International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 6 (2014) pp. 116-125
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Characterizing and Preventing Chargebacks in Web Payment Services

Evandro Caldeira[1], **Gabriel Brandão**[1] **and Adriano C. M. Pereira**[2]

[1] Federal Center for Technological Education
of Minas Gerais (CEFET-MG)
Computing Department (DECOM)
Belo Horizonte, MG, Brazil
{*evandrocaldeira, gabriel.brandao.comp*}@gmail.com

[2] Federal University of Minas Gerais (UFMG)
Dept. of Computer Science (DCC)
Belo Horizonte, MG, Brazil
adrianoc@dcc.ufmg.br

Abstract:
The volume of electronic transactions has raised a lot in last years, mainly due to the popularization of e-commerce, such as online retailers. We also observe a significant increase in the number of fraud cases, resulting in billions of dollars losses each year worldwide. Therefore it is important and necessary to developed and apply techniques that can assist in fraud detection, which motivates our research. This work aims to apply and evaluate some computational intelligence techniques to identify fraud in electronic transactions, more specifically in credit card operations. In order to evaluate the techniques, we define a concept of economic efficiency and apply them in an actual dataset of the most popular Brazilian electronic payment service. Our results show good performance in fraud detection, presenting significant gains in comparison to the actual scenario of the company.

Keywords: Fraud Detection; e-Business; Data Mining; Computational Intelligence; Bayesian Networks; Logistic Regression; Neural Networks; Random Forest;

I. Introduction

We have witnessed a significant increase in electronic transactions during the last decades, mainly due to the e-commerce popularization. This popularity, coupled with the large amounts involved and the sensitive information such as Social Security and credit card number, has attracted the criminal's attention. According to Bhatla et al. (2003) [1], the rate at which fraud occurs on the Internet is 12-15 times higher than in the "physical world", so that the sales on the Web represent a significant threat to merchants. According to the Mindware Research Group (2011) [2], it is estimated that the total North America on-line sales revenue loss associated with frauds in 2011 was approximately US$ 3.4 billion, an increase of US$ 700 million compared to 2010.

Considering that the fraud in electronic commerce has been increasing consistently and represents significant losses for business, the prevention and detection of fraud is becoming key to the success of the e-markets. There are a number of challenges in combating fraud, such as the volume of data to be analyzed. Sales through the Internet involve millions of transactions per day. In Brazil, according Mindware Research Group (2011) [2], the number of buyers on-line was 23 million in 2010 and this represented a growth of 35% over the previous year. This huge volume of information makes the manual analysis of each transaction in order to decide whether or not it is fraudulent unfeasible. Moreover, this is clearly a classification problem that is hard to solve, since a fraud transaction does not occur frequently and conditions that characterize fraud may vary significantly between fraudsters and over the time.

Thus, there is a need for novel computational theories and tools to help human beings in this non-trivial classification task. Moreover, many fraud detection problems occur in huge amounts of data. For instance, the credit card company Barclaycard has about 350 million transactions per year just in the United Kingdom. The Royal Bank of Scotland, which has the largest credit card market in Europe, has more than one billion transactions per year [3]. The processing of these datasets looking for fraudulent operations requires fast and efficient algorithms.

In this context, data mining techniques have been relevant in solving these problems since it can deal with a large amount of data. In this work we apply and evaluate computational intelligence techniques to identify fraud in electronic transactions, more specifically in credit card operations. In order to evaluate the techniques, we propose a concept of economic efficiency and apply them in an actual dataset of the most popular Brazilian electronic payment service, which has thousands of transactions per day. We check that imbalanced classes are a factor that directly impact on the prediction quality. Some results present significant gains, up to almost 27%, when compared to the actual fraud detection procedures used in the company. This work is an extended version of a paper published in NWeSP'12 [4]. In this article we have improved the experiments, providing new results and conclusions about them. Moreover, we also provided more details about the case study and explanations about the computational intelligence techniques that we have applied in this actual scenario.

The remainder of this paper is organized as follows. Section II describes some related work. Section III describes the fundamentals of the most relevant computational intelligence techniques used here. Section IV describes our case study, which uses a representative sample of actual data from a large Latin American Internet Service Provider, where we present a dataset overview and some dataset characterizations. Section VI describes our experiments using computational intelligence techniques and the results. Section V describes our Methodology and the process to evaluate the techniques performance. Finally, section VII presents the conclusions and future work.

II. Related Work

There are several researches that develop methods to detect fraud [5, 6, 7] and we can realize that these methodologies can differ significantly due to the peculiarities of each fraud type. However, it can be noticed that the data mining techniques have been widely used in fraud detection regardless of the methodology adopted. This is because these techniques allow the useful information extraction in databases with large volumes of data.

Due to the importance of the fraud detection problem, we may distinguish several works that discuss the subject. Thomas et al. (2004) [8] propose a very simple decision tree that is used to identify general fraud classes. They also propose a first step towards fraud taxonomy. Vasiu and Vasiu (2004) [9] propose a taxonomy for computer fraud and, to build it, employ a five-phase methodology. According to the authors, the taxonomy presented was prepared from a fraud preventing perspective and may be used in various ways. Chau et al. (2006) [10] propose a methodology called *2-Level Fraud Spotting (2LFS)* to model the techniques that fraudsters often use to carry out fraudulent activities and to detect offenders preventively. This methodology is used to characterize the auction users on-line as honest, dishonest, and accomplices.

