INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Classification for Fraud Detection with Social Network Analysis

Miguel Pironet San-Bento Almeida

Dissertation for the obtaining of a Masters Degree in
Engenharia Informática e de Computadores

Júri
Presidente: Nuno Mamede
Orientador: Cláudia Antunes
Vogal: Arlindo Oliveira
Vogal: Pedro Moura

Outubro 2009

Abstract

Worldwide fraud leads to large losses to State treasuries and to private enterprises. Due to this fact, efforts to detect and fight fraud are high. Despite continuous endeavourers, this mission is far from being accomplished. The problems faced when trying to characterize fraudulent activities are many, with the specificities of fraud of each business type leading the list. For this reason, the differences, developing a classification model for fraud detection almost always requires dealing with unbalanced datasets, since fraudulent records are usually in a small number, when compared with the non-fraudulent ones. This dissertation describes two types of techniques to deal with fraud detection: techniques deployed at a pre-processing level, where the goal is to balance the dataset, and techniques applied at a processing level, where the objective is to apply different errors costs to fraudulent and non-fraudulent cases. Besides this, as organizations and people, more often, establish themselves as associations in order to commit fraud, a new method to make use of that information to improve the training of classifiers for fraud detection is purposed. This new method identifies, in particular, patterns among the social networks for fraudulent organizations, to use them to strengthen the description of its entity. The enriched data will then be used, jointly, with balancing techniques to produce a better classifier to identify fraud.

Keywords: Fraud Detection, Classification, Unbalanced Datasets, Social Networks, Patterns

Resumo

Em todo o mundo a fraude conduz a grandes perdas nos orçamentos de Estado e de companhias. Devido a isso, as motivações para detectar e combater a fraude são elevadas, mas, apesar de esforços contínuos, esta tarefa ainda está longe de ser concluída com sucesso. Os problemas que são necessários enfrentar ao tentar caracterizar as actividades fraudulentas são muitos, sendo as especificidades de cada tipo de fraude, em cada diferente área de negócio, um dos principais. Apesar das diferenças, a construção de um classificador para a detecção de fraude quase sempre necessita de lidar com conjuntos de dados não balanceados já que os registos fraudulentos estão em muito menor número quando comparado com o número de registos não fraudulentos. Este trabalho descreve dois tipos de técnicas para lidar com a

detecção de fraude: técnicas ao nível do pré-processamento em que o objectivo é tornar o conjunto de dados mais balanceado e técnicas ao nível do processamento em que o objectivo é usar um custo de erro diferente à classe fraudulenta e à não fraudulenta. Para além desses problemas como as organizações e as pessoas muitas vezes se associam para cometer fraude, é proposto um novo método em que se faz uso da informação sobre essas associações para melhorar o treino dos classificadores. Em particular, este método identifica padrões entre as redes sociais das organizações fraudulentas e usa essa informação para enriquecer a descrição daquela entidade. Os dados enriquecidos vão ser usados juntamente com as técnicas de balanceamento de dados para gerar um melhor classificador para identificar fraude.

Palavras-Chave: Detecção de Fraude, Classificação, Conjunto de Dados Não Balanceado, Redes Sociais, Padrões

Contents

Figures Contents

1 Introduction

Fraud is a crime where the purpose is to appropriate money by an illegal form. Fraud leads to large losses to businesses or States. Since its detection is complex, there is not yet a fraud detection framework that can detect and prevent fraud in an efficient way. Almost any business enterprise that involves money and services can be compromised by fraudulent acts. Such areas as Insurance, Telecommunications, Financial and Credit Cards are examples of where fraud may occur and where, in the recent past, there has been an effort to develop methods to combat this type of financial fraud. Each one of the areas has specific characteristics; therefore, a single solution that may be deployed to fight fraud in Credit Cards cannot be applied to Insurance industry. As a consequence, one of the approaches to fight fraud is to pursue a model that describes fraudulent behaviors, or, better, create mechanisms that distinguish fraudulent from non-fraudulent behaviors. Technically, these mechanisms can be created using a data mining classification, making use of a set of historical records, i.e., records of past "customers" already known as fraudulent and non-fraudulent (the training set). However, applying classification techniques for fighting fraud always deal with a particularity: the existence of an unbalanced training dataset – a set that has many non-fraudulent records and only a few fraudulent cases. Because of this characteristic, the result of applying traditional classification techniques, like decision trees or neural networks, are not enough for obtaining a good classifier.

In the context of classification, there are three major approaches that may used to create a classifier for fraud detection: balance the data, consider different errors costs and detect outliers. This paper describes defines fraud and what the problems associated with classification algorithms, when one attempts to detect and predict fraud. Along with this description, it presents a set of most used techniques to face the stated problems. Lastly, it comments on the particularities of some business areas that are affected by fraud.

The specific problem that will be addressed is this paper is fraud detection at the "Direcção Geral de Impostos" (that is in a free translation to the State Taxes Organization) on the Value Added Tax (VAT), in particular on the detection concerning Carrousel fraud cases. Thus, one of the objectives of this paper is to develop a fraud classifier, with an acceptable accuracy level for fraudulent cases. Moreover, this must be accomplished by solving or minimizing financial fraud detection issues. In order to obtain this goal, several classifiers will be created, implementing the most relevant approaches

for fraud detection, as techniques to balance the dataset or algorithms that learn with different errors costs.

Along with this, it is important to remember that fraud is perpetuated by people or by organizations that cannot detach themselves from the rest of the world. As a result, fraudulent persons and organizations are connected among themselves and to the rest of all other honest and fair organizations. It is, in fact, is important not to deal with entities separately, when looking for fraud but, also, to deal with the relationships between these entities. As an example, the owner of a fraudulent organization is probably the person responsible for the fraud itself. If he owns another organization, it is more likely that this second organization is also fraudulent. Another issue is the fact that some take advantage of carrousel fraud, which can only be accomplished with a link of several organizations. Because of this, the analysis of social networks within fraudulent organizations and its people can be extremely important when searching for fraud.

In this manner, with the work presented here, the social networks of each organization will be analyzed in order to detect patterns that are common to fraudulent organizations. These patterns will enforce the data in VAT declarations, in order to create a dataset with more useful information. The dataset is the base support for the creation of new classifiers, which will be much more accurate than the original ones. With this new dataset that, although having more information about each organization, is still an unbalanced dataset, to which it will be also applied certain techniques do deal with that specific problem.

The rest of this document is structured as follows: section 2 presents the problem statement, section 3 shows a comparative study of techniques to deal with the unbalanced dataset problems, in section 4 is described the approach to improve fraud detection with social networks, section 5 presents the case studies with the use of Social Networks information and the techniques to handle unbalanced datasets and, finally, the last section presents the conclusions.

2. Problem Statement

Fraud is defined by the Association of Certified Fraud Examiners as “the use of one’s occupation for personal enrichment through the deliberate misuse or application of the employing organization’s resources or assets” and by the Cambridge Advanced Learner’s Dictionary as “the crime of obtaining money by deceiving people”. These definitions indicate that fraud is, in fact, a crime in which the its objective is to obtain money in an illegal form, simply by taking advantage of a position or power in a business organization and use them for inappropriate intentions along with the resources and tools of the organization. This in appropriation of goods and services can be accomplished by a single person, with the intention of personal gains or collectively by an organization itself. It is a given that fraud is a crime as old as time. Nowadays, with the evolution of technology, supported by global and speedy communication network, is easier and faster to perpetuate fraud, being much more complicated to be detected. Fraudulent activities lead to larges financial losses worldwide for both business and States. For this reason, it is extremely important at the present in time of economic crises to detect and combat fraud, in order to possibly recover money taken. Along with this, with the evolution of technology, new types of fraud have arisen, such as telecommunication fraud, computer intrusion, credit cards and insurance fraud. Each business area carries its own special features and so the methods for detecting and fighting fraud must be different, when dealing with each type of business. To complicate matters even more, huge amounts of data is generated every year and, so, more efficient techniques to mine that data searching for fraudulent information is necessary (Kou, 2004) (Hilas, 2005).

One way to analyze and combat fraud is by applying Data Mining, which is the exploratory phase of gaining knowledge in databases process (KDD). This task is defined as “the non-trivial extraction of implicit, previous unknown, and potentially useful information from data” (Adriaans, 1996).

Data mining employes a set of techniques that help to reach at vital information that may lead to fraud data. One of these techniques is the construction of a classification model, which attempts to identify indicative patterns of fraud in different areas of bussiness (Weatherford, 2002).

The ultimate goal of applying data mining to fraud detection is to create a classification model that can label a record, person or company as being fraudulent or not. The model is created by analyzing a dataset with records classified as fraudulent and non-fraudulent. This process goes on to locate sets of functions that may describe and distinguish those

records that seem to be inappropriate in an efficient manner. After creating the model, one may use it to classify future and unknown data as fraudulent or non-fraudulent, permitting the user to acknowledge which persons or entities are perpetuating fraud. (Zhao, 2006).

The data used to enable fraud detection has peculiar characteristics that cause three main problems worth addressing, which must be dealt with when working with the data. The first main problem is caused by the unbalanced distribution of the data, where one has to deal with datasets that have very few fraudulent records, compared with the non-fraudulent ones. In these datasets, where a class has the majority of the cases, standard classifiers tend to ignore the minority data class (Vo, 2007). The second problem is the different error costs of the misclassification of the data itself. The cost of misclassifying a fraudulent record as non-fraudulent is much higher than misclassifying a non-fraudulent as fraudulent (Hall, 2001). The third problem, but not least important, is the detection of records that deviate from the normal distribution of the dataset. These records are known as *outliers* and can, in fact, be indicators of potential fraudulent activities.

There are two different approaches to dealing with the unbalanced dataset problem. The first one, normally done in the pre-processing phase, consists in the reconfiguration of dataset into a more balanced one. This can be done by eliminating some of the instances in the majority class or by generating more instances of the minority class. The second approach in dealing with the unbalanced dataset is the use of boosting algorithms. Boosting algorithms are iterative algorithms that use different weights on the training set distribution on each iteration. Adjusting the weights or by increasing the instance of the misclassified and decreasing the occurrence of the correct classified instances, the algorithm will focus more on the misclassified instances, which are more often fraudulent instances.

Because the error cost of misclassifying a fraudulent instances is higher than the error cost of misclassifying a non fraudulent occurrences, it is important to study not only the precision (percentage of correct classified instances) but, also, to the sensibility (percentage of correct classified fraudulent instances) of each case.

Besides these more technical problems, there is a conceptual issue that is not taken into consideration when building classifiers using a dataset that is formed exclusively by data taken from tax forms. This issue is supported by the fact that an organization is not unique and stands alone in the world. Organizations are built on relationships between other organizations and people themselves. Keeping this in mind, it is important to look at the social network of each organization. Due to this dynamic condition, the detection of

social network patterns, in the fraudulent organizations, can help the construction of a better data classifier. For this reason, tools are needed to evaluate if each organization has a pattern of social relationships similar to the fraudulent ones and, as a result, uses the information to assist in the classification phase.

In conclusion, the use of techniques to balance the dataset and the use of social network patterns to enrich the dataset are two of the techniques that, when combined, produce better results in the identification of fraudulent organizations.

2.1 Applications Domains on Fraud Detection

Fraud detection presents itself as World of its own, with many specific problems for each area of the fraud equation. Consequently, it is not surprising that fraud detection is a complex task. Different systems may be needed for different kinds of fraud and each of these systems has different procedures and parameters to tune databases interfaces (Cahill, 2002).

The problem of fraud detection is primarily caused by extremely large databases with a very unbalanced class distribution and a non-uniform cost per error. There are several areas where this problem occurs, such as in credit cards, insurance, telecommunications industries and financial fraud detection. Due to the numbers of fraudulent transactions being small, when compared with the legitimate ones and the inherent cost of misclassification being high, the detection of fraud is an extremely complicate task.

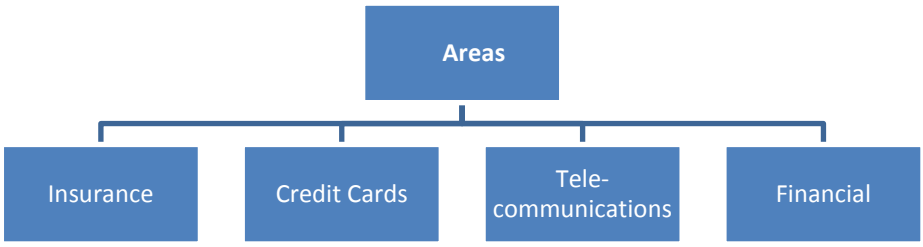


Figure 1 - Areas of Application of fraud detection techniques

Insurance

Insurance fraud is one of most frequent types of fraud to undertake. This type of fraud can take place in many forms with the simple objective of gaining money. The methods used to gain this money are known to be as: multiple insurance policies, suppression of information, faking of loss and the filing for deliberate losses. Another situation to

consider is multiple insurance policies, in which an item is insured more than once, in order to claim a higher premium than the cost of the item itself. The suppression of information is often used with health insurance and it occurs when a person omits a particular health problem in order to pay for a lower premium. Another examples is the faking a loss and occurs when an item is reported to the insurance carrier as being stolen but, in reality, the owner is being deceitful. Lastly, insurance fraud takes place as deliberate loss is filed when, for example, the owner of a factory sets fire to the installation because the amount of money to received from insurance is larger than the worth the factory itself.

All these types of frauds are criminal actions and lead to huge financial losses for the insurance industry. In the past, businesses were not interested in fighting this type of fraud because the fraud fighting task was not an easy one. For a business, it was preferable to increase premiums to cover the losses. Nowadays, the situation is quite different. Insurance companies are applying cross validation of various databases to detected fraud (Radcliffe, 1995). An example of this is with health insurance, where a person visited the dentist twice to remove the same tooth. Because the visits took place during a two-year period, the insurance carrier did not question the outcome. However, had it compared the database records, it would have been extremely easy to observe that a fraudulent act had taken place.

Credit Card

Credit card fraud is significantly different from insurance fraud. Credit cards fraud always begins with the act of acquiring the credit card number and using it to purchase product items that can be resold to gain cash or withdraw cash. In internet phishing, credit card skimmer and recording and system intrusion are the main techniques to gather the credit card numbers (Wang, 2006) (Shen, 2007).

With a high number of transactions taking place each day, a reduction of 2.5% in fraudulent acts will provide savings of one million dollars per year (Brause, 1999). ,As seen, credit card fraud detection has two peculiar characteristics: a high number of transaction to be handled and a short time to decide what to do with the information. Nowadays, credit cards are widely used to purchase items or to pay for almost anything. Consequently, millions of people perform millions of transactions every minute of the day. All these transactions must be analyzed in order to determine if they are fraudulent or not. The problem is not only the huge amount of data to be analyzed but, also, the short time in which this must be done. After a credit card transaction is initiated, the system must be

very rapid in sending an acceptance or rejection notice. Size and speed are the two main differences between credit card and insurance fraud. (Dorronsoro, 1997).

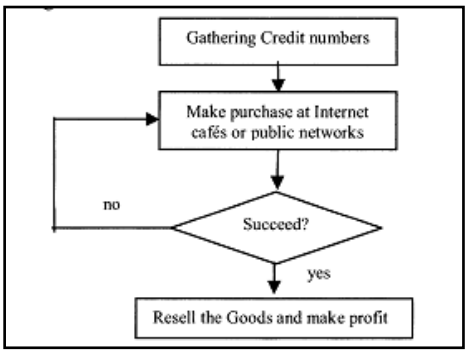


Figure 2 - The modus operandi of credit card frauds (Wang, 2006).

In view of that, there are two types of credit card transactions: the ones that are not authorized, which refer to transactions that deal with accounts to be known have been misused and the ones that are authorized. It is this former type of transactions that is a potential to those concerned because some of them deal with fake or stolen credit cards. As a result, the task here is to avoid a fraud transaction by use of a credit card before it is labeled as illegal (Brause, 1999).

Telecommunications

In the telecommunications industry, fraud refers to the illegal access to a network and the use of its services. This situation constitutes a problem to the telecommunications industry. It is here that this type of fraud causes losses as much as millions of dollars per year (Wang, 2004). Besides the loss of revenue, this type of fraud brings another concern: the waste of the network capacity. In this instance, there is an increase in the volume of data travelling in the network system, without the company receiving any payment in return for the unauthorized use.

