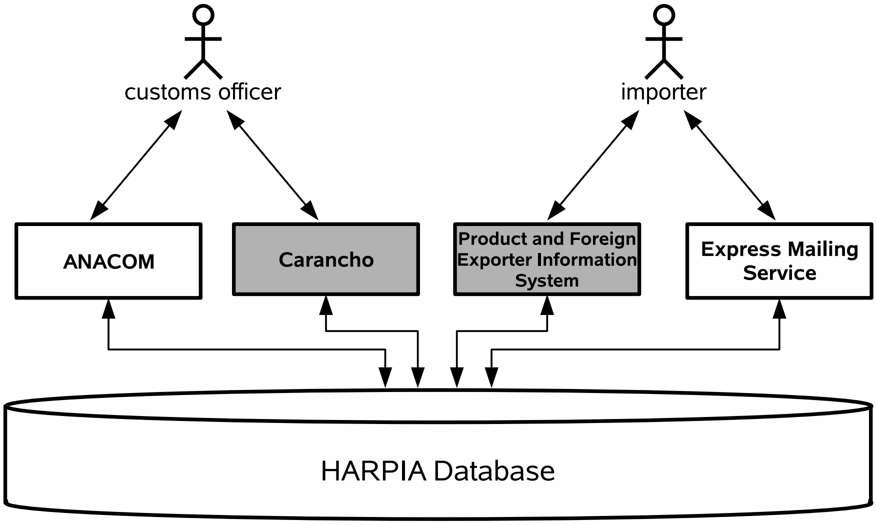
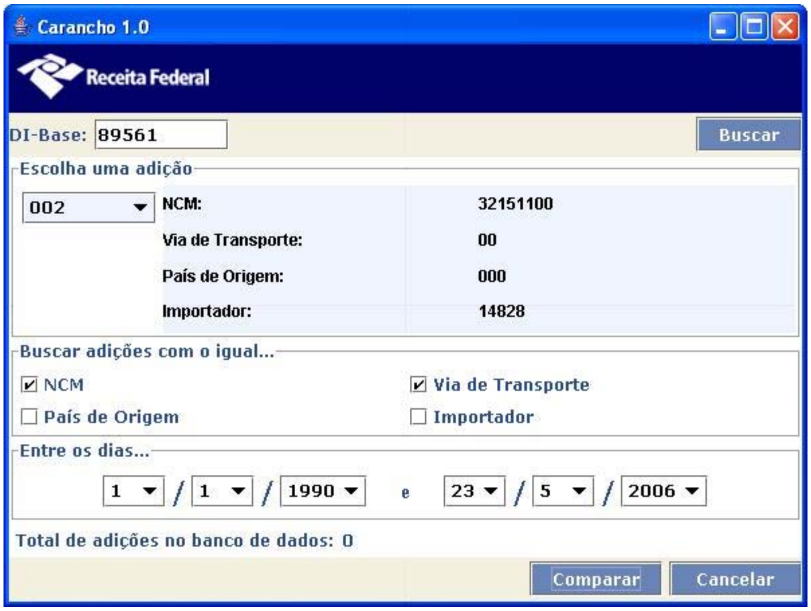


Figure 1: Part of HARPIA's AI architecture.



DI-Base: 89561

Escolha uma adição

002

NCM:	32151100
Via de Transporte:	00
País de Origem:	000
Importador:	14828

Buscar adições com o igual...

☑ NCM ☑ Via de Transporte

□ País de Origem □ Importador

Entre os dias...

1 / 1 / 1990 e 23 / 5 / 2006

Total de adições no banco de dados: 0

Figure 2: The system's interface.