There are several studies that develop methods for fraud detection [5, 7], and their analysis show that, as a consequence of frauds specificities, these methodologies may differ significantly as a function of the particularities of each fraud type. However, we can notice that the data mining techniques have been widely used in fraud detection regardless of the methodology adopted. This is because these techniques support the extraction of useful information in databases with large volumes of data. Phua et al. (2005) [11] perform a comprehensive study of numerous works related to fraud detection using data mining and present those methods and techniques as well as their limitations. According to the authors, there are three approaches on which these algorithms are based: supervised strategy, unsupervised strategy and hybrid strategy.

In the supervised strategy [12], learning algorithms examine all transactions, previously labeled, to mathematically determine the profile of a fraudulent transaction and to estimate their risk. In the unsupervised strategy [13], the methods do not require prior knowledge of the fraudulent and non-fraudulent transactions. On the other hand, they detect changes in behavior or unusual transactions. Another brief overview of different [14] fraud types also shows related works that use Genetic Programming to make fraud prediction.

In unsupervised strategy with unlabeled data, unsupervised methods do not require prior knowledge of fraudulent and not fraudulent transactions. On the other hand, changes in behavior are detected or unusual transactions are identified. Examples of these techniques are Clustering and Anomaly Detection. Netmap [15] describes how the clustering algorithm is used to form well-connected data groups and how it led to the capture of the real insurance fraudsters. Bolton and Hand [13] proposed a fraud detection in credit card using anomalies detecting techniques in transactions. Abnormal behaviors are identified in spending and how often they occur is used to determine which cases may be fraud.

In the hybrid approach (supervised and unsupervised) there are researches using data labeled with supervised and unsupervised algorithms to detect fraud in insurance and telecommunications. Unsupervised approaches have been used to segment data into groups to be used in supervised approaches. Williams and Huang [16] apply a three step process: k-means for detecting groups, C4.5 for decision making, and statistical summaries and visualization tools to evaluate the rule.

The Artificial Immune Systems (AIS) was used by Gadi et al. [17] and Wong et al. [18], providing good results to detect fraud in credit card transactions. [17] also uses Neural Nets, Bayesian Nets, Naive Bayes and Decision Trees. SVM is another technique that can be used in the credit card context. Hens and Tiwari [19] used SVM to create a credit scoring model to reduce the risk on to the news applicant. In order to reduce the computational time the authors made a stratified sampling of their data.

These related works motivate us to study and choose some techniques to apply to our fraud detection scenario and evaluate them using a real dataset.

III. Fundamentals

This section describes the techniques we apply and evaluate in this work: Bayesian networks (BN) (Section III-A), logistic regression (LR) (Section III-B), Radial Basis Function (RBF) neural networks (NN) (Section III-C), random forest (RF) (Section III-D), Support Vector Machines (SVM) ande Sequential Minimal Optimization (SMO).

RBF is a combination of radially symmetric nonlinear basis functions. In a supervised training the RBF creates a discriminant function for each class [20]. SVM is a Kernel-based machine learning technique that constructs hyperplanes to use in regression or classification [21]. SMO takes, heuristically, two points at each step so that the problem could be solved analytically [21].

A. Bayesian Networks

Bayesian Networks (BN) are directed acyclic graphs that represent dependencies between the variables of a probabilistic model, where each node in the graph represents a random variable and the arcs represents the relationships between these variables [22], as showed by Figure 1, where the event A depends directly of event D that depends directly of event B and H, and so on. And *e* is an independent event.

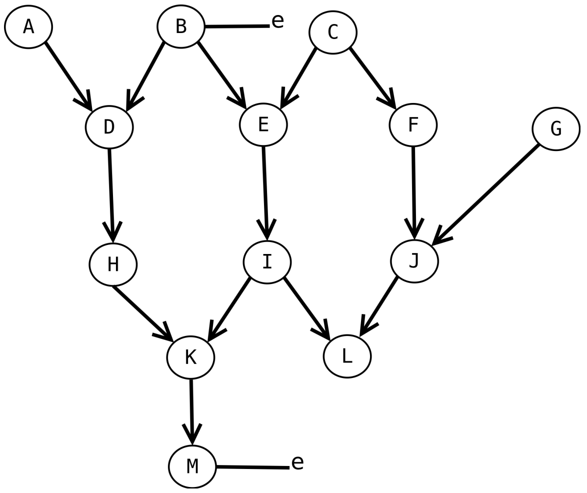


Figure. 1: Bayesian Network - Description.

The mathematical definition for BN is derived of Bayes theorem, which shows that conditional probability of an event A_i given an event B, can be calculated by Equation 1.

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \quad (1)$$

where $P(A_i|B)$ is the probability of A when B occurs.
In fraud detection problem the BN is unknown, therefore to build the BN graph it is need to learn it from the data. From the BN graph, we can calculate the set of dependent variables to happen a fraud (conditional probability), using Equation 1. Before calculating the conditional probability, we can find the probability of fraud applying Equation 2 [23].

$$P(x_i, ..., x_n) = \prod_{i=0}^{n} P(x_i|Parents(X_i)) \quad (2)$$

, where $Parents(X_i)$ are determined by a graph as showed by Figure 1.