Telecommunication fraud can be classified in two categories known as subscription and superimposed fraud. Subscription fraud occurs when someone subscribe to a service using a false identity. In this particular case, the perpetrator does not intent not to pay for service that is unlawfully used. On the other hand, superimposed fraud is the appropriation of a service without having the necessary permission this to do so. This situation can be easily detected by identifying unauthorized calls appearing in the telephone statement. This type of fraud happens in several ways, such as with mobile

phone cloning (much like credit cards cloning) and ghosting (technology that deceives the network in order to obtain free calls).

Hence in situations of large revenue losses, companies have to develop methods to detect the fraudulent activities. Data mining is an example of one of the method to detect fraud. With the assistance of a fraud detection models, fraudulent activities can be spotted, thus actions being beneficial to the telecommunication companies from an economic point of view (Taniguchi, 1998).

Financial

Financial fraud is one of the most complex and difficult types of fraud to detect. A business financial statement contains a balance sheet, income and cash flow for a certain period of time. These statements are based on the company's accounting standard. Fraud, in this area, consists in a somehow manipulation of these financial statements, in order to demonstrate the overstatement of revenue or the financial loss is understated (Kim, 2008). As a consequence, the detection of fraud in the financial industry serves to identify suspicious transfers, felonious statements or other illegal activities.

Outliers, which are known as records that deviate from the rest of the data, are seen as suspicious records and might be looked at very careful, when dealing with fraud detection. The identification of outliers can lead to the discovery of useful knowledge in the fraud detection areas (Ferdousi, 2006).

A specific type of fraud is known as Carrousel Fraud. In this case, it is very common in the European Union for this type o fraud to occur, Here, companies from two or more countries are linked together, in order to make transactions between themselves. The main objective of these companies is for neither of them to pay taxes to the respective State and, therefore, present a larger profit at the end of the business cycle. It is with this scheme that Nation States are harmed most.

3 Comparative study of balancing techniques

3.1 Literature Review

Indeed the problems described are just guides to the different mining approaches that presently exist to deal with the problem of fraud detection.

In particular, the problems of skewed datasets with minority classes can be handled by applying two different approaches: the algorithm or data approaches (Kumar, 2007).

The data approach is used at the pre-processing phase of the KDD process. This task consists of re-sampling the unbalanced classes. Re-sampling data is a technique that allows one to create more record examples of a class (over–sampling) or eliminate some records of a class (under-sampling). There are three forms for achieving a more balanced dataset. That is under-sampling the majority class, over-sampling the minority class, or doing both. The great advantage of this method is that it can be used with any classification technique (Ahumada, 2008).

The algorithm approach is used in the processing phase of the KDD and consists of modifying existing data mining algorithms by adjusting the costs of the errors (Han, 2005). Thus, by assigning a superior cost to the false negatives (fraudulent records classified as non-fraudulent) than to the false positives (non-fraudulent records classified as fraudulent ones), one will get better performance with respect to the fraudulent class (Weiss, 2004). Another possible approach, used in this phase, is to combine and apply different algorithms to perform better results, by taking advantage of the strength of each one (Phua, 2004).

Those two different phase's approaches end with a usual classification model, where there is a need for them to be evaluated and the performance of the classifier to be measured. This cannot be done in terms of predictive accuracy where all errors costs the same because, as we have seen before, because the minority class is the most significant one (Yue, 2007) (Yang, 2006).

Figure 3 shows twenty instances of a dataset, in which fifteen normal and five fraudulent transactions, where it is difficult to separate the fraudulent from the non-fraudulent ones. As a result, one of the ways would be to apply a basic classification technique to this dataset. Here, it would be difficult to build a good model. In a worse case situation, one would compile a model with a 75% of precision, in that it would classify all instances as normal. This is the main problem when dealing with unbalanced

data because it returns a classification model with a high precision rate. One of the concerns here, would be that this conditions cannot really identify any fraudulent case.

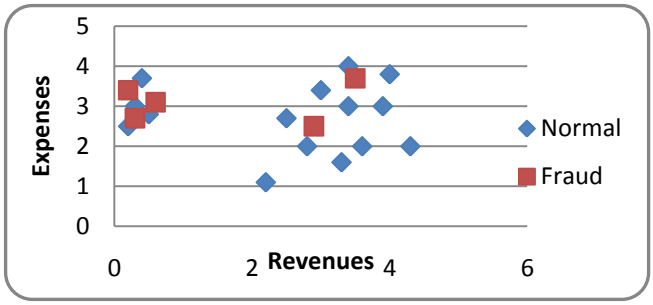


Figure 3 - An example of the distribution of a dataset with 20 instances with 2 attributes (Expenses and Revenues). There are 5 fraudulent instances and 20 non-fraudulent.

The last important method that can be applied to detect fraud is outliers detection. "*An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*" (Hawkins, 2002). This means that an outlier is a data record that has attribute values that a very different from the majority of the other records. This situation leads to the determination that the values were not naturally originated. As a result, the approaches to detect outliers are concentrated on the discovery of patterns that occur in a infrequently way in the data, as opposed to the most traditional techniques of data mining that have the goal to find patterns that occur frequently in the data (Koufakou, 2007).

Outliers are very often treated as errors that need to be removed during the data cleaning phase of KDD, in order for the specified algorithm to succeed. As an example of this would be for one to think of a sensor that reads hourly air temperature in a building. Every day, during a 24-hour period, the temperature oscillates between 18ºC and 22ºC, but at a particular hour, the temperature is recorded to be at 40ºC. This value would have to be considered as an outlier because it deviates, significantly, from the other records. Consequently, to have efficient model that are generated from a data source, it is necessary to remove all outlier record.

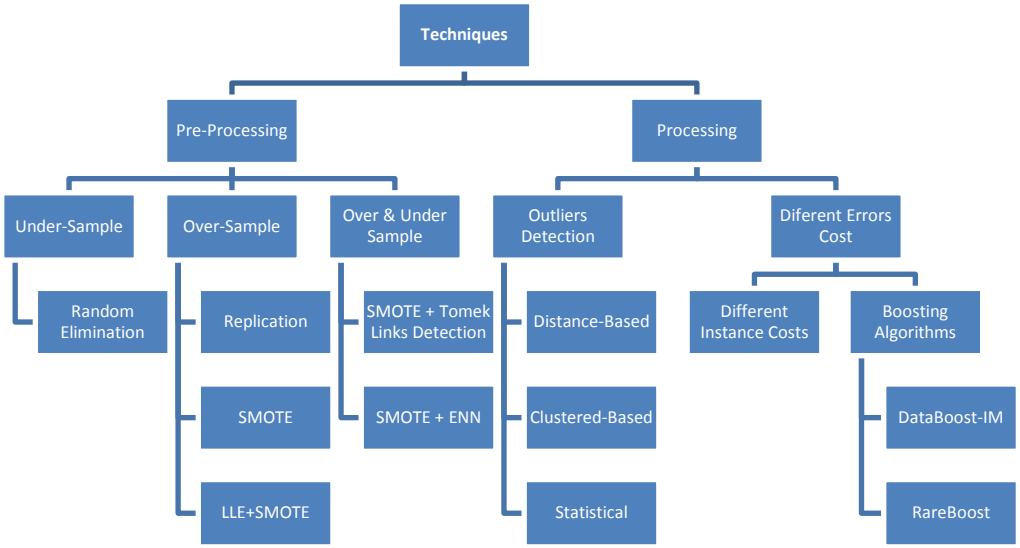


Figure 4 - Data mining techniques to detect fraud

Notwithstanding, outliers can sometimes lead to the discovery of important information. In the specific case of fraud detection, outliers can be notable errors or the can, in fact, be an indication of a fraudulent activity. Indeed, using the outlier techniques, one can, also, detect fraud, which is another way of fighting it.

Figure 4 presents the various methods of fraud detection that is described in foregoing document.

Balancing the dataset

As seen before, in a real setting, the natural distribution of data is not the most adequate solution in the application involving classifiers. In this case, one of the possible solutions is to change the data, in order to create a more balanced dataset. To accomplish this task, there are three strategies that may be applied to balance data. These strategies are over-sampling, under-sampling or both. Re-sampling the dataset is a method that creates a more balanced dataset by eliminating some of the records of the majority class (*under-sampling*) or by generating more records of the minority class (*over-sampling*) (Han, 2005).

Elimination

In the example of figure 5, a basic under-sampling technique is applied to the original dataset (figure 3) by random elimination of some instances from the non-fraudulent

records class. Although that condition can bring better results, it also brings many more problems with it. By eliminating instances of the majority class, the problem of over-fitting can occur. An *overfitted* model is a model too specialized that can have a higher precision rate in a specific dataset but, when it is used with a new set of data, will not produce any useful results. That is a common problem occurring in predictive models.

Removing some of the instances of the majority class, one can create important information loss, as seen in the example, the predictive model that classifies all data with values of revenues smaller than 3 and values of expenses bigger than 2 as fraudulent data has a predictive rate of 87% but the same model applied to the original dataset has only a predictive rate of 70%. As seen here, it is understandable that when under-sampling occurs, the majority class, by randomly removing samples, can experience a loss of some of the useful information (Chawla, 2003).

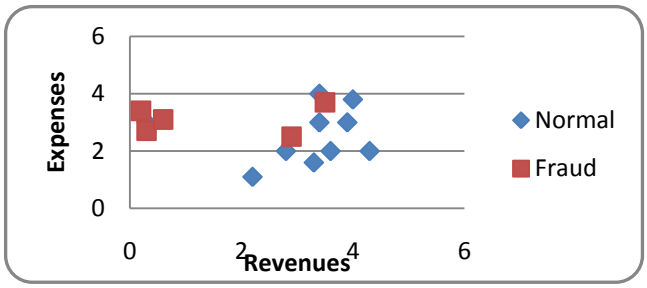


Figure 5 - Demonstration of the basic technique of under-sampling the majority class by random elimination of some of its own instances.

Replication

In the example of figure 6, a basic over-sampling technique is applied to the original dataset (figure 3) by duplicating the samples of the fraud class. This condition returns a more balanced dataset that can create a better predictive model, after applying a classification algorithm. This algorithm can learn from more fraudulent examples.

This is a basic and simple solution to balance the dataset but as also under-sampling by elimination, it can leads to over-fitting. Another issue that can occur in this situation is when applying this technique in order to increase the size of the dataset, which can lead to an enormous problem, when dealing with already large and unbalanced datasets (Chawla, 2004).

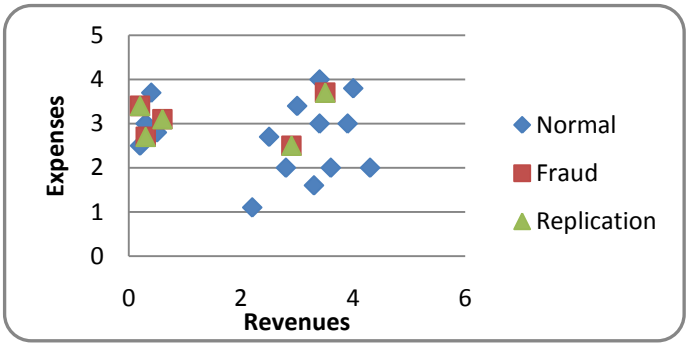


Figure 6 - Demonstrations of the basic technique of over-sampling a minority class by replication its own instances.

SMOTE

Although re-sampling the data is a good idea to originate a more balanced dataset, in order to help the classifiers build a better predictive model, this is a simple task because it can create other problems that will disrupt the quality of this model. For this reason, the basic techniques for modeling are not efficient and, therefore, other techniques have to be addressed.

One of these main techniques, which serve to deal with over sampling, is known as the Synthetic minority over-sampling technique (SMOTE). SMOTE over-samples the minority class by introducing into the equation synthetic examples, along the line of segments that join the k, minority class and nearest neighbors (k-NN). Depending upon the amount of over-sampling required, neighbors from the k –NN are arbitrarily chosen (Chawla, 2002). Using SMOTE, the inductive learners are able to expand the minority class decision regions without leading to the over fitting problem (He, 2005) (Pelayo, 2007). Synthetic samples are generated by taking the difference between the feature vector (sample) under consideration and its nearest neighbor, multiplying this difference by a random number (between 0 and 1) and adding it to the feature used feature vector. This will cause the selection of a random point, along the line segment between two specific features. This approach causes, in an effective way, the decision region of the minority class to become more general. Applying this sampling method, because it does not use the random duplication of samples, does not create a more specific minority class, Rather, it causes the classifier to build a larger decision regions, which contains nearby minority class points (Chawla, 2002).

Instance	Nearest Neighbor	Distance Vector	New Synthetic Instance	Random Factor
(0,3;2,7)	(0,6;3,1)	(0,3;0,4)	(0,3;2,7)+rand(0-1)*(0,3;0,4)=(0,45;2,9)	0,5
(0,6;3,1)	(0,2;3,4)	(-0,4;0,3)	(0,6;3,1)+rand(0-1)*(-0,4;0,3)=(0,48;3,01)	0,3
(0,2;3,4)	(0,6;3,1)	(0,4;-0,3)	(0,2;3,4)+rand(0-1)*(0,4;-0,3)=(0,48;3,19)	0,7
(2,9;2,5)	(3,5;3,7)	(0,6;1,2)	(2,9;2,5)+rand(0-1)*(0,6;1,2)=(3,14;2,98)	0,4
(3,5;3,7)	(2,9;2,5)	(-0,6;-1,2)	(3,5;3,7)+rand(0-1)*(-0,6;-1,2)=(3,02;2,74)	0,8

Figure 7 - Calculation of the new SMOTE instances

Figure 7 represents the generation of the new instances by applying SMOTE to the dataset example of figure 1. Each instance of the minority class is selected, and then it is identified the fraudulent nearest neighbor and calculated the distance vector between them. After this is done, the formula to generate the new instance is applied:

$$(f1', f2') = (f1_1, f1_2) + rand(0-1) * (f1_1 - f2_1, f1_2 - f2_2)$$

This creates new instances in the line that links the original instance to its nearest neighbor, as seen in figure 7 where it is represented the result of this application of SMOTE to the original dataset (figure 3).

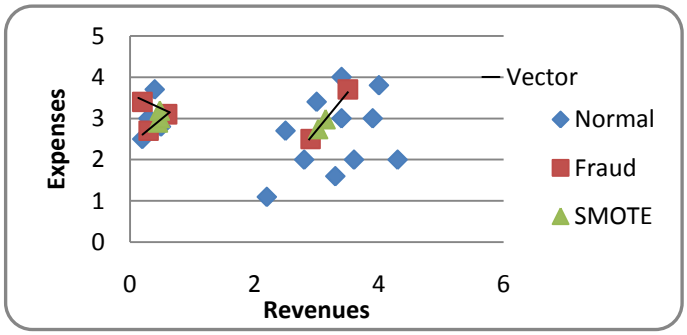


Figure 8 - Demonstrations of the SMOTE technique of over-sampling the minority class.

BorderlineSMOTE

SMOTE, as a bases for the need to conduct some modifications to the model, was added to increase the accuracy of the classifier of the minority class.

Borderline SMOTE is one of those new techniques in which SMOTE is only applied to the borderline of the minority class, by just creating new samples on this border. Figure 8 shows the application of Borderline SMOTE. Figure. 8 (a) is the dataset representation, where the black dots are the instances of the majority class and the red ones are the

minority class. Figure 8 (b) represents the identification of the instances in which the minority class is the borderline within the ones of the majority class contained in the blue squares. Figure 8 (c) is the final dataset after applying SMOTE, applied only to the instances of the blue square. The proposed solution returns better results than applying SMOTE to all instances, because the ones that are in the borderline of the two classes are the ones that have higher probability of being misclassified (Han, 2005). Another advantage of this technique, when compared with SMOTE, is that by only creating instances that will really help build a better classifier in which the dataset size does not increase with un-useful instances. This condition contributes to a more efficient construction of the model.