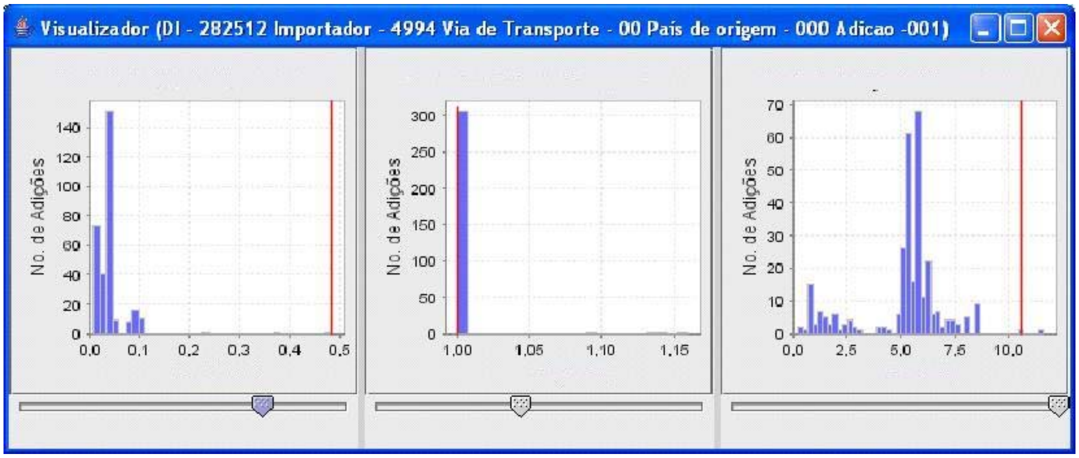
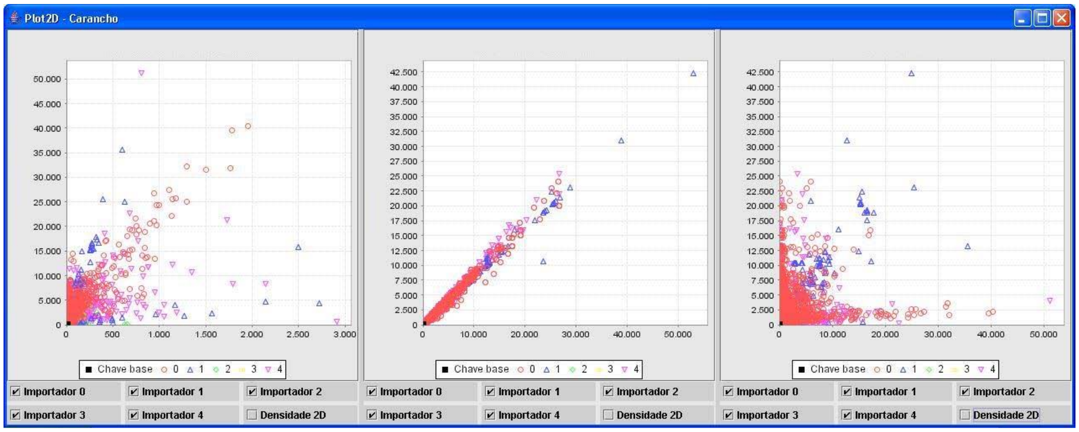


Figure 3: Output of the system's first version.

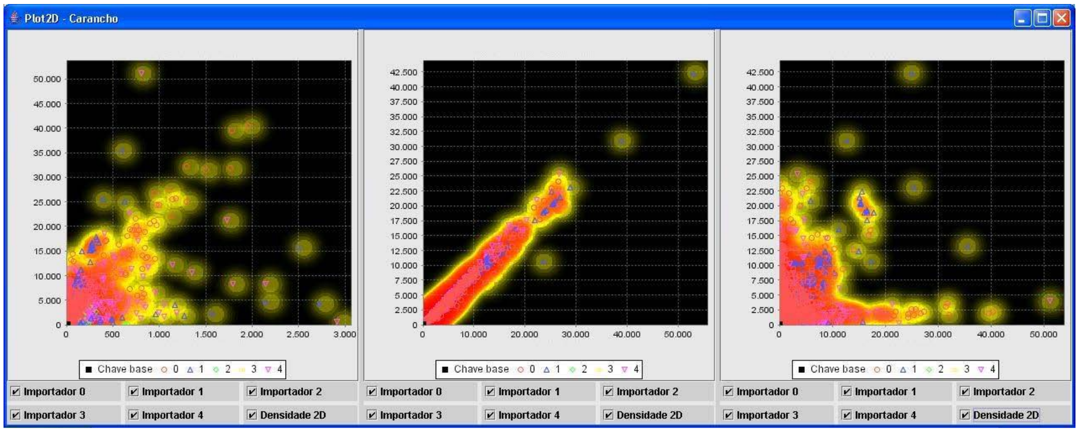


☑ Importador 0 ☑ Importador 1 ☑ Importador 2 ☑ Importador 3 ☑ Importador 4 □ Densidade 2D