B. Logistic Regression

Logistic Regression (LR) is a statistical technique that produces, from set of explanatory variables, a model that can predict values taken by a categorical dependent variable. Thus, a regression model is used to calculate the probability of an event, through the *link* function described by the following Equation:

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...\beta_i x_i)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...\beta_i x_i)}}, \quad (3)$$

where $\pi(x)$ is the probability of success when the value of the predictive variable is x. β_0 is a constant used for adjustment and β_i are the coefficients of the predictive variables [24].

In order understand LR, it is important to explain the concept of *Generalized Linear Models* (GLM). This consists of three components [25]:

- A random component, which contains the probability distribution of the dependent variable (Y).
- A systematic component, which corresponds to a linear function between the independent variables.
- A *link* function, that is responsible for describing the mathematical relationship between the systematic component and random component.

The binary LR model is a special case of the GLM model with the *logit* function. This function is used to get the estimation of coefficients [26]. Then, we apply these coefficients in Equation 3 that result in our fraud probability.

C. Neural Networks

A Neural Network (NN) is an interconnected assembly of simple processing elements, units or nodes, whose functionality is loosely based on the animal neuron [27]. The processing ability of the network is stored in the inter-unit connection strengths, or weights, obtained by a process of adaptation to, or learning from, a set of training patterns.
Generically, the processing in a neuron consists of a linear combination of entries (x_j), which can be described by Equation 4:

$$\begin{aligned} net &= w_1 * x_1 + w_2 * x_2 + ... + w_D * x_D \\ &= \sum_{j=1}^{D} w_j x_j = \underline{w}^T * \underline{x} \end{aligned} \quad (4)$$

, where w_j is a weight associated with the input (x_j). This weight shows the intensity wherewith a particular input influences the output value. The calculated value (net) is applied in an activation function that can be Linear, Step, Ramp, Sigmoid, Hyperbolic Tangent or Gaussian. [28]
The NN model used was MultiLayer Perceptron (MLP), which has the ability to classify non-linearly separable regions [29], appropriate for our fraud detection approach.
The training was done using the Levenberg-Marquardt algorithm [30], because it is fast and can achieve good results. We perform a set of experiments to determine the best NN configuration, that is, a network with two layers: the first (hidden layer) containing ten neurons and the second (output layer) containing one neuron.

D. Random Forest

The Random Forest (RF) algorithm was proposed by Breiman [31] based on the use of trees to product classification. Breiman's definition to algorithm is: "A RF is a classifier consisting of a collection of tree-structured classifiers $h(x, \theta_k), k = 1, ...$ where the θ_k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x".
The classifier quality or performance can be measured by a high value of probability $\mathcal{P}(h(X) = Y)$. The vector X represents the variables of the problem and Y is the response.

Given a observed dataset

$$((x_{1,1},...x_{1,n}),(x_{2,1},...x_{2,n}),...,(x_{k,1},...x_{k,n}))=D$$

and let B be the number of trees and m the number of features. The Algorithm 1 describes the RF.

Algorithm 1 Random Forest Algorithm

```
for $N = 0, .., B$ do
    $D_i \leftarrow$ Bootstrap sample from $D$
    $T_i \leftarrow$ Contruct tree using $D_i$
    for $node = 1, .., No.Nodes$ do
        $node_i \leftarrow$ choose random subset $m$ of all features.
    end for
end for
$X \leftarrow$ take the majority vote for all trees
```

Other ensemble methods can be seen in [32].

IV. Case Study: Characterization

This section presents our case study, where we apply our techniques to characterize electronic transactions. These characterizations allow a better understanding of the problem and provide a way to select some attributes from the dataset, which will be used in the fraud detection process [33].
Analyzing the transaction values we decided to take off those ones with values greater than US$750[1], which can be defined as outliers. These values, which correspond to less than 1% of the whole dataset, generate noise on the sample that could let to a misunderstanding.

	Valid transactions	Chargeback transaction
Average value (US$)	40.92	99.57
Standard deviation (US$)	74.35	133.33
Median (US$)	15.14	44.20
Coefficient Of Variation	1.82	1.34

Table 1: Basic Statistics of April 2011

Table 1 presents the basic statistics of our dataset. It shows that average value of chargeback transactions is greater than valid transactions. This analysis motivates us to select this attribute as an important element to help classifying chargeback transactions.
Figure 2 shows the cumulative distribution function (CDF), where we can observe that valid transactions with values lower than US$50.00 correspond to 80%, and 55% for chargeback ones. Thus, we can see that in general valid transactions present lower values than chargeback ones.
Figure 3 presents the average transaction value per hour for chargeback and valid transactions, where we can note that the average values of chargeback and valid transactions have a different behavior over the time of the day. The first one varies during the hours of the day, while the other one is almost constant. As we have already observed only a small change in transaction values, it is important to note here that chargeback presents different behavior during the time of the day, which can be considered important to help in fraud classification process. In the figure the thin lines represent the standard deviation for each column value.