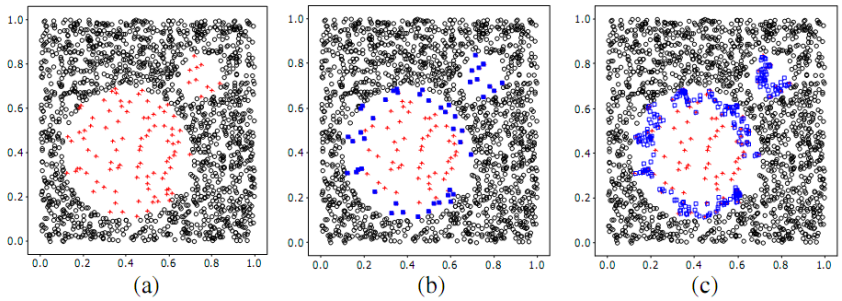


Figure 9 - Demonstrations of the application of Borderline SMOTE (Han, 2005)

SMOTE+LLE

Another possible use of the SMOTE is the combination of it with the Locally Linear Embedding (LLE) technique. The reason for doing this is the fact that SMOTE cannot be applied when the original data is not linearly separable. Therefore, when this happens, the LLE technique can be applied to mapping the original data into a new linearly separable feature space, followed by the application of SMOTE. This technique avoids the creation of positive instances on top of negative ones resulting in a better classifier (Wang, 2006).

SMOTE+TomekLinks

Another good solution is to detect Tomek links to under-sample the majority class after the application of SMOTE. Tomek links are pairs of samples of different classes that are either noise or on the borderline. Indeed, a pair of samples that form a Tomek link does not have any other sample of either the classes between them. A problem frequently observed is that some datasets do not have well defined classes, where some of the elements of the majority class crossover into the domain of minority class. When this

condition occurs, combining Tomek with SMOTE is very good. In view o that, the advantage is to calculate Tomek links after applying SMOTE to over-sample a class.

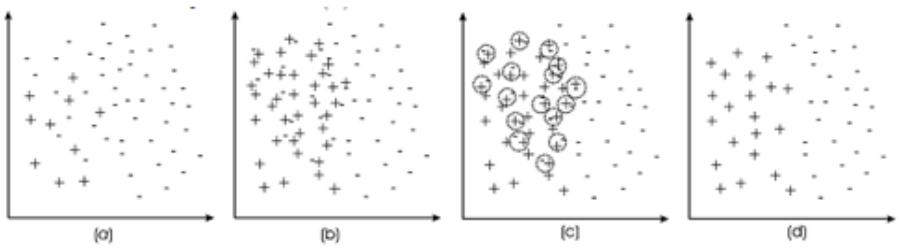


Figure 10 - Demonstrations of the application of SMOTE combined with Tomek links detection (Batista, 2004).

In figure 10 (a) it is represented the original unbalance dataset. The fig. 10 (b) is the result of SMOTE. Fig. 10(c) is the detection of the Tomek links and in fig 10. (d) is represented the final dataset result after the removal of the elements from the majority class of those links. The results are two well separated balanced classes (Batista, 2004).

SMOTE+ENN

The application of SMOTE with Wilson's Edited Nearest Neighbor Rule (ENN) is compared to SMOTE with Tomek links. But, instead of only removing the item from the majority class it also removes it from the minority one. ENN detects a sample that has been classified in a different form, compared to the majority of the three nearest neighbors. In figure 11 there are in "1" four instances of the normal class that are classified different from 2 of their 3 nearest neighbors and in "2" is a fraudulent instance that is classified different from his 3 nearest neighbors so this 5 instances will be eliminated. This leads to a more in depth data cleaning, as seen in figure 12 (Batista, 2004).

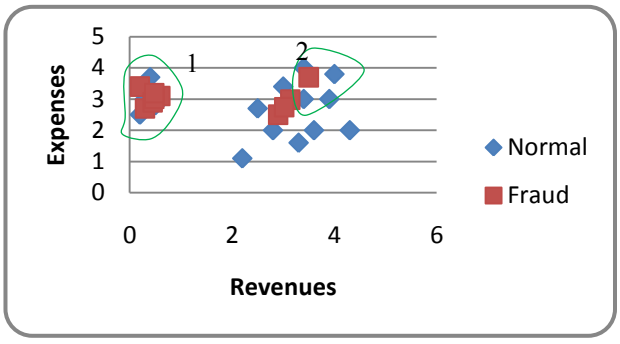


Figure 11 - Demonstration of ENN technique to select the instances that will be eliminated.

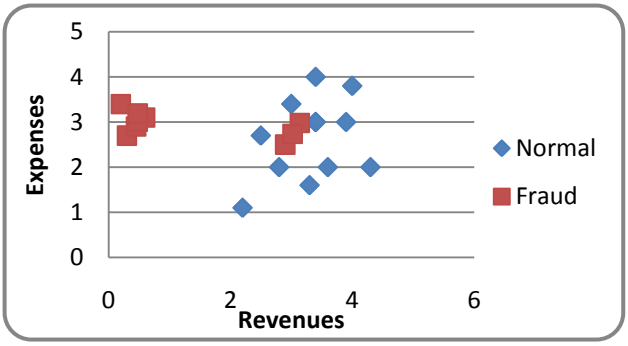


Figure 12 - Representation of the final dataset after the application of SMOTE with ENN.

Considering different errors costs

At an algorithm level, to deal with unbalanced datasets, the methods applied work in algorithms and not on the distribution of data, as seen before, and there are two different solutions to consider. One is to use cost-sensitive learning methods and the other is to apply boosting algorithms.

Cost-sensitive learning methods explore the fact that it is more important to correctly identify the elements from the minority class. Therefore, for fraud detection problems, the solution is to associate a greater cost to the false negative (fraudulent cases classified as non-fraudulent) than to the false positives (non-fraudulent cases classified as fraudulent) (Weiss, 2004).

As an example, let's look to figure 13 and define that the misclassification cost ratio to be 2:1, this meaning that a fraudulent instance has a value 2 and a non-fraudulent has a value 1. As consequence of that cost ratio, region 1 will be classified as fraud because it has 4 normal instance with ratio 1 and 3 fraudulent ones with ratio 2 what means , to the learning algorithm, that it has 6 fraudulent instances. Furthermore, region 2 will be classified as non-fraudulent because it only has 2 fraudulent instances that as value 4. To be classified as fraudulent the misclassification cost ratio as to be 6:1 (2*6=12 fraudulent instances versus 11*1=11 non-fraudulent instances)

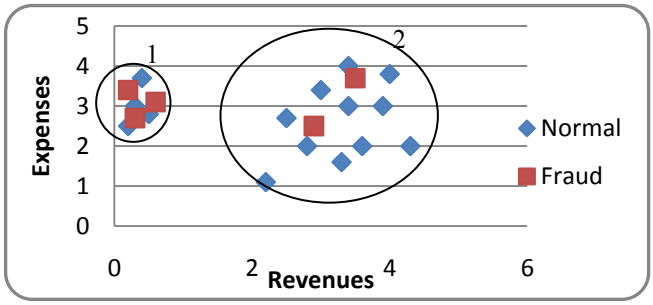


Figure 13 - An example of a dataset with two specific regions.

Although as reviewed in the literature, these cost-sensitive learning methods are referred to techniques to detect fraud at an algorithm level, it seems that this method is equal to the replication of the minority class at the pre-processing phase and, most likely, leads to the over-fitting problem that was explained earlier.

The other solution to deal with different errors costs is to apply boosting algorithms. These algorithms are iterative algorithms that use different weights on the training set distribution of each iteration. This means that after each iteration the algorithms increase the weight of the misclassified examples and decrease the weight of the correctly classified ones. The continuous correction of weights forces the learning process to focus more on the incorrectly classified examples in the next iteration, instead of treating all examples as equals. Thinking that the classifier, more often, misclassifies the minority class, it appears that these boosting algorithms may improve the accuracy of the fraudulent examples. (Weiss, 2004)

An example of a boosting algorithm is the DataBoost-IM which combines data generation and boosting procedures to improve the predictive accuracy on both classes without improving the minority one in detriment of the majority class. This is method produces a more reliable classifier because it learns from the greater number instances to classify (Guo, 2004).

The main disadvantage of the DataBoost-IM and all the other standard boosting algorithms is the fact that they increase the weights of misclassified examples and decreases those instances that are correctly classified using the same proportion, whilst not considering the imbalance of the dataset. Due to this condition, the traditional boosting algorithms do not perform well on the minority class. (Han, 2005).

This limitation can be seen as motivation for a possible solution to the higher cost of misclassifying the minority class is the use of RareBoost algorithm. As the name indicates, this was a boosting algorithm developed to create better classifiers of the rare

class examples. This algorithm updates weights of positive prediction (True Positives and False Positives) in a different way that updates the negative ones (True Negative and False Negative). Because of those different updates and given a higher weight to the positive prediction, the classifier produced will have greater accuracy on the positive (minority) class, when compared to the traditional boosting approach. (Joshi, 2001).

In conclusion, the solution of considering different errors costs is goes away by adjusting the weights of the misclassified examples, focusing on the construction of a better classifier, which can detect with more accuracy records as fraudulent or non-fraudulent.

Outliers

As seen before, there is also a third way of detecting fraud. This is the application of techniques for outliers detection. Outliers are records that deviate from the rest of the dataset and, therefore, providing from the data a clear indication or suspicions of fraud. In the literature, there are no specific methods for Outliers detection to be applied in fraud detection problems. For this reason, all the generic methods can be applied towards the goal of finding outliers. They can be divided in three approaches: Statistical-Based, Distance-Based and Density-Based.

The *Statistical-based* approach assumes that a parametric model describes the distribution of the data and that it has, normally, only one attribute. The problem and limitation of this method is the fact that it is difficult to find the right model for each dataset and the application of it because it efficiency declines as the data dimensionality increases.

The *Distance-based* approach does not make assumptions for the data distribution because it only computes distances between points. The biggest problem of this method is the fact that it is impossible to use for large datasets because it will cause large calculations of data.

The *Density-based* approach calculates the density distribution of the data and identifies outliers that are in low-density regions. This approach has the advantage that it can detect outliers that are not detected with the other techniques but it fails when working in a high-dimensional space because data is sparse (Koufakou, 2007).

There is a technique that can be applied is the Peer Group Analysis (PGA), which offers several advantages to the process of dealing with fraud detection problems. It is a technique used when a time series data is more difficult to detect fraud based on outliers' detection. For example, in the case of a person having a salary of 1000€, the bank credits

this person account for 1000€ each end of the month, but at Christmas this person receives a Christmas bonus of 1000€. In this case the bank account shows that during the month of December 2000€ was received. The December a amount deviates from all other 11 months of the year. In this case, fraud is flags, simply, based on a basic outlier detection algorithm. Although these transaction it might seem as suspicious acts, when compared with all other persons within the same employment and a wage schedule, it is easy to determine that they all received the same amount of money, Obviously, here, fraud is not the case. This basic rational of the PGA is that some outliers' values in a time series data can be classified as non-fraudulent, when the agents in the same peer group are affected by the same actions.

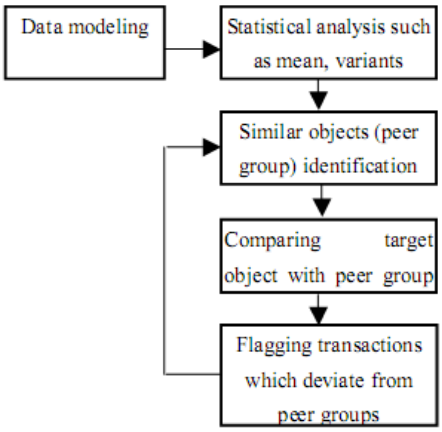


Figure 14 -The process flow of outliers' detection in a PGA (Ferdousi, 2006).

In this degree, the PGA consists of the creations of data groups compiled of people with the same characteristics and, then, analyzed in search of outliers that deviate from the peer group.

Before proceeding with a description of the different application areas of fraud detection, some particularities of the KDD process will be discussed, when applied to this domain.

3.2 Comparisons Studies

To make the studies was used two different datasets: one with artificial data called Ozone and the other with real data of organization's VAT declarations. To do the experiences was used a Weka 3.6.0 which is a data mining open source software. This software has in it 3 different techniques to balance the dataset in the pre-processing

phase. Those techniques are Resample, random subsample of the dataset using both elimination from the majority class and replication of the minority, Spread SubSample, that produces a subsample of the majority class, and SMOTE witch resample the minority class by applying the Synthetic Minority Oversampling Technique described in section 3.1. All these techniques are applied before the creation of the classifiers and produces different datasets distribution.

Besides that techniques was used also two different Boosting Algortihms, in the processing phase. Those algorithms are iterative and focus on the incorrect classified instance on each iteration and because of this focus they produce better classifiers when they learn from their errors. In Weka there are two different boosting algorithms Adaboost, that is a basis boosting algorithm, and MultiBoost that combines also wagging. In both case was used the j48 as the base learning algorithm.

Type	Literature Review	Weka 3.6.0
Balancing Techniques	Random Elimination	Spread SubSample
	Random Elimination + Replication	ReSample
	SMOTE	SMOTE
Different Errors Costs	Boosting Algorithms	AdaBoost
	Boosting Algorithms	MultiBoost

Figure 15 –Correspondence between the techniques described before with the ones implemented in Weka.

The table before represents de correspondence between the techniques described in section 3.1 in Literature Review and the techniques used with Weka. There was not used any different technique that was not explain in the Literature Review but, because not all the techniques described were implemented in Weka was impossible to test all methods. Although that the main techniques and methods to deal with the unbalanced dataset problem were tested.

To do all those test with those different techniques and different datasets was used the standard KDD process generic framework, called CRISP-DM, which is divide is three phases.

The first one is the *pre-processing* phase, where a table is created, containing all available data having one line per entity. After data collection, it is necessary to analyze the context where this data is inserted, to analyze de data to detect problems like miss and invalid values and to correct those data problems.

The second phase is called modeling and it is where the information is discovered. In this phase it is necessary to pick up a method like decision trees or neural networks to create the model.

To solve the fraud detection equation, the pre-preprocessing phase in acted upon with techniques to balance the dataset, along with the processing phase, with the detection of outliers and the creation of classification models.

After the creation of those classification models, we enter on the third phase, the evaluation phase. This is a decisive phase where we have to decide if the model is good enough to be a definitive fraud classifier. To decide that it is necessary to evaluate the model performance.

In a normal balanced dataset problem the quality of the classifier is given by the *accuracy* which is the rate of the correct classified instances. However, when dealing with an unbalanced dataset, as in fraud detection, it is not adequate to evaluate a classifier by its accuracy. Consider a problem where 99% of the data are non-fraudulent and only 1% is fraudulent: a classifier that says that every record is non-fraudulent would have an accuracy of 99%, or just an error rate of only 1%. This is an excellent model but un-useful to the problem. Therefore, it is important to look carefully at other numbers (Batista, 2004). The most important values are the true positive rates (sensibility) that correspond to the percentage of fraudulent cases classified as fraud and, therefore, must be as high as possible and the false negatives rates, that correspond to the percentage of fraudulent cases classified as non-fraudulent, must be as lower as possible.

The simplest way of doing performing this task is based on the analysis of the *confusion matrix* (figure 16). This matrix holds cases of positive instances classified as positive (*True Positive*) and negative numbers (*False Negative*) and the number of the negative ones classified as negative (*True Negative*) and as positive (*False Positives*).

	Predicted Negative	Predicted Positive
Actual Positive	FN(False Negative)	TP(True Positive)
Actual Negative	TN(True Negative)	FP(False Positive)

Figure 16 - Confusion Matrix

3.2.1 Artificial Data

The Ozone dataset is a set with 2534 instances with 2374 normal instances and 160 Ozone ones. This is a very unbalanced data in which the minority class is about 6% of the dataset. This dataset has 72 attributes with real values in which there are at most 250 missing values.

This unbalanced dataset is the input of several experiences with the objective of testing several techniques to deal with this characteristic of dataset.