☑ Importador 0 ☑ Importador 1 ☑ Importador 2 ☑ Importador 3 ☑ Importador 4 □ Densidade 2D

☑ Importador 0 ☑ Importador 1 ☑ Importador 2 ☑ Importador 3 ☑ Importador 4 □ Densidade 2D

Figure 4: Relationship between the data in each pair of dimensions.



☑ Importador 0 ☑ Importador 1 ☑ Importador 2 ☑ Importador 3 ☑ Importador 4 ☑ Densidade 2D

☑ Importador 0 ☑ Importador 1 ☑ Importador 2 ☑ Importador 3 ☑ Importador 4 ☑ Densidade 2D

☑ Importador 0 ☑ Importador 1 ☑ Importador 2 ☑ Importador 3 ☑ Importador 4 ☑ Densidade 2D

Figure 5: Relationship between the dimensions (and their density).

two catalogs are being developed: the *Product Catalog System* and the *Foreign Importer/Exporter Catalog System.*

National enterprises that trade with other countries are uniquely identified in Brazil by the CNPJ number, which is a unique identifier provided by the Federal Revenue. When these enterprises make an international transaction, the Brazilian Federal Revenue will have them register their international partners, following a specific protocol. First, the user designated by the national enterprise queries the Importer/Exporter Catalog looking for the partner company. The system, in turn, looks up the database, returning any match it finds, ranked according to a probability function.

If, on the other hand, no satisfactory match is found, the national company can create and register its foreign partner in the system. Once this operation is confirmed, the new foreign company is registered in the catalog and a unique identifier is created. This identifier can then be used by the national company whenever it makes an international transaction with the foreign company it represents. The same procedure will be followed whenever the importer tries to register a new product.

There is, however, more about these catalogs than a simple search engine. The users of a search engine are very interested in finding whatever they describe in their queries. Companies which want to commit frauds do not want their foreign partners or the products they are pursuing to be recognized. To carry out this task in a proper way, we need to care about spelling errors, *i.e.*, we must take into account, among other things, the possibility that the user mistypes his/her query. Also, the system must be able to identify and correct multiple instances of the same company or product. To do so, the catalogs have a built-in probabilistic spelling checker, along with methods for insertion, deletion, merging and correction of records, in an attempt to keep the database consistency.

The spelling checker's implementation is based on Markov Chains and n-grams. These techniques are used mainly for calculating a word similarity value, based on string matching operations, and to calculate the probability that a given string is, in fact, a valid word in a given domain. Observe that, in a multi-language domain of proper names, new words can be considered neither wrong nor right, for there is no proper lexicon to match them against.

Under these constraints, the system must deal with unreliable information, that is, a dataset that might also contain ill-formed strings, being potentially as problematic as the query string from the user. For this reason, our systems use a probabilistic model that takes into account the commonest misspelling errors, keyboard character position, and the semantics for a set of special words, like "international", "ltd" and "co", for instance. Whenever a new product or enterprise is inserted in the catalog, its words are added in the vocabularies and the probabilities and frequencies are updated.

4. CONCLUSIONS AND FUTURE WORK

Fraud detection systems in customs operations are very important to minimize the manual inspection of goods and maximize the number of frauds detected. They are complex systems that must deal with several problems, such as, high cardinality attributes, imbalanced databases, and misspelling problems.

In this paper, we presented some artificial intelligence approaches used in the Brazilian's customs fraud detection system. The main contributions are (i) the ability to help identify outliers (suspicious operations), and (ii) the products and foreign exporters information system (including databases and tools to identify redundancies and to suggest a category to each product).