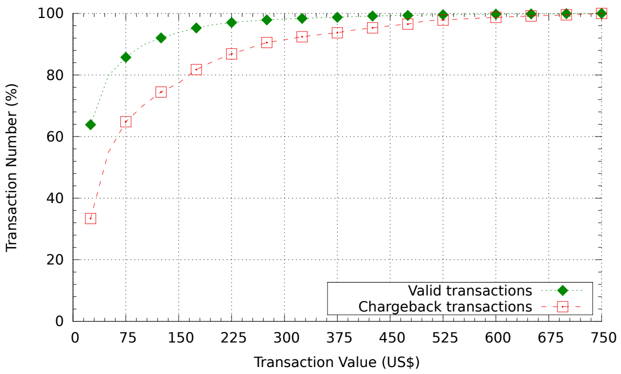


Figure. 2: Volume of transactions by value category

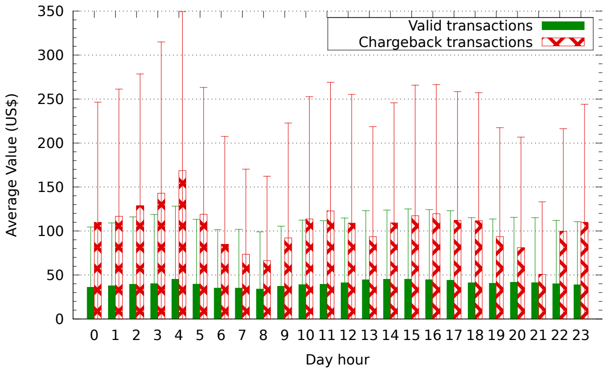


Figure. 3: Average Transaction Value by the time of the day

Figure 4 shows the influence of the buyer age in the fraud occurrence. It shows that users between 25 and 35 years old buy more than other age ranges, but the relative number of chargebacks is greater than valid transactions over 40.

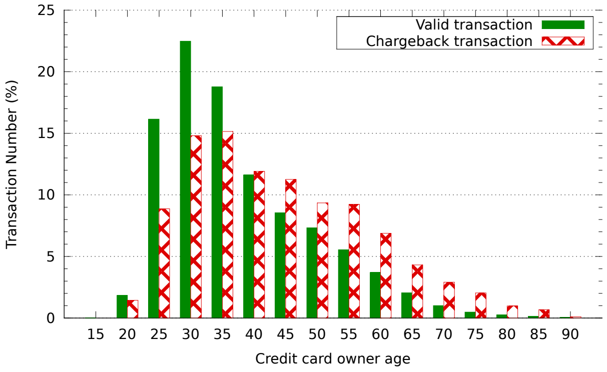


Figure. 4: Number of transactions by age

Figure 5 shows the risk of chargeback related with the credit card operator. A greater risk can be explained by the different rules that each credit card operator specifies.

[1] We use the R$2.00 as dollar quotation

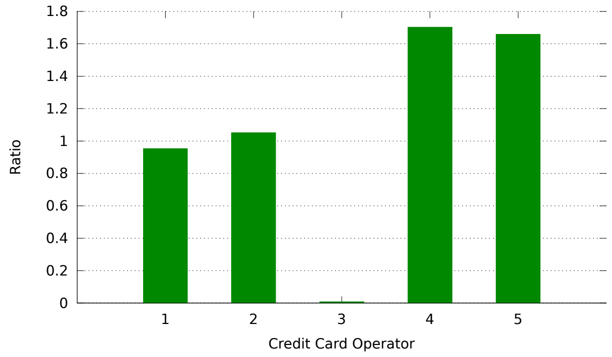


Figure. 5: Risk of chargeback per card operator

A high risk means that operators have a relation

$$\frac{Chargeback\_transaction}{Valid\_transactions}$$

greater than other ones.

These are some examples of characterizations that we have performed in our work, which can contribute with the efficiency of the computational intelligence techniques.

V. Methodology

We have in our dataset more than 30 attributes but not all of them are suitable to apply to fraud detection. Some attributes, such as user id, credit card number and phone number, were excluded to guarantee the model generality. The characterization presented in section IV has great importance to a better understanding the problem, the data related to it and thus exclude outliers values that could disturb the model. We also use the Information Gain, Chi-square (χ^2) [34] and Correlation-based Feature Subset Selection [35] to find the most relevant attributes to fraud detection. We conclude, after this process, that the following attributes are the best set to use in fraud detection:

- **Value**: numeric attribute with the transaction value.
- **Hour** : numeric attribute with the transaction hour.
- **Day of the Week:** numeric attribute representing the day of the week.
- **Flag CPF**[2]: binary attribute assigned to 1 when user's CPF differs from the CPF of the Credit Card Owner.
- **DDD**: Long Distance Call Number.
- **Flag Dispute**:binary attribute assigned to 1 when a registered user starts a dispute with a seller due to any problem with the transaction.
- **Flag Postal Code**: binary attribute assigned to 1 when the Postal Code of the Credit Card Owner differs from the delivery address.
- **Seller Main Category**: numeric attribute with the main category of the seller.
- **Credit Card Owners Age**: numeric attribute with the age in years of the credit card owner.
- **Status Serasa**[3]: numeric attribute in response to a Serasa request about the CPF of the Credit Card owner.
- **Number of Installments**: numeric attribute with the number of transaction installments.
- **Credit Card Operator**: numeric attribute to the Credit Card Operator.
- **Number of distinct items**: numeric attribute with the number of distinct items of the transaction.
- **Flag Registered User**: binary attribute assigned to 1 when this is a registered user.