Unbalanced Dataset Problems

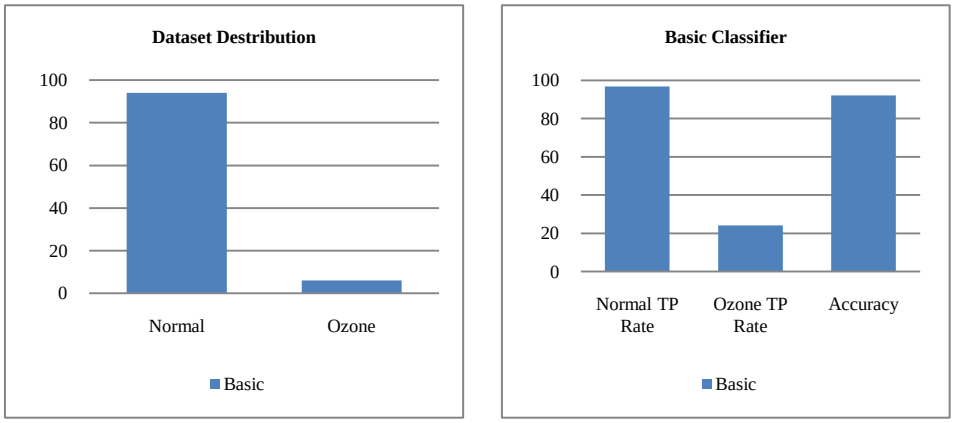


Figure 17 -Charts that shows the dataset distribution and the results of the basic classifier.

The basic classifier was built using j48 algorithm and produced the results presented in the chart above. The training set had a total of 861 instances which 54 of them were Ozone instances. The basic classifier classified correctly 795 instances (97%) but only 13 Ozone instances were correctly classified (24.1%).

Pre-Processing Techniques

The objective of the classifiers in this problem is to detect the Ozone instances but with the low Ozone TP Rate as seen with the basic classifier this objective cannot be achieved with a good score. Because of that were used the 3 different techniques to balanced the dataset that are implemented in Weka and were described before. All these techniques are applied before the creation of the classifiers and produces different datasets distribution. In figure 18 are classes distribution charts of both singulars and societies datasets before and after the appliance of those pre-processing technique.

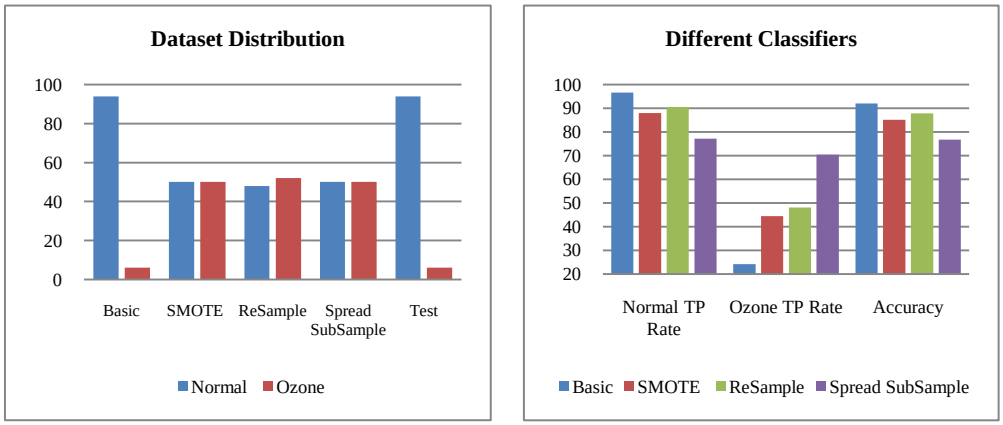


Figure 18 - Charts that shows the datasets distribution and the results of the classifiers with the different dataset balancing techniques.

All different techniques used to balance the dataset creates a balanced dataset that generates better classifiers. As shown in the graph above all classifiers created from the balanced datasets are better than the original one. Spread SubSample is the better technique when looking Ozone TP Rate but the accuracy is lower. ReSample looks the better technique because it not only increases a lot the Ozone TP Rate but also has a very high accuracy rate that is almost the same as the basic classifier.

	Correctly Classified Instances	Total Instances	Correctly Classified Ozone Instances	Total Ozone Instances
Basic	793	861	13	54
SMOTE	734	861	24	54
ReSample	756	861	26	54
Spread SubSample	660	861	38	54

Figure 19 -Table with the numbers of correctly classified instances with the different balanced datasets.

Processing Techniques (Boosting Algorithms)

The processing techniques are applied in the processing phase instead of pre-processing phase and, as described before, are iterative algorithms that use the learning algorithm j48 and in each iteration focus on the misclassified instances before. In figure 20 are the charts with the measures of each boosting algorithm in the two datasets used before.

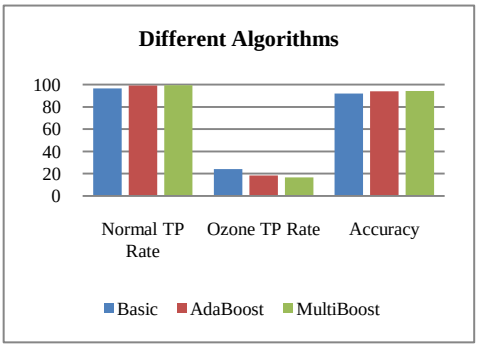


Figure 20 - Chart that shows the results of the classifiers with the different boosting algorithms.

In this specific dataset, both boosting algorithms produced worse results, compared to the basic classifier. Both classifiers presented more accuracy and better normal TP Rate yet works with Ozone TP Rates. This could be explained by the fact that these boosting algorithms are not specific to unbalanced datasets. They take into consideration the TP Rate of the minority class and the main goal is to reach highest accuracy possible.

	Correctly Classified Instances	Total Instances	Correctly Classified Ozone Instances	Total Ozone Instances
Basic	793	861	13	54
AdaBoost	810	861	10	54
MultiBoost	811	861	9	54

Figure 21 -Table with the numbers of correctly classified instances with the different boosting algorithms.

Conclusions

The Ozone dataset provided an unbalanced dataset to be tested with different techniques. With the techniques to change the dataset into a more balanced one all results were better than the original dataset and ReSample was the techniques that produced better results looking at Ozone TP Rate and Accuracy. On the other side boosting algorithms did not produced better results when classifying the minority class. Although that the classifier's accuracy with those algorithms was higher.

In conclusion it was shown that the techniques to create a balanced dataset produced better results but the use of boosting algorithms in this case created worst classifiers in terms of Ozone TP Rate.

3.2.2 Real Data on Fraud Detection

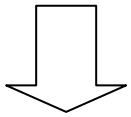
Data Understanding and Preparation

The first phase of work was to analyze the database in order to discover the important data to be used. The database is formed with data from organizations and people and its own description and taxes declarations of both from 2004 until 2006. There are also described the relationships between organizations, people and both. An example of those relationships is the person A that is the CEO of organization X and a partner of organization Y and he is also married with the person B. In this example is described the relation between organization X and Y, person A and B, and person A and organizations X and Y. In the database beyond the characteristics of organizations X and Y and persons A and B there are also the VAT and IRC declarations of X and Y and IRS tax declaration from A and B.

Because the financial fraud is a big world this work was focused on a small slice of it: the VAT declarations. As a result, the main objective was to detect and classify new VAT declarations as possibly fraudulent or non fraudulent. This classification will help to detect which organizations are much more prone to commit fraud. The VAT declaration indicates the amount of the transaction (purchase or sale) of a service or products offered by an organization. It is the method used for the calculation of the amount taxes that the organization must pay or to receive from the State. The objective when doing fraud at VAT declarations level is to pay less taxes or receive more refunds.

The second phase was to prepare the data to apply the data mining techniques. The organizations could be divided in two different types: societies and singulars and because each one has its own particularities was decided to create different tables to each type. Singulars are green receipts or organizations with only one person to opposite to societies that are the organizations with more than one owner and employees. Each organization fills a VAT declaration each trimester and the database has declarations from the first trimester of 2004 until the fourth of 2006. The data from those declarations was used to build the data sets. There were selected all VAT declarations of year 2004 and 2005 to create the train table and of year 2006 to create the test table. The train table was used build the classification model and the test table were used to evaluate this model. The attributes from those tables are the fields of VAT declarations. The explanation to the built of this table is explained on figure 22.

Org.	Year	Trimester	Total to Pay	Total to Receive
A	2004	3	500	0
A	2005	1	1000	400
A	2005	2	200	1440
A	2005	3	5400	3250
B	2005	2	990	100
B	2005	3	0	1500



Org.	T. Pay Trimester 1	T. Receive Trimester 1	T. Pay Trimester 2	T. Receive Trimester 2	T. Pay Trimester 3	T. Receive Trimester 3
A	-	-	-	-	500	0
A	1000	400	200	1440	5400	3250
B	-	-	990	100	0	1500

Figure 22 - A demonstration of the transformations in dataset.

In this phase of data preparation was used SQL Server to manipulate the data and create the tables. Next it was necessary to add the classification labels Fraud and Non Fraud to the tables. Those labels were not given but there was present a Risk Factor Grid to calculate the risk factor to each organization. For that reason, each organization has its own risk value from which that value is deduced the classification. Having as the objective an unbalanced dataset, the labels of singular organizations were calculated differently from the associations. As is in the singulars case, if the risk factor was zero, then label would be considered as “Non Fraud”. In the case of the factor being greater than one hundred, the label would be considered as “Fraud”. In the case of the factor being between 1 and 100, the factor was the percentage do be labeled as “Fraud”. In figure 2 is an example of the calculation of singulars’ labels. In the case of Societies if the factor was superior to two hundred it was labeled as “Fraud” and if it was between one and two hundred it the factor divided by two was the percentage of being fraud.

Org	Risk Factor	Formula	Label
A	0	= 0	Non Fraud
B	150	> 100	Fraud
C	75	Factor > Random(1…100)	Fraud

Figure 23 - A demonstration of how labels were calculated.

The result of the data preparation described before is two similar tables with the same attributes to the Singular and Societies. The only physical difference between the singulars and societies table was the presence of attribute C_CIRS_PROFISSAO in the singulars table that is applied only to the “Green Receipts”. In figure 3 is a table with all the attributes and is description. These are the attributes that are present in the VAT declaration. The attributes’ name that start with letter “C” are codes and with letter “V” are values.

Attribute	Description	Type	Values Range	Missing Values
X_NIF	ID	Nominal		0%
LABEL	Class Label	Nominal	[Non Fraud; Fraud]	0%
C_CAE	Area of industry	Integer	[0; 96091]	0%
C_CIRS_PROFISSAO	Job (only present in singulars table)	Integer	[0; 9011]	0%
C_LOCAL_ACTIVIDADE	Local of activity. 4 Attributes from this type.	Integer	[1; 3]	15.4%
C_DP_MODELO	VAT model. 4 Attributes from this type.	Integer	[1; 6]	15.4%
C_DP_CLASSIFICACAO	Classification. 4 Attributes from this type.	Integer	[0; 5]	15.4%
V_ATTRIBUTES	Values from de VAT declaration. 104 Attributes from this type.	Real	[0; 61074]	15.4%

Figure 24 - Description of dataset.

In the next tables are the correspondence between code attributes and its real value.

Code	LOCAL_ACTIVIDADE	DP_MODELO	DP_CLASSIFICACAO
0	-	-	Imposto a Entregar s/ Cheque Enviado
1	Mainland	Model A	Imposto a Entregar c/ Cheque Enviado
2	Azores	Model B	Reporte de Crédito
3	Madeira	Model C	Pedido Reembolso
4	-	Appends	Reporte + Reembolso

5	-	Model 383	Blank
6	-	Model 420	-

Figure 25 - Description of attributes LOCAL_ACTIVIDADE, DP_MODELO, DP CLASSIFICACAO.

Code	CAE
00000	Undefined
01111	Cerealicultura (excepto arroz)
01112	Cultura de leguminosas secas e sementes oleaginosas
01120	Cultura de arroz
01130	Culturas de produtos hortícolas, raízes e tubérculos
01140	Cultura de cana-de-açúcar
01150	Cultura de tabaco

Figure 26 - Description of attribute CAE.

Code	CIRS_PROFISSAO
0000	Undefined
0201	Architects
0202	Civil Constructors
0203	Engineers
0204	Technical Engineers
0205	Topographers
0206	Designers

Figure 27 - Description of attribute CIRS_PROFISSAO.

The options and decisions that were made conduct to some limitations of this solution. First of all the labeling of the VAT declarations was made by using a Risk Factor Grid that were created to evaluate the Organizations and not the VAT declarations. Another issue that had to be dealt with was the fact that some of the organizations did existed only during part of the year. Consequently, the table was missing values. As an example of this is an organization that was established during the 2nd quarter of 2004 and ended the business during the 3rd quarter of 2005. This organization I question did not fill the VAT declaration for the 1st quarter of 2004 and the 4th quarter of 2005. These informational gaps did, in fact, interference with the classifier quality. The solution was to ignore all organization that didn't file all the quarterly VAT declarations. For that reason, the

situation created a concern: a very small amount of data to be worked. Another possible solution for this problem was to use data from trimesters and not years. Instead of having a table with 4 trimesters (and 4 declarations) the solution was having a table of 1 trimester. The problem encountered here was the fact that a company that, for example, produces Christmas decoration products will have much higher values of transactions in the last trimester that in the first one. This fact could lead also to a bad classifier.

To conclude, the data preparation phase, consisted in the selection, creation and cleaning of the data towards the creation of data tables to create and test models of classification to detect fraud in the VAT declarations. The next chapters explain all the work done with this data, the methodology and solutions. In a future work the data preparation could be much more correct with the presence of domain experts that could create a Risk Factor Grid to be applied in the classification of VAT declarations.

Classification

In this chapter is described the methodology and solutions to deal and solve the unbalanced dataset problems. Using the unbalanced dataset described before were applied different techniques to change the learning dataset to a more balanced one to build a better classifier. This study describes the different techniques of pre-processing and makes a comparison of them. It was also used to different boosting algorithms to test if they produce better classifiers too.

Unbalanced Dataset Problems

As described in those previous chapter was used two different sets, singulars and societies, both very unbalanced. In figure 28 is a chart of the two datasets distribution.

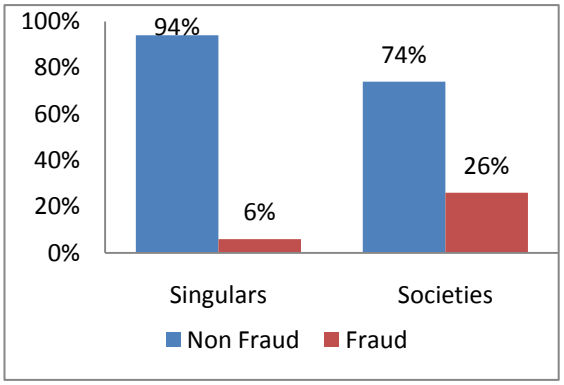


Figure 28 - Charts that shows the class distribution of both datasets: Singulars and Societies.

Those datasets were the study starting point. The first step was the building of a classifier with those datasets distribution to have a comparison point with the next classifiers. All classifiers were constructed using the Weka 3.6.0 software with the classification tree algorithm j48 which is the Weka C4.5 clone. To evaluate those classifiers was measured the Non Fraud True Positive Rate (Non Fraud TP Rate), the Fraud True Positive Rate (Fraud TP Rate) and the Accuracy. Those measures are calculated with formulas shown above.

$$Non\ Fraud\ TP\ Rate = \frac{Non\ Fraud\ Instances\ Correctly\ Classified}{Number\ of\ Non\ Fraud\ Instances}$$

$$Fraud\ TP\ Rate = \frac{Fraud\ Instances\ Correctly\ Classified}{Number\ of\ Fraud\ Instances}$$

$$Accuracy = \frac{Instances\ Correctly\ Classified}{Number\ of\ \ Instances}$$

The problem when dealing with unbalanced datasets is that it classifies very bad the minority class instances. Because of that the classifiers of an unbalanced dataset fraud detection problem have, very often, high Non Fraud TP Rates and Accuracy and a very low Fraud TP Rate. In figure 29 is the evaluation of the classifiers built with the figure 1 datasets.