As for future work, we are currently developing some automatic outlier detection techniques, which will be used in conjunction with the visual techniques to show the user both the graphics and the probability values. These values, in turn, would represent the system's confidence that, according to the current dataset, a given product actually costs the amount declared by the importer.

5. ADDITIONAL AUTHORS

Additional authors:
Everton R. Constantino (Institute of Computing, UNICAMP, email: constantino.everton@gmail.com),
Rodrigo Rezende (Institute of Computing, UNICAMP, email: rcrezende@gmail.com),
Bruno C. Brandao (Institute of Computing, UNICAMP, email: brunocedraz@gmail.com),
Helder S. Ribeiro (Institute of Computing, UNICAMP, email: helder@gmail.com),
Pietro K. Carolino (CLE-IFCH, UNICAMP, email: helder@gmail.com),
Antonella Lanna (Brazil's Federal Revenue, email: antonella.lanna@gmail.com),
Jacques Wainer (Institute of Computing, UNICAMP, email: wainer@ic.unicamp.br) and
Siome Goldenstein (Institute of Computing, UNICAMP, email: siome@ic.unicamp.br) .

6. REFERENCES

[1] A. Deshmukh and T. Talluru. A rule based fuzzy reasoning system for assessing the risk of management fraud. *Journal of Intelligent Systems in Accounting, Finance & Management*, 4:669–673, 1997.

[2] M. M. Eining, D. R. Jones, and J. K. Loebbecke. Reliance on decision aids: an examination of auditors assessment of management fraud. *Auditing: A Journal of Practice and Theory*, 16(2):1–19, 1997.

[3] K. Fanning and K. Cogger. Neural network detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 17(1):21–24, 1998.

[4] M. A. C. Ferreira. Uso de redes de crença para seleção de declarações de importação. Master's thesis, Instituto Tecnológico de Aeronáutica, 2003.

[5] B. P. Green and J. H. Choi. Assessing the risk of management fraud through neural network technology. *Auditing: A Journal of Practice and Theory*, 16(1):14–28, 1997.

[6] V. Hodge and J. Austin. A survey of outlier detection

methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[7] J. Jambeiro Filho and J. Wainer. Analyzing Bayesian networks with local structure and cardinality reduction over a practical case. In *Proceedings of the Workshop on Computational Intelligence (WCI)*, 2006.

[8] J. Jambeiro Filho and J. Wainer. Using a hierarchical Bayesian model to handle high cardinality attributes with relevant interactions in a classification problem. In *Proceedings of the International Joint Conference of Artificial Intelligence (IJCAI)*. AAAI Press, 2007.

[9] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Englewood Cliffs, New Jersey, 2000.

[10] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[11] S. Maes, K. Tuyls, B. Vanschoenwinkel, and B. Manderick. Credit card fraud detection using Bayesian and neural networks. In *Proceedings of the 1st International NAISO Congress on Neuro Fuzzy Technologies*, 2002.

[12] Mercosul/Mercosur – Southern Common Market. http://www.mercosur.int/msweb/ (as of 2007-10-25).

[13] B. W. Paleo, C. G. G. Hita, J. C. Lima, C. H. Ribeiro, and J. Jambeiro Filho. A modified edit-distance algorithm for record linkage in a database of companies. In *Proceedings of the 2nd Workshop em Algoritmos e Aplicações de Mineração de Dados (WAAMD)*, 2006.

[14] N. T. Roman, E. R. Constantino, H. Ribeiro, J. J. Filho, A. Lanna, S. K. Goldenstein, and J. Wainer. Carancho – a decision support system for customs. In *Proceedings of ECML PKDD Workshop on Practical Data Mining: Applications, Experiences and Challenges*, pages 100–103, September 2006.

[15] World Customs Organization. http://www2.wcoomd.org/ie/index.html (as of 2007-10-25).

[16] K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery*, 8(3):275–300, 2004.

[17] D. Yue, X. Wu, Y. Wang, Y. Li, and C.-H. Chu. A review of data mining-based financial fraud detection research. In *International Conference on Wireless Communications, Networking and Mobile Computing (WiCom)*, pages 5514–5517, September 2007.