Beyond the attribute selection we also have the problem of imbalanced dataset with the minor class with 0.58% of the samples. To minimize this, we create two different datasets. Real Set (RS) is a dataset composed by all transactions in weeks 1 to 3 for training and the remained weeks of the month for test. The Modified Set (MS) was formed taking all chargebacks in weeks 1 to 3, but only 10% of the valid transactions. This increases the chargebacks for all transactions ratio to 5.98% and was random to keep the generality. The test set is the same for both datasets and has 0.63% of chargebacks.

In our problem, Precision and Recall is not the best measures to evaluate the performance of the techniques. Thus, in conjunction with ours Fraud Specialists we proposed the concept of Economic Efficiency, as presented by Equation 5 and Equation 6. The Gain (G) represents the financial value of the true positive transactions, rate (r) is a percentual that the company gains in a successful transaction, the Lost (L) is the financial value of false negative transactions and **n** is the number of transactions that we have.

In Equation 5 EE_{NP} is the Economic Efficiency with **N**o **P**enalty

$$EE_{np} = \sum_{j=1}^{n} G_j * r - L_j * (1 - r) \qquad (5)$$

,

and on Equation 6 EE_{WP} is the Economic Efficiency **W**ith **P**enalty.

$$EE_{wp} = \sum_{j=1}^{n} G_j * r - (L_j * (1 - r) + NG_j * p) \qquad (6)$$

, the **p**enalty ratio was defined by the Fraud Specialists to be 1%. **NG** (No Gain) represents the transaction financial value of the misclassification. This is an improvement from a previous work [4].

In this equation there is a **p**enalty ratio to each misclassification of the algorithms. The approach chosen here is different than Gadi et al. [17] that use an average loss of $100 for every undetected fraud. The major difference is that here we consider the loss as a percentual value of the transaction. Proceeding this way the behavior of each one of the algorithm

[2] "Natural Persons Register", is a number provided by the Brazilian revenue agency to identify a taxpayers

[3] Serasa Experian is leader in analysis and information for credit decisions and business support.

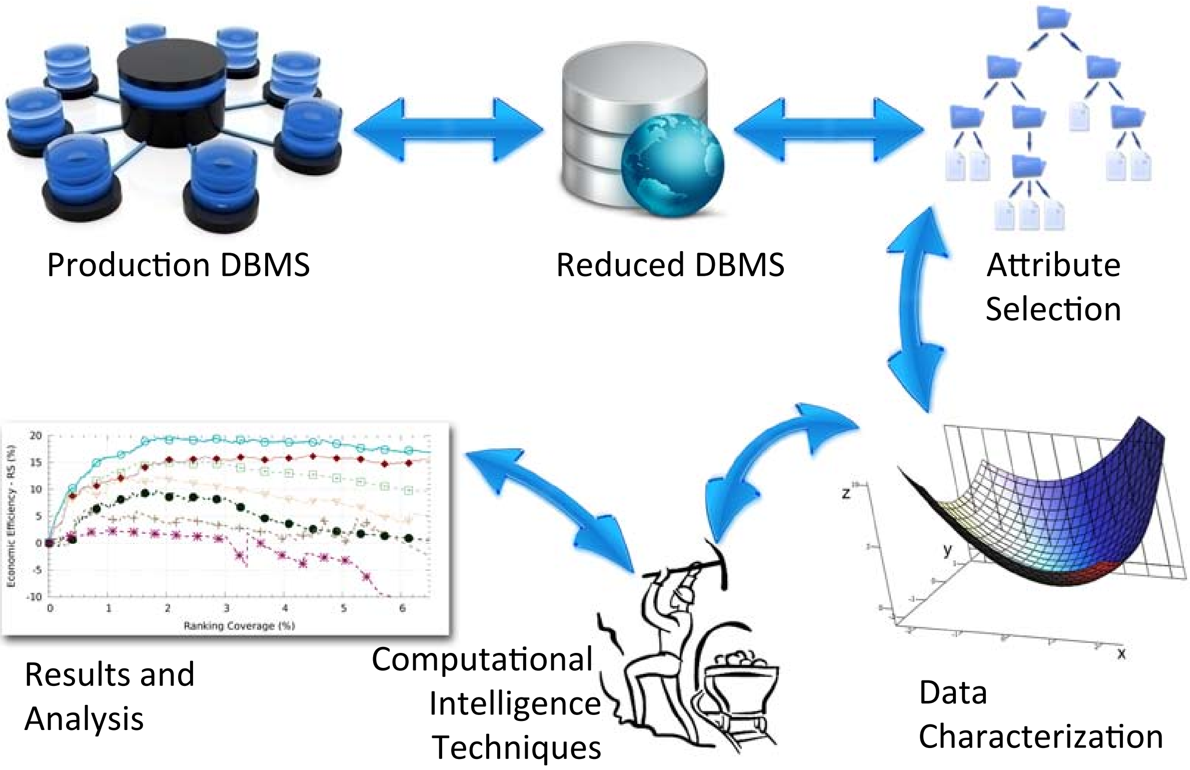


Figure. 6: Model describing the main methodology steps

can be measured in terms of the real gain or loss in a real case study, as the presented in the paper.

We apply these equations in a ranking sorted by fraud probability assigned to each transaction. At the top of it are those with the highest fraud probability. Applying these formulas in many ranking ranges provides us a position that maximizes the profit for a given algorithm.