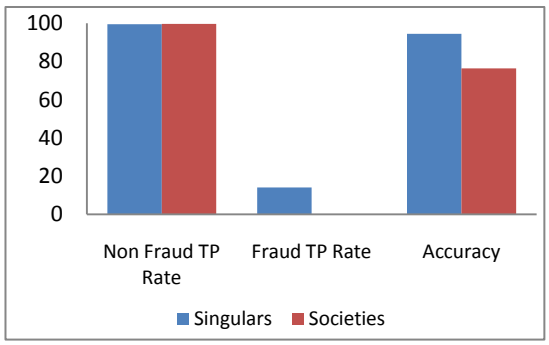


Figure 29 - Charts that shows the basic classifier measures.

As expected, the Non Fraud TP Rate was almost 100% in both cases but the Fraud TP rate as about 10% the singulars case and 0% in the Societies case. Although that low

Fraud TP rate the accuracy in both cases was extremely good, that because the number of instances Non Fraudulent are much more than the Fraudulent. So the total number of correctly classified instances was very high. The next table shows the number of instances that support that high value of accuracy and low value o Fraud TP Rate.

	Correctly Classified Instances	Total Instances	Correctly Classified Fraud Instances	Total Fraud Instances
Singulars	1041	1202	9	64
Societies	296	388	0	91

Figure 30 - Table with the numbers of correctly classified instances in both datasets with a basic classifier.

Pre-Processing Techniques

The objective of the classifiers in this problem is to detect the fraud instances but with the low Fraud TP Rate this objective cannot be achieved with a good score. For this reason, three different techniques were used to balance the dataset implemented in Weka and were described and used before. All these techniques are applied before the creation of the classifiers and produces different datasets distribution. Figure 31 presents classes distribution charts for both individual and associative datasets before and after the application of the pre-processing techniques.

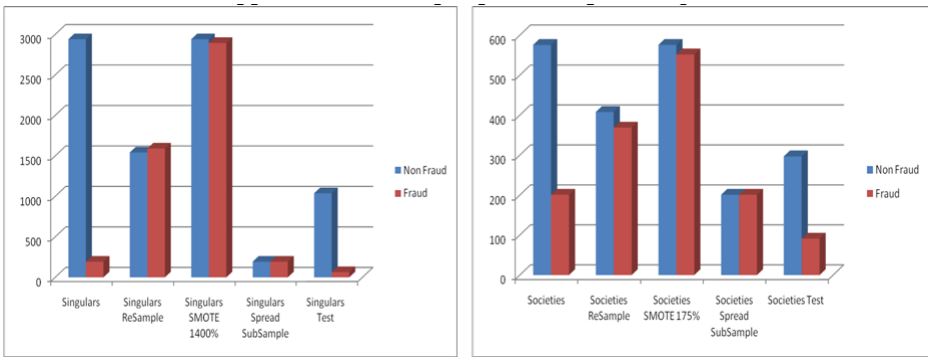


Figure 31 - Classes distribution charts of both datasets with the appliance of balancing datasets techniques.

The application of these different techniques produces all balanced datasets. SMOTE produces instances to make the fraud class have the same number of instances of the non fraud class and Spread Subsample does the opposite, eliminates instances from the non fraud class until it have the same number of instances that the fraud class. Resample

produces a dataset balanced and where the fraud class has more instances than the original dataset and the non fraud have fewer instances than the original one.

After having those balanced that sets were created different classifiers for each using always the j48 algorithm in Weka. Figure 32 has the charts that represent the measure of classifiers quality for each dataset: original dataset, SMOTE (percentage = 1400%), Resample and Spread Subsample.

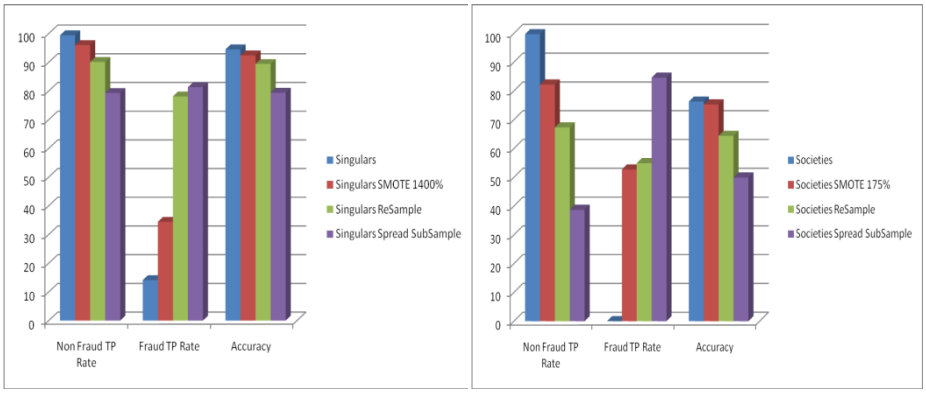


Figure 32 - Classifiers evaluation charts of both datasets with the appliance of balancing datasets techniques.

The results shows us that the right classification of Fraud instances (Fraud TP Rate) has increased in a very expansive form and by opposite the Non Fraud TP Rate has decreased but in a lower rate than the Fraud one. This fact happens because the learning algorithm now looks to the two classes in the same way and both have the same importance. Because the goal of the learning algorithm is to achieved the higher Accuracy possible in the original dataset it was not a problem to classify all instances as Non Fraud because the Fraud ones were very few. Now with the balanced dataset in place, the condition is not true anymore. In this case, the learning algorithm needs to construct a classifier that has a better accuracy rate of both classes. Because it is normal to produce classifiers with less accuracy than the original, the correct classification of Fraud instances, which represents the Fraud Detection, represents a much more important measure in this specific problem.

The next tables have the values of correctly classified instances and that permits to understand that SMOTE is the technique that generates a dataset which produces a classifier that classifies correctly almost the same number of instances than the original dataset and that Spread Subsample is the technique that allows the generation of the

dataset that creates a classifier that better classifies the Fraud Instances, allowing a better system for the detection of fraud.

	Correctly Classified Instances	Total Instances	Correctly Classified Fraud Instances	Total Fraud Instances
Singulars	1041	1202	9	64
Singulars Resample	985	1202	50	64
Singulars SMOTE 1400%	1018	1202	22	64
Singulars Spread Subsample	875	1202	52	64

	Correctly Classified Instances	Total Instances	Correctly Classified Fraud Instances	Total Fraud Instances
Societies	296	388	0	91
Societies Resample	250	388	50	91
Societies SMOTE 175%	292	388	48	91
Societies Spread Subsample	192	388	77	91

Figure 33 - Tables with the numbers of correctly classified instances for each classifier in both datasets with several different techniques to balance the dataset.

Processing Techniques (Boosting Algorithms)

The processing techniques are applied in the processing phase instead of pre-processing phase and, as described before, are iterative algorithms that use the learning algorithm j48 and in each iteration focus on the misclassified instances before. In figure 34 are the charts with the measures of each boosting algorithm in the two datasets used before.

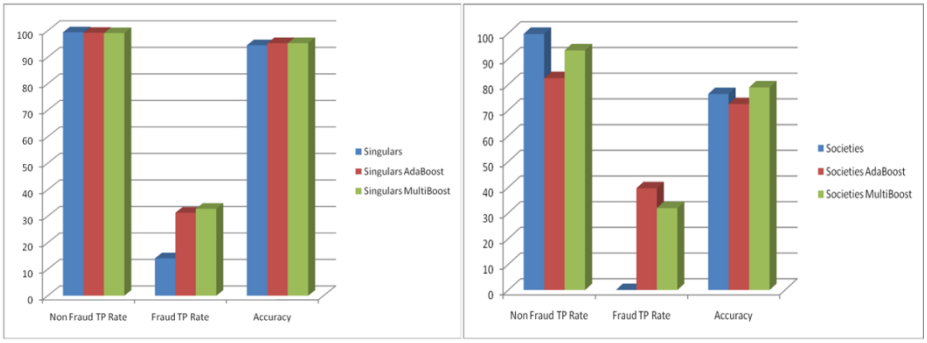


Figure 34 - Classifiers evaluation charts of both datasets with the appliance of two different boosting algorithms.

The results shows us that both boosting algorithms produces better results than the basic classifier. In the Singulars dataset the use of those algorithms achive a much higher

Fraud TP Rate and approximitly the same Non Fraud TP Rate and that generates a classifier with a bett accuracy and a much better fraud detection. On the associative dataset the results are not that perfect. Also revealing, there is more accuracy when MultiBoost is used and in this case, the result is lower in the Non Fraud TP Rate, in which it is compensated by the big raise in the Fraud TP Rate. This conclusions can be supported in full with the number of Correct Classified Fraud Instances and all instances. In the individual case both boosting algorithm produced better accuracy and almost the same correct classification of Fraud Instances and in the associative case, as described before AdaBoost can detect an higher number of Fraud Instances but is the MultiBoost who produce a better classifier.

	Correctly Classified Instances	Total Instances	Correctly Classified Fraud Instances	Total Fraud Instances
Singulars	1041	1202	9	64
Singulars AdaBoost	1050	1202	20	64
Singulars MultiBoost	1050	1202	21	64

	Correctly Classified Instances	Total Instances	Correctly Classified Fraud Instances	Total Fraud Instances
Societies	296	388	0	91
Societies AdaBoost	281	388	36	91
Societies MultiBoost	306	388	29	91

Figure 35 - Tables with the numbers of correctly classified instances for each classifier in both datasets the two different boosting algorithms

Balancing the dataset for Boosting Algorithms

As seen before, the techniques to balance the dataset and the boosting algorithms, when used separately, produce better results than a basic classifier. The next step was to combine both pre-processing and processing techniques. This means that the first used one of the techniques to balance the dataset (Resample, Spread Subsample or SMOTE) and then boosting algorithm was used to produce better results: MultiBoost. The results are shown in figure 36.

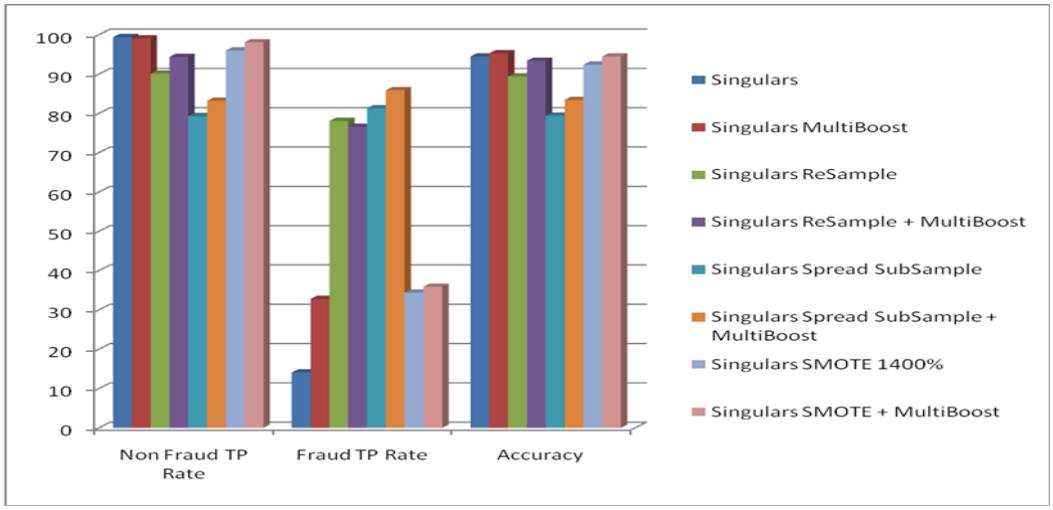


Figure 36 - evaluation chart of different techniques combination.

This graph shows that the aplication of a balance technique by its self generates better results, nevertheless the combination with Multiboost generates even better results. Resample and Multiboost produces better results than just Resample, Spread Subsample and Multiboost produces better results than just Spread Subsample and the same to SMOTE with MultiBoost. In the next table are the results of correctly classified instances from each test. The solution that achives better accuracy classifing is the combination between Spread Subsample and Multiboost but when looking to both correctly classified instances and correctly classified fraud instances is evident that the better solution is mix between Resample and Multiboost.

	Correctly Classified Instances	Total Instances	Correctly Classified Fraud Instances	Total Fraud Instances
Singulars	1041	1202	9	64
Singulars MultiBoost	1050	1202	21	64
Singulars ReSample	985	1202	50	64
Singulars ReSample + MultiBoost	1029	1202	49	64
Singulars Spread SubSample	875	1202	52	64
Singulars Spread SubSample + MultiBoost	919	1202	55	64
Singulars SMOTE 1400%	1018	1202	22	64
Singulars SMOTE + MultiBoost	1041	1202	23	64

Figure 37 - Tables with the numbers of correctly classified instances for each classifier.

Conclusions

This study case presents a real case of fraud detection with the problem of an unbalanced dataset. It was proved that any technique that transforms the dataset into a more balanced one conducts to a better classifier. It was also shown that boosting algorithms produces better results with the same dataset than the construction of a basic classifier. In the pre-processing phase Spread Subsample is the technique that permit to achieve a high rate of the minority class true positives. But it is important also to preserve a high accuracy classifier and in this matter Resample is the technique that shows better result in classifying correctly both minority and majority class. In the processing phase both boosting algorithms produces better classifiers but in the specific case of Societies dataset Multiboost revels a high performance. This separate study of pre-processing and processing techniques has served to prove that both solutions generate better results. The next logical step was to test a mixture of both techniques and the results are even better. In this test the use of Resample technique together with Multiboost produces a classifier with accuracy of 94.4% and a Fraud TP Rate of 76.6%. The results presented here means that combining both techniques is, in fact, the best way to solve Fraud Detection problems, when dealing with an unbalanced dataset.

4 Improving Fraud Detection with Social Patterns

The detection of fraud, as shown before, is a very hard task and there is no efficient way of detecting it. As what concerns to classification there are two big problems: the datasets that are very unbalanced, where the fraudulent instances are very outnumbered facing the non fraudulent ones, and the fact that the cost of the misclassification of a fraudulent instance is higher than the cost of the misclassification of a non fraudulent one. Because of these two issues applying classification techniques to detect fraud is not trivial. Besides that, the majority of fraud types are too hard to execute alone and that leads to the creation of fraudulent associations with the goal of committing fraud in a "safer" way. The social structure of those associations is nothing more than social networks (Gomes, 2008). Keeping in mind those fraudulent and non fraudulent organizations can have different relationships, the main goal is to detect patterns that are specific to fraudulent organizations, in order to better distinguish among them.

There are various factors that suggest the use of social networks when identifying and fighting fraud. Very often, fraud is not only committed inside an organization but also between organizations. An example of fraud of this type is the transaction of goods or services without paying the VAT. Besides that, there is another type of fraud called Carrousel that consists in a set of transactions between several organizations where some of them reclaims the refund of VAT from state, when others do not pay the VAT to the state. These two cases are examples of possible fraud committed with relations between organizations. Another important case to look at a social network is the fact that if someone owns a fraudulent organization and has another independent organization there are high probabilities that this second organization could also be a fraudulent one. Apparently, the use of information retrieved from the social network built with the relationships between entities may result in a better classification and consequently in an improved means for the detection of fraud.

From a sociological point of view, every people and organizations are not alone in our society: they are organized groups and have relationship between them. The social network is the structural and hierarchical representation of those relationships between entities. Is a fact that two entities are only connected if they have something in common and they share some kind of information and this share of information is the relationships that could be commercial (transactions of goods or services), professional (worker, employee) or personals (marriage, offspring) (Jamali, et al., 2006). The importance of entities in the network is usually classified by the number of relationships that they own.

An entity that has a higher number of links is a more central entity than one with fewer connections and this centrality represents the importance of it in the network.

Keeping in mind a famous Portuguese saying: “Diz-me com quem andas e dir-te-ei quem és” (Tell me who your friends are and I will tell you who you are), meaning that the people with whom a person relates to defines who this person really is. For this reason, people who are close to each other may commit fraud, it is more probable that you also commit it, is possible to imagine that entities connected with fraudulent organizations have higher probability to be fraudulent too. Moreover, it is important to pay attention at the “Small World Phenomenon” (Milgram, 1969) that defends that all the people of world are connected between them with very few intermediaries. Also known as the “six degrees of separation”, it means that there is only six relationships max between any two different people. Indeed, the world is a giant social network with short distances between people and organizations. This means, in fact, that the real world social network is a very complex networking system made-up of influence of an entity in the social network is only real to the nearest connections.

4.1 Social Pattern-based Classification

Considering those issues, we propose a new method to detect fraud that can be applied when there is information about the relationships between entities. This new method is based on the enrichment of the characterization of instances with information about their social relations. In particular, the method starts by considering the data about entities that are described in some data source, and the social relations between them. From the set of relations, we identify the patterns that are common to fraudulent entities. Using these patterns as binary attributes, each entity becomes to be described by its original features plus another one for each pattern. These new instances may then be entered in the training process, using balancing techniques as described in previous chapters.

Figure 38 represents the process that originates a dataset enriched with social network patterns. Each step is described in detail below.

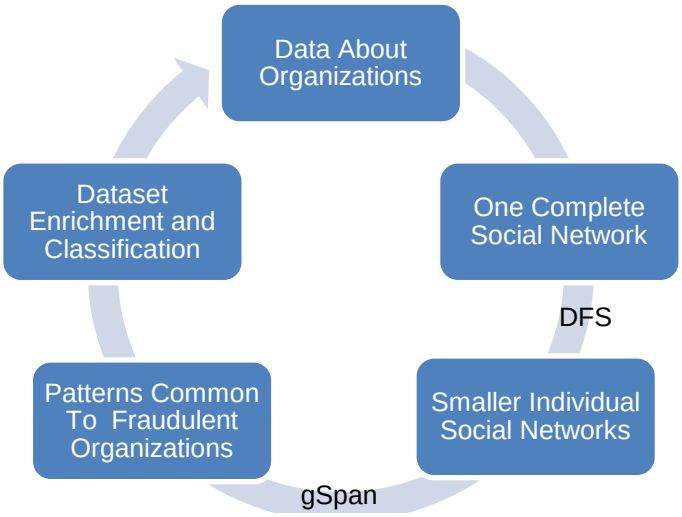


Figure 38 - Process of data enrichment with social networks patterns.

4.1.1 Data Concerning Organizations

In organizations there can be type of data, which are known as static and social data. Static data corresponds to the data resulting from any attribute that describes static characteristics of entities, such as physical characteristics. By social data we mean, attributes that can be used to identify the social relationships of an entity. Notwithstanding, it is possible with static attributes to located sociological attributes, such as the age of a person. Nevertheless, static attributes cannot be used to establish social relations, as social data does.

Deriving from social data describing the relationships between entities (people and organizations), a global social network is created, including all the relations known between entities. Note that this data can be retrieved either from the organizations' databases or from State databases.

4.1.2 One Complete Social Network

The global social network can be seen as a graph formed by nodes (which are the entities) and edges, the connections between nodes which correspond to the relations between entities. Formally, a graph G is a five element tuple G = (V, E, ∑v, ∑e, L) where V is a set of nodes and E ⊂ V x V is a set of undirected edges. ∑v and ∑e are the sets of vertex labels and edge labels, respectively. The labeling function L defines the mappings V → ∑v and E → ∑e. In a perfect world, all the nodes should be labeled, as fraud and non-fraud in the case of fraud detection, but in the real world only a few of them are

classified. Those classified nodes as fraud are the most important because is from them that the common patterns to fraudulent entities are detected.

4.1.3 Smaller Individual Social Networks

From the “Small World Phenomenon”, it is clear that the impact of each entity is limited to its closest neighbor and this being the reason not for it to be completely used in the global social network. It is divided in smaller social networks where each central entity corresponds to each organization that will be classified as fraud or non fraud.

The next phase in the process consists in the partition of the global social network into several sub-networks for each fraudulent entity. Because the goal is to detect patterns that are common to the fraudulent entities, this partition provides a set of social networks in which the patterns are present. To do this partition is necessary to use a search algorithm as Deep First Search algorithm (DFS). DFS starts the search in the root (nodes) and goes through each branch before turning back (Pearl, 1984). With this algorithm is possible to detect all entities that are connected with the one that is being analyzed and is in this phase that must be defined the depth considered for each social network. An organization that relates with a person is a social network with one level. If the person relates to another organization is a two levels social networks. The DFS when start searching in a branch goes as deep as it can but we can change that by defining a variable depth that determinates how deep the DFS can go. By defining the depth of the DFS we establish the maximum number of levels that the social network has. The result of the application of DFS to all the nodes classified as fraud is a set of several smaller social networks.

4.1.4 Common Patterns to Fraudulent Organizations

The set of those social networks is the input for the pattern detection phase, where common patterns to fraudulent entities are identified.

Since social networks can be seen as graphs, one of the pattern mining algorithms that can be used is the gSpan. In this particular case, gSpan detects parts of graphs, called sub graphs, which are common to several of the graphs present in the original set (Yan, et al., 2002). gSpan is to graphs as the same way that Apriori is to items. Apriori is an algorithm that detects Association Rules and this means that Apriori detects which items appear frequently in a set with various lists of items. For example Apriori detects that beer and dippers appear several times together is the list of products bought in a supermarket and gSpan detects that a fraudulent organization do transactions with another fraudulent

organization. The main variable to this algorithm is frequency that means how frequent a pattern has to be to be detected. This is measured in terms of percentage, meaning that a percentage of organizations had to have this pattern applied to their social networks for it to be detected. The appliance of gSpan to all the fraudulent entities results in a set of frequent sub graphs. This means that the result is a set of patterns that are frequently present in the social networks of fraudulent entities.

4.1.5 Dataset Enrichment and Classification

The last step is to detect if an entity has or not those patterns. In this phase a table is created where each row describes a different entity. Moreover, each column either corresponds to a static attribute or represents a different pattern. Those last attributes are filled with 0 and 1 meaning if the pattern is present or not in the social network of the entity.

This table is then used as the input for the training process to learn an accurate classifier. This train can be made using any of the known classification methods (from decision trees to support vector machines), and even considering balancing techniques as described above.

4.2 Instantiation of the process to the VAT fraud problem

In VAT fraud detection problem it is necessary to have a set of data with information about the relations between entities. This data is retrieved from the organization taxes declarations and records. In the organization's VAT declaration is possible to know the transactions between organizations which were made, besides all other organizations financial data. From the person's IRS declaration is possible to know who paid her salary (the organization that employees he) and who is the rest of her family (if the person is married and have offspring's). Because the goal of this work is to help in the classification phase this set of data must have entities labeled as fraud and non-fraud.

In conclusion in the VAT fraud detection the basic dataset is formed with data from the VAT declaration form and the social network is built with information contained in the VAT and IRS databases.

4.2.1 Example

As seen before in an unbalanced dataset, the classifier can classify all instances within the majority class, creating a classifier where the accuracy is the same that the

percentage of instances from the majority class. In an example where there are 8 non fraudulent instances (80%) and 2 fraudulent ones (20%), a classifier that classifies all instances as non fraudulent has an accuracy of 80%. Because of this and because the organizations have relations between them when this data is available it can be used to help in the fraud detection problem.

Consider the example in Figure 39, where organizations are represented with letters and people with numbers: there are 10 organizations and 2 of them are fraudulent. Also consider that there is data about the relations between entities. In this manner, this data can be worked to form a social network that links most of the organizations. Let's assume that organizations A and C are the fraudulent ones and B and D are some of the non fraudulent. All other 6 non fraudulent organizations don't have relations relevant for the example.

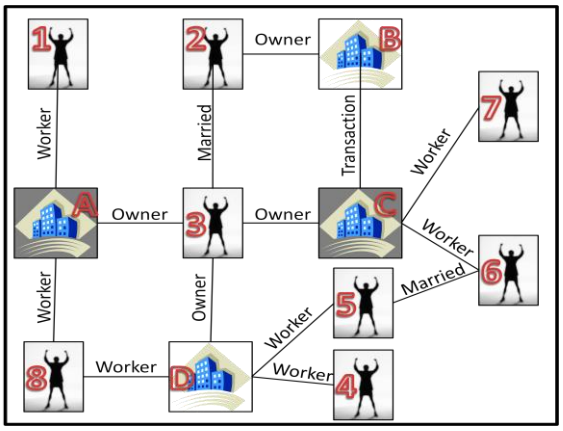


Figure 39 - An example of a social network.

After creating the social network is necessary to generate the smaller, individual social networks to each organization. This is done by applying the DFS to each organization and generating the social networks of each organization. In the example above the DFS is applied with the depth equal to 2. This means that the social networks will only have 3 levels, the root and the two next links. Figure 40 shows all 4 individual social networks.

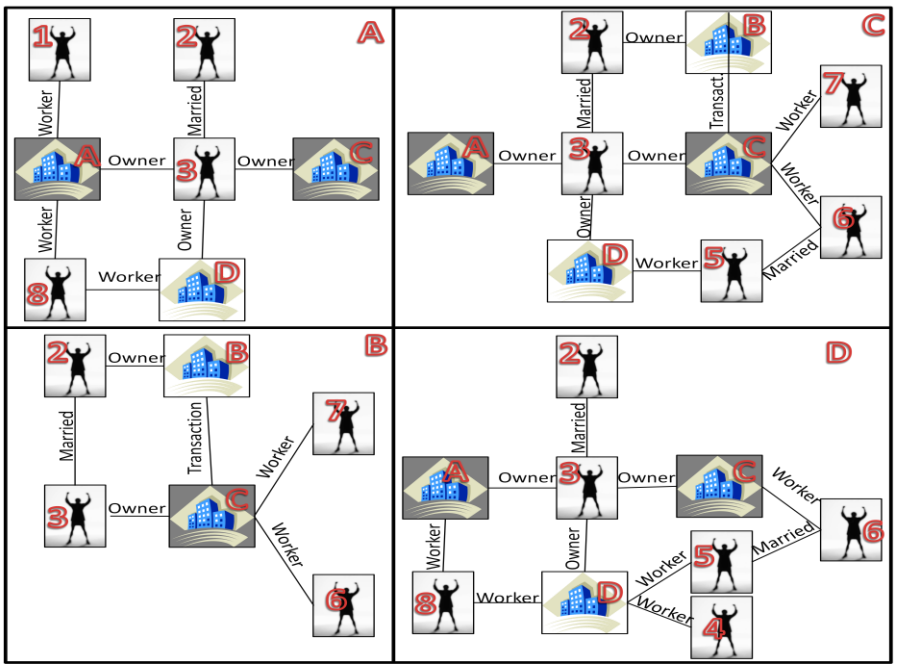


Figure 40 - The 4 social network of each organization.

The next step is the pattern detection. This is done by applying gSpan to those individual social networks. Here gSpan will find patterns that are common to various social networks. In the application of gSpan, it is necessary to define the minimum frequency of a pattern for it to be detected. In this example, because 50% of social networks are from fraudulent organization, the minimum frequency will be set to 50%, in order to detect only patterns that are common to at least two of the graphs.

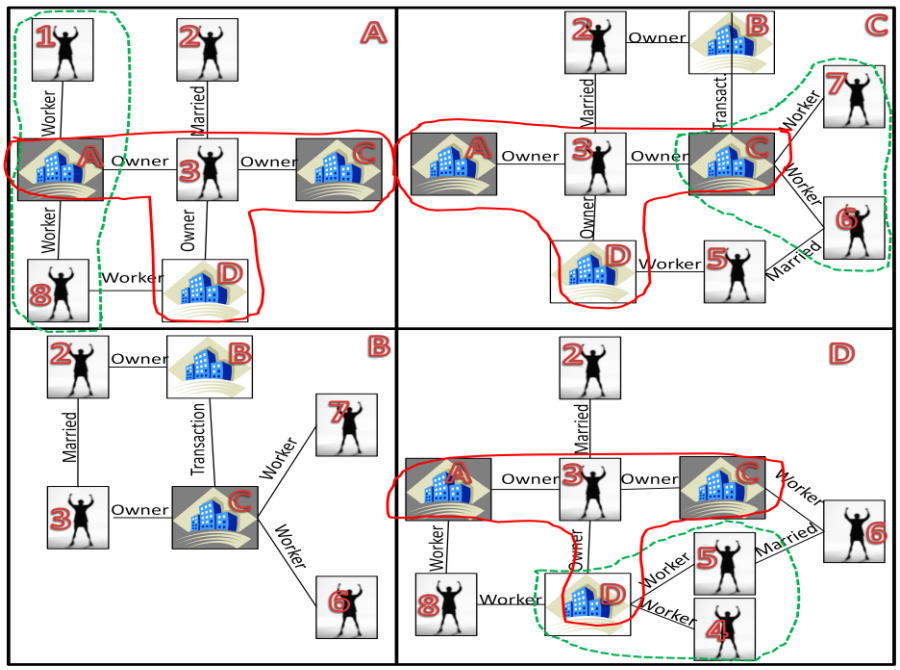


Figure 41 - Some patterns detected with gSpan..

Because this is a small example of a social network with very common branches, a lot of patterns are detected. Patterns like an organization with a worker, an organization with an owner or, to say the least, an organization with an owner that is married are patterns present in all four social networks. For this reason, no value is added to the problem solution. In this example there are no patterns that appear only in the fraudulent organizations social network but there are two patterns shown in figure 41 that appear in the two fraudulent organizations and in one non fraudulent. These patterns are important and could help in the classification process. The patterns “one organization that has an owner that own another two organizations” and “one organization with two workers” are two important patterns in this example. These issues will help to exclude organization B as a fraudulent organization.

Original Data		
	Class	Profit Declared
1	Non Fraud	17000€
2	Non Fraud	20000€
3	Non Fraud	10000€
4	Non Fraud	10000€
5	Non Fraud	35000€
6	Non Fraud	19000€
B	Non Fraud	7500€
D	Non Fraud	19000€
A	Fraud	15000€
C	Fraud	5000€

Enriched Data			
	Class	Profit Declared	Pattern 1
1	Non Fraud	17000€	-
2	Non Fraud	20000€	-
3	Non Fraud	10000€	-
4	Non Fraud	10000€	-
5	Non Fraud	35000€	-
6	Non Fraud	19000€	-
B	Non Fraud	7500€	0
D	Non Fraud	19000€	1
A	Fraud	15000€	1
C	Fraud	5000€	1

Figure 42 - Some patterns detected with gSpan.

The last phase is the enrichment of the original dataset includes information about patterns in the social network. Figure 42 has two tables where the first one depicts the original dataset and the second one has de enriched dataset. A classifier built with data from table one can classify as non fraud the organizations that declare a profit higher than 5000€ and as fraud the ones with profit equal or lower than 5000€ and it has an accuracy of 90% because it classifies correctly 9 out of 10 organizations. But with the enriched data (second table) with information about one of the patterns, the classifier could be perfect, which is to say 100% accurate, because it classifies all instances with profit declared lower than 19000€ and with pattern 1 with value 1 as fraud and the rest of the instances as non fraud.

This simple example shows and illustrates the steps to use data about relationship of organizations to help in the detection of fraud. As obvious in the real world it is impossible to have a classifier that is 100% accurate. Nevertheless, the system can, definitely, improve the real world classifiers. The following chapter addresses real world and real data case studies pertinent to the problem stated, along with the application of a possible solution.

5 Case Studies

This chapter presents case studies with real data to which it is applied the techniques described in chapter 3 to deal with unbalanced dataset. Also, the datasets used in chapter 3 are enriched with Social Network Patterns common to fraud, as explained in the preceding chapter. This chapter presents the results of the experiences made that serve to prove that the use of Social Networks information can improve the detection of fraud.

5.1 Social Networks

First of all, a comparison between classifiers built only with VAT attributes and classifiers built with data from social network patterns will be done. The basic dataset are the same as the ones used in chapter 3: the Singulars dataset, which is very unbalanced, with 94% of non fraud instances and 6% of fraud instances in a total of 3133 instances and the Societies dataset , less unbalanced, with 74% of non fraud instances and 26% of fraud instances in a total of 777 instances.