Unsorted			
Id	**F.**	**FP**	**NPF**
1	0	0,20	0,80
2	0	0,40	0,60
3	1	0,45	0,55
4	1	0,47	0,53
5	0	0,60	0,40
6	0	0,40	0,60
7	0	0,30	0,70
8	1	0,99	0,01
9	0	0,25	0,75
10	1	0,95	0,05

(a)

Sorted			
Id	**F.**	**FP**	**NPF**
8	1	0,99	0,01
10	1	0,95	0,05
5	0	0,60	0,40
4	1	0,47	0,53
3	1	0,45	0,55
2	0	0,40	0,60
6	0	0,40	0,60
7	0	0,30	0,70
9	0	0,25	0,75
1	0	0,20	0,80

(b)

Table 2: Table with the ranking to fraud detection. In (a) are the unsorted results and in (b) are the results ranked by the fraud prabability. "Id" is a transaction identifier, "F" is a label that indicates if the transaction is a fraud or not, "FP" is the Fraud probability and "NFP" is the Non-fraud Probability

Table 2 exemplifies proposed ranking for a hypothetical set of 10 transactions. In Table 2a the registers are disposed in same order in which they were provided to an algorithm. In order to facilitate the visualization the fraud transactions are colored with red background and in Table 2b the registers are sorted by the Fraud Probability (FP). In this sample the transaction with Id 3, background in blue, is the point in the ranking with the major gain for the technique that produced this classification.

Equation 7 gives a relative gain where 100% represents the maximum possible gain and 0% is the actual scenario without the use of any technique. The EE_{Max} indicator is the maximum gain that the company could have when all frauds are detected and valid transactions are accepted. We use this equation to compare all the techniques.

$$EE = \frac{EE' - EE_{Real}}{EE_{Max} - EE_{Real}} \tag{7}$$

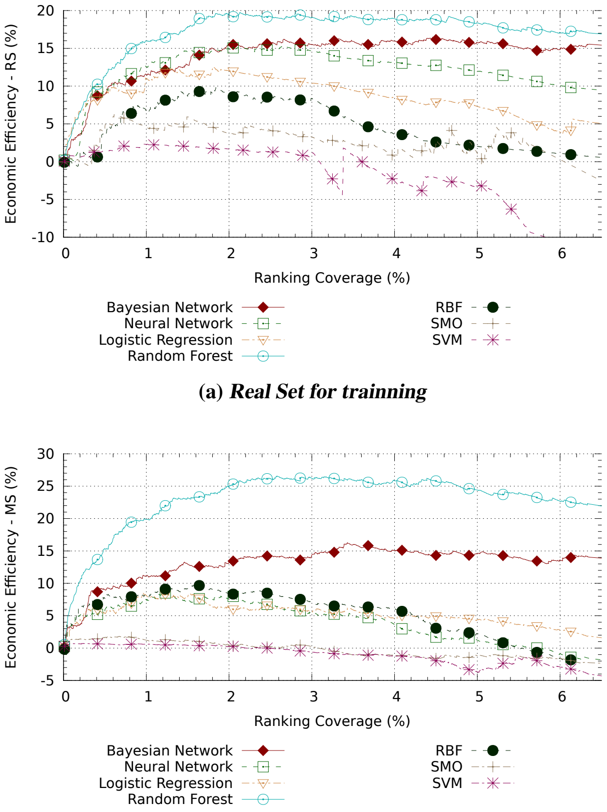
, and EE′ is one of these $\mathrm{EE_{NP}}$ or $\mathrm{EE_{WP}}$.

Figure 6 describes the main steps of our experimental methodology, from data extraction to result analysis of the computation intelligence techniques. Due to the impossibility and the risk of access the database of the system in production, the proposed solution uses a reduced database of the original one. The knowledge of the problem domain, acquired previously, in conjunction with the assistance of the application designer make possible to create this reduced database to be used by the data mining and fraud detection team.

Many companies also have restricted security rules to access their production databases. In these cases the access of a third part team is forbidden to avoid the risk to the operation of the Web service. The attribute selection is made in conjunction with the characterization process. This step is performed to enforce the knowledge of the application's domain and the data object of study. The next step consists of the execution of the data mining and other techniques and the evaluation of results. This is not a one way process, since the results from step i can provide useful information to modify the step $i - 1$, providing insights and new elements to the fraud detection process.

VI. Case Study: Chargeback Prevention - Results

Figure 7a shows the results with the RS. LR and RF have similar results until 0,4% of the ranking, then RF outperforms all techniques. We can observe that SVM and SMO don't present good results.