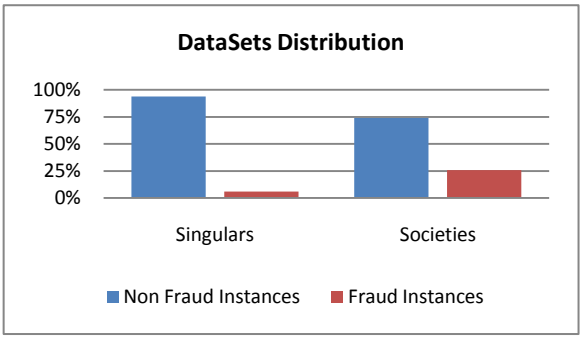


Figure 43 - The distribution of Singulars and Societies datasets.

Each dataset has three different sets for testing: The first one has only VAT attributes (static attributes), the second on has only Social Networks Patterns attributes (social attributes), and the last one has both VAT and Social Networks Patterns attributes. The goal is to identify which of these 3 different sets of attributes generates the better classifier. The first test is the comparison between the classifier built with the 3 different set of attributes and the results are presented in the next figure.

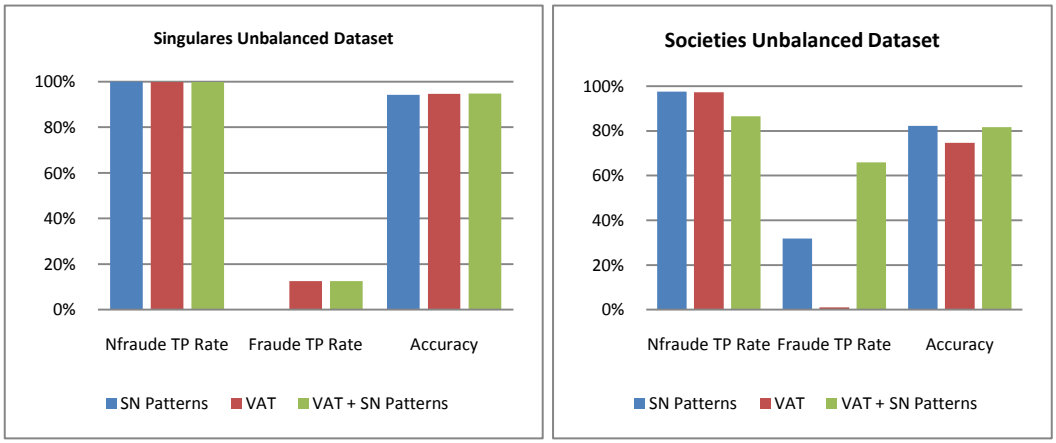


Figure 44 - Results from the classifiers built whit the 3 different kinds of attributes.

Analyzing the results above, it is not complicated to observe that with a very unbalanced dataset, such as the Singulars one, the classifier is not acceptable and even with the introduction of attributes from the Social Networks the results, it does not improve it very much. In case of the Societies chart, with a less unbalanced dataset, the use of information about social network patterns increases the quality of the classifier. In this situation, the combination of the VAT and Social Network attributes produces better results.

5.2 Social Networks and Balancing Techniques

With the results obtained before it is impossible to conclude that the classifier built with attributes from social networks was better than the one built with only VAT attributes. Therefore it is important to study how those different datasets behaves when combining it with techniques to balance the dataset (Resample, SpreadSubSample and SMOTE) and with boosting algorithms (Adaboost and Multiboost). Those techniques are the same described and used in chapter 3.

Singulars

The first dataset used was the Singulars dataset , which is the more unbalanced one. In this first comparison, using this dataset the classifiers built with VAT attributes and with VAT and Social Networks attributes, were the most accurate when detecting fraudulent instances. In figure JSF are the results from all experiences done combining the 3 datasets with different kinds of attributes.

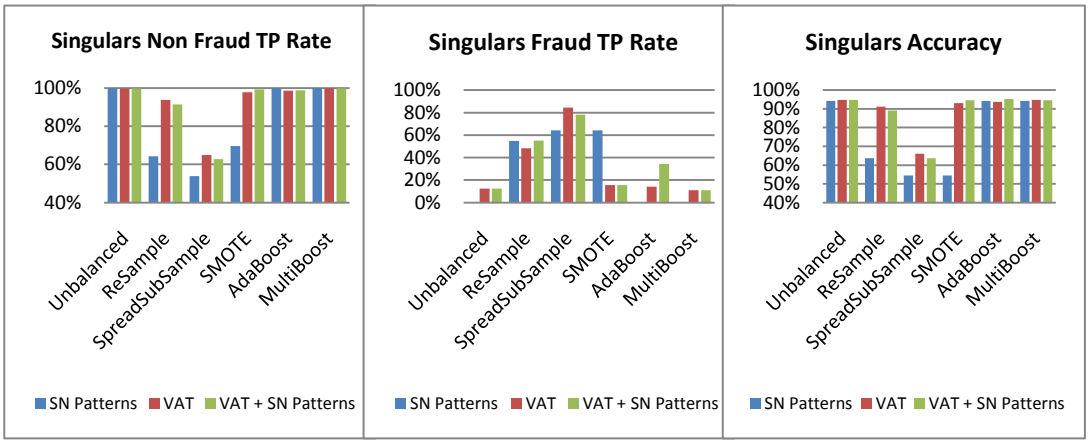


Figure 45 - Results from the Singulars classifiers built whit the 3 different kinds of attributes and with the techniques to deal with unbalanced datasets.

It can be said that the use of Social Networks (SN) attributes in the Singulars dataset does not produce a high improvement in the classifiers. The use of ReSample to balance the dataset is a better technique and, in this case, the dataset with VAT and SN attributes is the one that produces the most accurate classifier with a Fraud True Positive Rate above 50%.

The following table presents the testing results of the Singulars dataset and helps to identify the different results with the different techniques.

Singulars		Unbalanced	ReSample	SpreadSubSample	SMOTE	AdaBoost	MultiBoost
Non Fraud TP Rate	**SN Patterns**	100%	64,30%	53,90%	69,60%	100%	100%
	VAT	99,80%	93,80%	64,90%	97,80%	98,70%	100%
	VAT + SN Patterns	99,90%	91,50%	62,70%	99,40%	98,90%	99,80%
Fraud TP Rate	**SN Patterns**	0%	54,70%	64,10%	64,10%	0%	0%
	VAT	12,50%	48,40%	84,40%	15,60%	14,10%	10,90%
	VAT + SN Patterns	12,50%	55,00%	78,10%	15,60%	34,40%	10,90%
Accuracy	**SN Patterns**	94,20%	63,70%	54,50%	54,50%	94,20%	94,20%
	VAT	94,70%	91,20%	66,10%	93,00%	93,80%	94,80%
	VAT + SN Patterns	94,80%	89,10%	63,60%	8,00%	95,20%	94,60%

Figure 46 - All tests results with the Singular dataset.

Societies

In the case of Societies the results obtained are very different. In this case, the use of information about Social Networks (SN) produces a classifier with higher rates in fraud

detection. The same tests done in the Singulars case were done in this case too and the results are presented in figure 47.

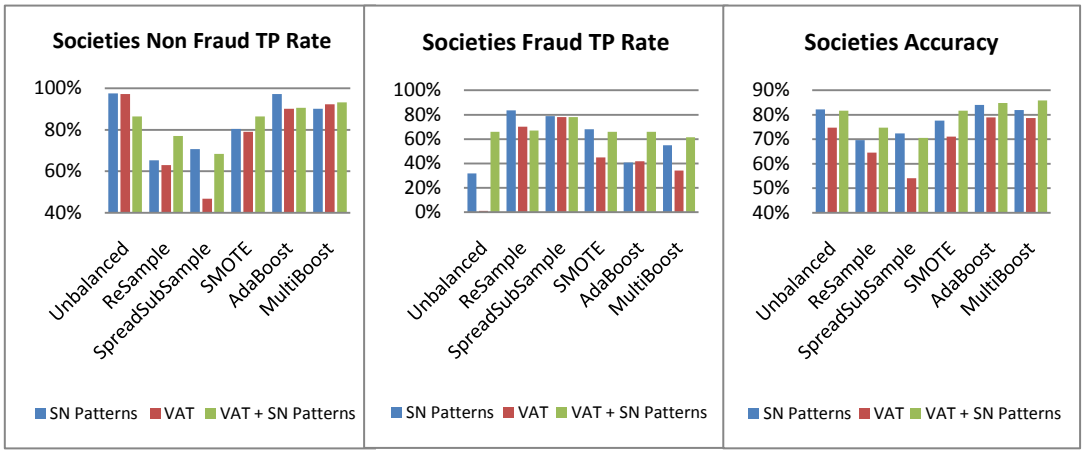


Figure 47 - Results from the Societies classifiers built whit the 3 different kinds of attributes and with the techniques to deal with unbalanced datasets.

The results in this case show that the use of SN attributes helps a lot in the classification stage. The use of Resample to balance the dataset with only SN attributes is the classifier that has the higher Fraud TP Rate but it has a low Non Fraud TP Rate which results in a low accuracy. Although that there are better results in terms of accuracy and high Fraud TP Rate. With AdaBoost and a dataset composed with VAT and SN attributes is produced a classifier with 85% of accuracy and a Fraud TP Rate of 67%.

All other values for the tests realized are presented in the next table.

Societies		Unbalanced	ReSample	SpreadSubSample	SMOTE	AdaBoost	MultiBoost
Non Fraud TP Rate	**SN Patterns**	97,60%	65,30%	70,70%	80,50%	97,30%	90,20%
	VAT	97,30%	63%	46,80%	79,10%	90,20%	92,30%
	VAT + SN Patterns	86,50%	77,10%	68,40%	86,50%	90,60%	93,30%
Fraud TP Rate	**SN Patterns**	31,90%	83,50%	78,70%	68,10%	40,70%	54,90%
	VAT	1,10%	70,30%	78,00%	45,10%	41,80%	34,10%
	VAT + SN Patterns	65,90%	67,00%	78,00%	65,90%	65,90%	61,50%
Accuracy	**SN Patterns**	82,20%	69,60%	72,40%	77,60%	84%	82%
	VAT	74,70%	64,60%	54,10%	71,10%	78,90%	78,60%
	VAT + SN Patterns	81,70%	74,70%	70,60%	81,70%	84,80%	85,80%

Figure 48 - All tests results with the Societies dataset.

5.3 Discussion

In this chapter several experiments were conducted to test if the use of Social Network (SN) information could help on improving the detection of fraud. In the Singulars case the use of SN attributes alone or with VAT attributes do not improve very much the classifiers' quality and this can happen due to the very unbalanced state of the dataset. Although in the Societies case the results were very different. The use of information retrieved from common patterns of fraudulent organizations contributed significantly to the creation of a better classifier. The graphs presented in figure 49 show the difference between the uses of SN information in comparison to the use of only VAT attributes.

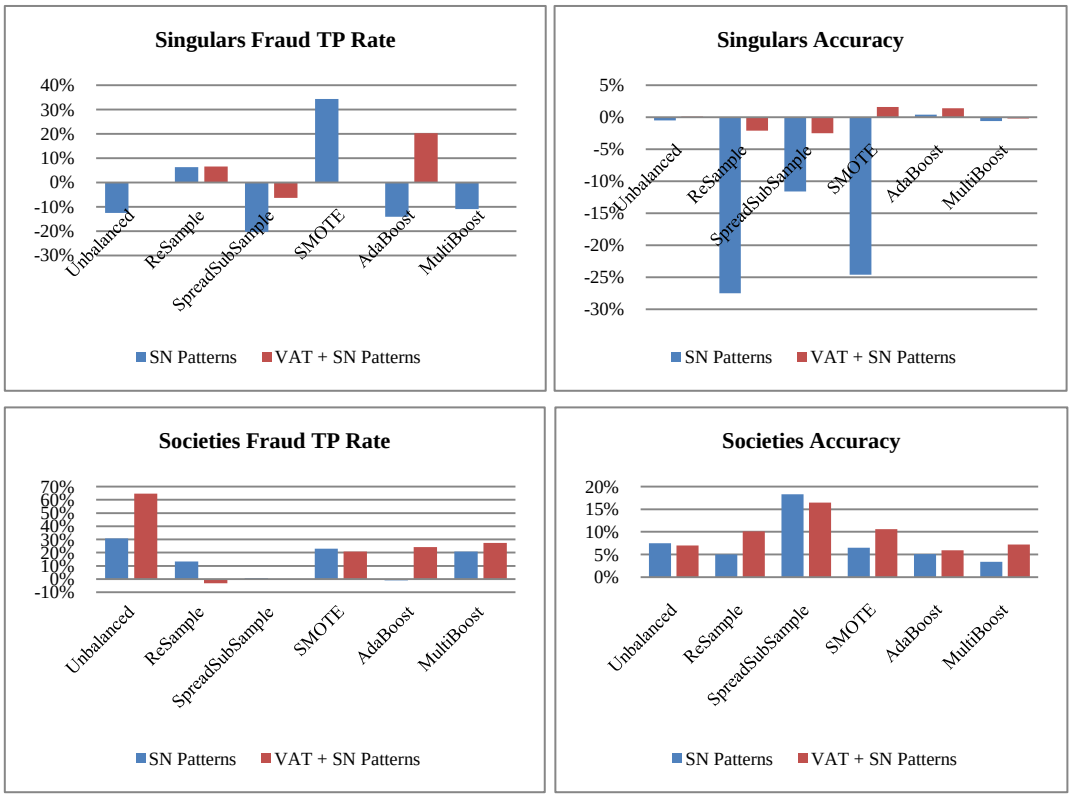


Figure 49 - The charts show the gain or loss in terms Fraud TP Rate and Accuracy using SN attributes and SN and VAT attributes in comparison with only VAT attributes.

As seen in the charts, the values associated with the Societies dataset when using SN attributes only or with VAT attributes are positive, thus revealing a better classifier than the one constructed with only the VAT attributes. The same does not happen when using the Singulars dataset where some of the results are negative, this meaning that the classifiers constructed with that kind of data are less worthy than the ones constructed

with only VAT attributes. This chart does not indicate which technique is the best one because it only compares the results with the one obtained using the same technique but a different dataset.

In conclusion, with the results obtained in this set of tests, it is possible to state that the use of a dataset enriched with information about the social network of an organization can lead to better results in some cases. Although that it is impossible to say that it work with all the cases as it was shown in the Singulars case.

6 Conclusions

There are several techniques and methods that can be deployed to assist in fraud detection. The unbalanced dataset present a problem. Nevertheless, there are two possible solutions to handle the situation, including the introduction of artificial samples at a pre-processing phase and the changing of error costs at a processing level. Besides these, it is possible to apply outliers as detection techniques to flag fraudulent records that deviate from non-fraudulent ones. The work presented here presents the application of the referenced techniques to a real-work dataset and the comparison of the results obtained. All the experiments and tests conducted prove that when classifiers are used within unbalanced datasets they are poor indicators, but when applied along with adequate and specific techniques to deal with this problem, the results are, in fact, improved.

Besides that, another issue is studied here: the use of social networks to help improve the fraud detection classifiers. This done, keeping in mind that organization and people are connected between themselves and that some types of fraud are perpetuated not by individuals alone but also by an association of these, when proposed the use of information concerning the social networks of organizations and people , in helping in fraud detection. This study demonstrates that a dataset when strengthen with information concerning patterns detected in social networks and that are common to fraudulent organizations produces some interesting results. As a result, the use of this sort of information, in some case, may lead to a better classifier and, indeed, to better results in fraud detection.

When trying to identify fraud of organizations with a lot of social connection the better solution is to use a dataset enriched with pattern discovered in the social networks and use SMOTE to balance the dataset. This is the example of using SMOTE in the VAT + SN patterns dataset which produced the best classifier for Fraud Detection.

7 References

Adriaans P., Zantinge, D. Data Mining [Book]. - Harlo. England : Addison-Wesley, 1996.

Ahumada H., Grinblat,G., Uzal, L., Granitto, P., Ceccatto, A. A new hybrid approach to highly imbalanced classification problems [Conferência] // Eighth International Conference on Hybrid Intelligent Systems. - [s.l.] : IEEE, 2008. - pp. 386--391.

Batista G., Prati, R., Monard, M. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data [Conferência] // Sigkdd Explorations. - [s.l.] : ACM, 2004. - Vols. Volume 6, Issue 1. - pp. 20—29.