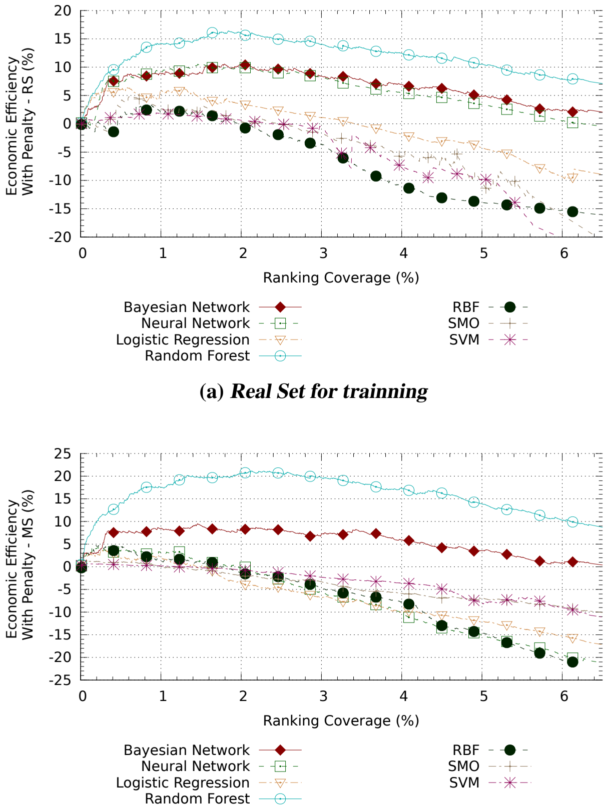
(a) ***Real Set for trainning***

(b) ***Modified Set for trainning***

Figure. 7: Economic Efficiency with No Penalty

The penalty case can be seen in Figure 8. Figure 8a presents the results to the Real Set. The Random Forest algorithm is the best choice all over the ranking. The Bayesian Network and Neural Network having very similar curves up to 6% of the ranking. Logistic Regression stay close to the previous techniques up to 0.4% of the ranking and after 1.2% it start to decrease its performance moving away from the previous techniques. Just at the beginning of the ranking RBF got negative results and at 0.4% it goes to the positive area but did not maintain its growth. SVM and SMO go poor results in a similar way as RBF.

Figure 8b shows the results with penalty to the Modified Set. The Random Forest has a fast advantage comparison with the other techniques. It stays with approximately 20% of EE_{MS} from 1.1% to 3.2% of the ranking. The Bayesian Network was oscilating around 8% of EE_{MS} from 0.5% to 2.5% of the ranking. The RBF and Neural Network have similar behavior in their curves. SVM and SMO did not show significant results.



(a) ***Real Set for trainning***

(b) ***Modified Set for trainning***

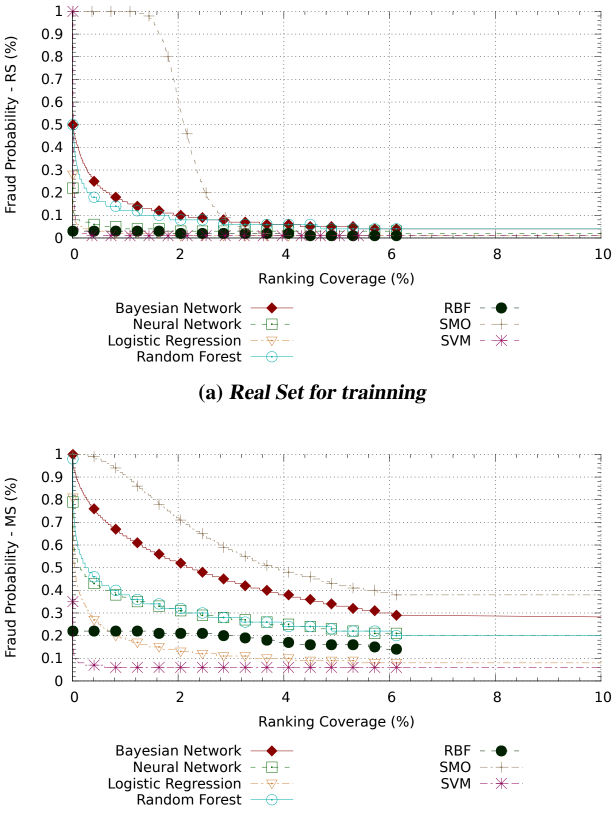
Figure. 8: Economic Efficiency with Penalty

Figure 9 can be used to analyze the fraud probability over the ranking. To the case of EE_{RS}, see Figure 9a, the SMO set the highest probalistic values up to 1% of the ranking and after 3% the values are more close to the values of the other techniques. Despite the high probalistic value the SMO presented the worst results together with the SVM. The Bayesian Network e the second with the higher probalistic value to fraud followed by the Random Forest. These two techniques start with 0.5% but Random Forest decrease faster than Bayes. The RBF did not show high values in these analyses and Neural Network and Logistic Regression have a faster decay of probability value compared to the others.

To the case EE_{MS} (Figure 9b) the SMO also show the highest values despite the quality of the final result. The Bayesian Network also has higher probalistic value but present better result than SMO. The Random Forest also got higher probalistic values like Bayes but here the decay of value was higher. The Neural Network presented a very similar behavior, in terms of probalistic value, as the Random Forest. The Logistic Regression also has a fast decay and going to stay below the RBF after 1% of the ranking.

The RBF and BN techniques presented similar results, around 9% and 16%, respectively, in both experiments; therefore we can conclude that the MS dataset has quite no effect on its performance. The best result of NN was 15.33% of EE_{RS}, which was the opposite of RF.