Bradley A. The Use of the Area Under the ROC Curve in the Evaluation of the Machine Learning Algorithms [Conference] // Pattern Recognition Society. - [s.l.] : Elsevier Science Ltd, 1996.

Brause R., Langsdorf, T., Hepp, M. Neural Data Mining for Credit Card Fraud Detection [Conferência] // Tools with Artificial Intelligence. - [s.l.] : IEEE, 1999. - pp. 103-106.

Cahill M., Lambert, D., Pinheiro, J., Sun, D. Detecting Fraud in the Real World [Secção do Livro] // Handbook of massive data sets. - USA : Kluwer Academic Publishers, 2002.

Chawla N. C4.5 and Imbalanced Data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure [Conferência] // Workshop on Learning from Imbalanced Datasets II. - USA : ICML, 2003.

Chawla N., Bowyer, K., Hall, L., Kegelmeyer, W. :SMOTE: Synthetic Minority Over-sampling Technique [Jornal] // Journal of Artificial Intelligence Research 16. - [s.l.] : AI Access Foundation and Morgan Kaufmann Publishers , 2002. - pp. pp. 321--357.

Chawla N., Japkowicz, N., Icz, A. Special Issue on Learning from Imbalanced Data Sets [Conferência] // Sigkdd Explorations. - [s.l.] : ACM, 2004. - Vol. 6. - pp. 1--6.

Dorronsoro J., Ginel, F., Sanchez, C., Cruz, C. Neural Fraud Detection in Credit Card Operations [Conferência] // Transactions on Neuronal Networks. - [s.l.] : IEEE, 1997. - Vols. 8, nº4. - pp. 827--834.

Ferdousi Z., Maeda, A. Unsupervised Outlier Detection in Time Series Data [Conferência] // Proceedings of the 22nd International Conference on Data Engineering Workshops. - [s.l.] : IEEE, 2006. - pp. 51--56.

Gomes João Nascimento Social mining no combate à fraude [Journal]. - [s.l.] : TECH&BIZZ - Novabase, 2008.

Guo H., Viktor, H Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach. [Conferência] // Sigkdd Explorations. - [s.l.] : ACM, 2004. - Vols. 6, Issue 1. - pp. 30--39.

Hall L. Data mining from extreme data sets: Very large and/or very skewed data sets [Conferência]. - [s.l.] : IEEE, 2001. - p. 2555.

Han H., Wang, W., Mao, B-H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning [Conferência] // ICIC 2005, Part I, LNCS 3644. - [s.l.] : Springer-Verlag Berlin Heidelberg, 2005. - pp. 876--886 .

Hawkins S., He, H., Williams, G., and Baxter, R. Outlier Detection Using Replicator Neural Networks [Conferência] // Procedings of the 5º Internacional Conference Data Warehousing and Knowledge Discovery. - [s.l.] : ACM, 2002. - pp. 170-180.

He G.,Han, H.,Wang, W. An Over-sampling Expert System for Learning from Imbalanced Data Sets [Conferência]. - IEEE : [s.n.], 2005. - pp. 537--541.

Hilas C., Sahalos, J. User Profiling for Fraud Detection in Telecommunication Networks [Conferência] // 5th International Conference on Technology and Automation . - Greece : IEEE, 2005.

Jamali Mohsen e Abolhassani Hassan Different Aspects of Social Network Analysis [Jornal]. - [s.l.] : IEEE, 2006.

Jensen David e Nevile Jennifer Data mining in social network [Jornal]. - 2002.

Joshi M., Kumar, V., Agarwal, C. Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements [Conferência] // 1st International Conference on Data . - USA : IEEE, 2001.

Kim H., Savoldi, A., Lee, H., Yun,S., Lee, S., Lim, J. Design and Implementation of a Tool to Detect Accounting Frauds [Conferência] // International Conference on Intelligent Information Hiding and Multimedia Signal Processing. - [s.l.] : IEEE, 2008. - pp. 547--552.

Kotsiantis S., Koumanakos, E., Tzelepis, D., Tampakas, V. Forecasting Fraudulent Financial Statements using Data Mining [Conferência] // Proceedings of International Conference on Signal Processing . - [s.l.] : International Journal of Computational Intelligence, 2006. - Vols. 3, Nº2. - pp. 1304--2386.

Kou Y., Lu C-T., S. Sirirat. Survey on Fraud Detection Techniques [Conference] // International Conference on Networking, Sensing and Contro. - USA : IEEE, 2004. - Vol. 2. - pp. 749- 754 .

Koufakou A., Ortiz, E.G., Georgiopoulos, M., Anagnostopoulos, G.C., Reynolds, K.M. A Scalable and Efficient Outlier Detection Strategy for Categorical Data [Conferência] // 19th International Conference on Tools with Artificial Intelligence. - [s.l.] : IEEE, 2007.

Kumar A., Nagadevara, V. Development of Hybrid Classification Methodology for Mining Skewed Data Sets – A Case Study of Indian Customs Data [Conferência]. - [s.l.] : IEEE, 2007. - pp. 584--591.

Milgram Stanley An Experimental Study of the Small World Problem [Secção do Livro]. - 1969.

Pearl J. Heuristics: Intelligent Search Strategies for Computer Problem Solving [Livro]. - [s.l.] : Addison-Wesley, 1984.

Pelayo L., Dick, S. Applying Novel Re-sampling Strategies To Software Defect Prediction [Conferência] // Fuzzy Information Processing Society. - [s.l.] : IEEE, 2007. - pp. 69--72 .

Phua C., Alahakoon, D., Lee, V. Minority Report in Fraud Detection: Classification of Skewed Data [Conferência] // Sigkdd Explorations / ed. 50--59. - [s.l.] : ACM, 2004. - Vols. 6, Issue 1.

Radcliffe J. The insurance industry's use of databases to prevent and detect fraud and improve recoveries. [Conferência] // European Convention on Security and Detection Conference publication No. 408. - [s.l.] : IEEE, 1995. - pp. 2216--224.

Shen A.,Tong R., Deng, Y. Application of Classification Models on Credit Card Fraud Detection [Conferência] // International Conference on Service Systems and Service Management. - IEEE : [s.n.], 2007. - pp. 1--4.

Taniguchi M., Haft, M., Hollmkn, J.,Tresp, V. Fraud detection in communication networks using neural and probabilistic methods [Conferência] // International Conference on Acoustics, Speech and Signal Processing. - [s.l.] : IEEE, 1998. - Vol. 2. - pp. 1241--1244.

Vo N., Won, Y. Classification of Unbalanced Medical Data with Weighted Regularized Least Squares [Conferência] // Frontiers in the Convergence of Bioscience and Information Technologies. - [s.l.] : IEEE, 2007. - pp. 347--352 .

Wang D., Wang, Q., Zahn, S., Li, F., Wang, D. A Feature Extraction Method for Fraud Detection in Mobile Communication Networks [Conferência] // Proceedings of the 5'h World Congress on Intelligent Control and Automation. - China : IEEE, 2004. - pp. 1853--1856.

Wang J., Xu, M., Wang H., Zhang, J. Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embeddin [Conferência] // International Conference on Signal Processing. - [s.l.] : IEEE, 2006.

Wang J-H., Liao,Y-L., Tsay, T-M., Hung, G. Technology-based Financial Frauds in Taiwan: Issues and Approaches. [Conferência] // International Conference of Systems, Man and cybernetics. - Taiwan : IEEE, 2006. - pp. 1120--1124.

Weatherford M. Mining for Fraud [Jornal] // IEEE Intelligent Systems. - [s.l.] : IEEE, 2002. - pp. 4--7.

Weiss G. Mining with Rarity: A Unifying Framework. [Conferência] // Sigkdd Explorations. - [s.l.] : ACM, 2004. - Vols. 6, Issue 1 . - pp. 7--19.

Yan Xifeng and Han Jiawei gSpan: Graph-Based Substructure Pattern Mining [Journal]. - 2002.

Yang Q., Wu, X. 10 challenging problems in data mining research. [Conferência] // International Journal of Information Technology & Decision Making . - [s.l.] : IEEE, 2006. - Vol. 5. - pp. 597--604 .

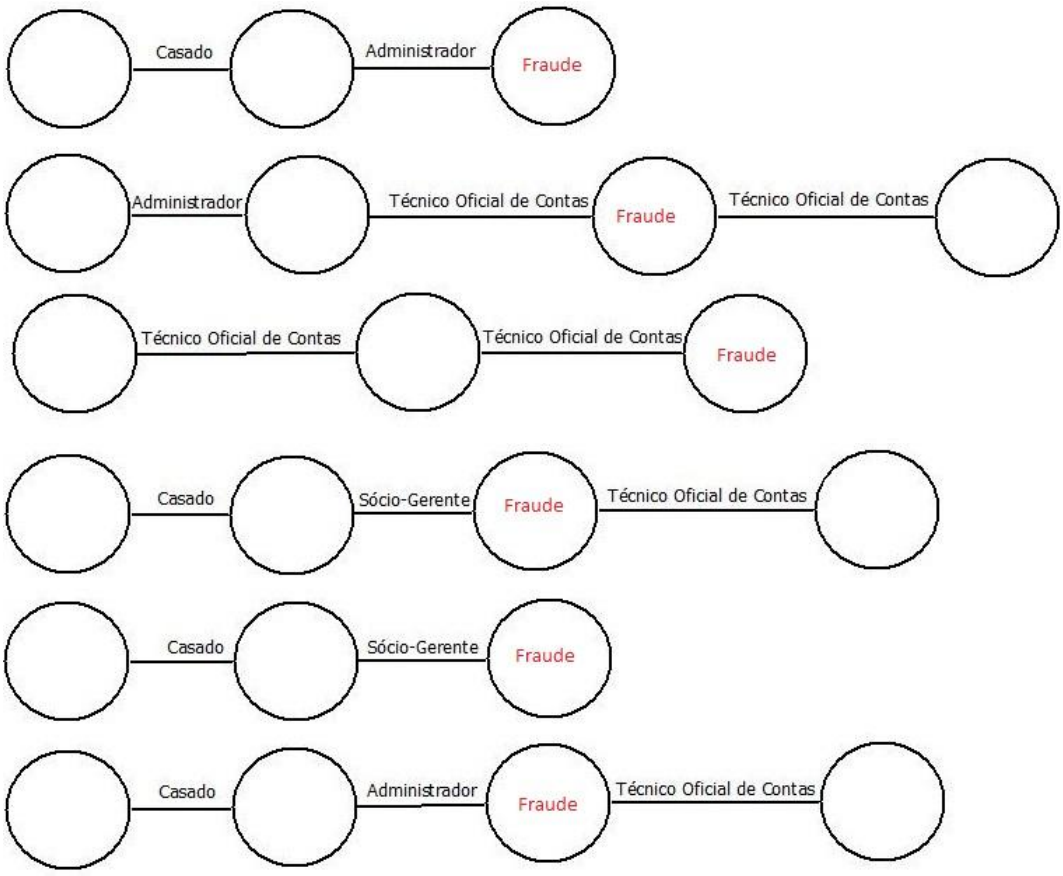
Yue D.,Wu, X.,Wang, T., Chu, Y. Review of Data Mining-based Financial Fraud Detection Research [Conferência] // International Conference on Wireless Communications, Networking and Mobile Computing. - [s.l.] : IEEE, 2007. - pp. 5519--5522.

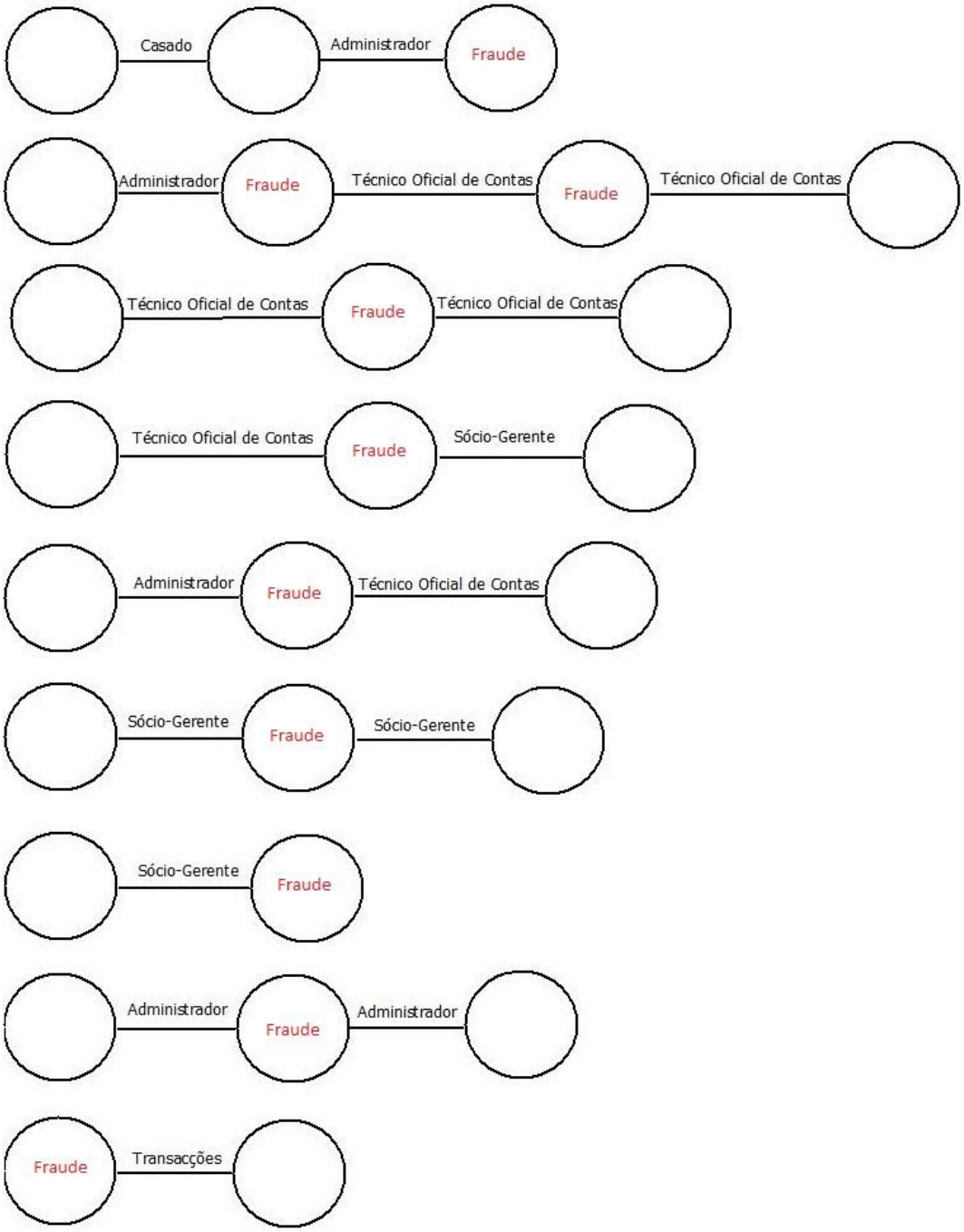
Zhao Y., He, Q. An Unbalanced Dataset Classification Approach Based on Support Vector Machine. [Conferência] // Proceedings of the 6th World Congress on Intelligent Control and Automation. - China : IEEE, 2006. - pp. 10496--10501.

8 Appendixes

8.1 Patterns

Here is shown the list of patterns common to fraudulent entities discovered in the social network analysis phase. All labels are in Portuguese the translation to English could originate some errors. Although that here are the list of terms in Portuguese and the correspondent translation to English: “Casado” is Married,” Técnico Oficial de Contas” means Statutory Auditor, “Administrador” means Administrator, “Sócio-Gerente” is a Partner Manager, “Transações” are Transactions and “Divorciado” means Divorced.




Casado
Administrador
Fraude
Administrador
Fraude
Técnico Oficial de Contas
Fraude
Técnico Oficial de Contas
Técnico Oficial de Contas
Fraude
Técnico Oficial de Contas
Técnico Oficial de Contas
Fraude
Sócio-Gerente
Administrador
Fraude
Técnico Oficial de Contas
Sócio-Gerente
Fraude
Sócio-Gerente
Sócio-Gerente
Fraude
Administrador
Fraude
Administrador
Fraude
Transacções