Table 3 summarizes the results presented before. Table 3a



(a) ***Real Set for trainning***

(b) ***Modified Set for trainning***

Figure. 9: Fraud Probability over the ranking

shows the results with No Penalty. RF shows the best results with $EE^{NP}{}_{RS}$ of 19.74% and $EE^{NP}{}_{MS}$ of 26.55%. On the other hand, SVM presents an EE_{RS} of 2.36% and an EE_{MS} of 0.71%, thus there is quite no improvements compared to the actual scenario. BN has the highest $Recall_{RS}$ of 28.92% and RF presents the highest $Recall_{MS}$ of 29.13%. Table 3b presents the results with Penalty. In these scenario Random Forest also was the best with 16.41% of $EE^{WP}{}_{RS}$ and 21.14% of $EE^{WP}{}_{MS}$.

VII. Conclusion

We presented in this work an analysis of computational intelligence techniques applied to predict fraud in Internet transactions, more specifically in credit card operations. We evaluated Bayesian Networks (BN), Logistic Regression (LR), Radial Basis Function (RBF), Neural Network (NN), Random Forest (RF), Support Vector Machines (SVM) and Sequential Minimal Optimization (SMO). We presented the concept of Economic Efficiency (EE) that is a more confident measure to a Fraud Specialist analysis than Precision and Recall. This concept was presented in two variations, the first, EE^{NP} is our cost function without penalty to any misclassification, the second, EE^{WP} address a penalty ration of 1% to any misclassification of the algorithm. We also proposed an alternative to the problem of imbalanced classes, evaluating a modified dataset as training set that presents 10 times more transaction chargebacks. The results could be used by e-commerce companies to improve the efficiency of their fraud detection processes.

We apply our experiments in an actual dataset containing thousands of transactions per day of the most popular Brazilian electronic payment service, called *PagSeguro*. The work shows that imbalanced classes were a factor that impacts on the prediction quality. The achieved results present significant gains when compared to actual scenario of the company that adopts some fraud detection procedures. RF reaches the best results with No Penalty of $EE^{NP}{}_{RS}$ of 19.74% and $EE^{NP}{}_{MS}$ of 26.56%. With the Penalty equation Random Forest also achieved the best results with $EE^{WP}{}_{RS}$ of 16.41% and $EE^{WP}{}_{MS}$ of 21.14%.

The training set modification did not have a great impact on BN and RBF, showing very similar results in both scenarios. One explanation about NN behavior could be that our training set was not enough to identify regular transactions causing misclassifications and losses. As was expected all techniques get worst results with the use of penalty and a loss in Recall. Some techniques such as BN, LR and RBF lose about 6 percentual points with the use of penalty showing that in more rigid conditions some techniques are better than the others. SMO and SVM had not presented good results and with the disadvantage to be much more time consuming than the other approaches although the penalty has a smaller impact on its performance. The use of actual data, instead of synthetic samples, also shows the applicability of this work in similar Web application scenarios.

One of the challenges of this research is the nature of the data, since they are very imbalanced with the minor class with less than 1% and alternatives to deal with this problem should be investigated in more detail. The use of a cost sensitive evaluation [36, 37] is one possible alternative. Beyond this, we could segment the dataset, by price or age, for example, and analyze the performance of the techniques in these different scenarios. We can also apply some undersampling techniques to remove redundant noise and examples or oversampling to introduce instances of the minority class [38].

Acknowledgments

This research was sponsored by UOL (www.uol.com.br), through its UOL Bolsa Pesquisa Program, Processes Number 20120217214100. We also was supported by the Brazilian National Institute of Science and Technology for the Web (CNPq grant no. 573871/2008-6), CAPES, CNPq, Finep, and Fapemig.

References

[1] V. P. Tej Paul Bhatla and A. Dua, "Understanding Credit Card Frauds," Tata Consultancy Services, Tech. Rep., Jan. 2003. [Online]. Available: http://www.popcenter.org/problems/credit_card_fraud/PDFs/Bhatla.pdf

[2] C. Mindware Research Group, "2011 Online Fraud Report," CyberSource, Tech. Rep., 2011. [Online]. Available: http://www.cybersource.com

[3] R. J. Bolton and H. David J., "Statistical fraud detection: A review," pp. 235–255, 2002.

Tech.		EE (%)	Prec. (%)	Recall (%)	Rank. (%)
BN	RS$_{NP}$	16.46	4.12	**28.92**	4.46
	MS$_{NP}$	16.24	4.35	23.48	3.43
NN	RS$_{NP}$	15.33	3.56	14.87	2.65
	MS$_{NP}$	8.55	4.25	8.01	1.20
LR	RS$_{NP}$	12.53	4.06	11.61	1.82
	MS$_{NP}$	8.51	3.89	7.63	1.24
RF	RS$_{NP}$	**19.74**	6.62	22.19	2.13
	MS$_{NP}$	**26.55**	7.16	**29.13**	2.58
RBF	RS$_{NP}$	9.89	3.5	10.15	1.84
	MS$_{NP}$	9.72	3.77	9.68	1.63
SMO	RS$_{NP}$	6.21	6.54	7.08	0.61
	MS$_{NP}$	1.85	5.15	6.13	0.75
SVM	RS$_{NP}$	2.36	5.31	8.85	0.94
	MS$_{NP}$	0.71	6.18	2.19	0.22

(a) *Results with No Penalty*

Tech.		EE (%)	Prec. (%)	Recall (%)	Rank. (%)
BN	RS$_{WP}$	10.55	5.49	15.47	1.79
	MS$_{WP}$	9.50	5.59	12.81	1.46
NN	RS$_{WP}$	10.53	4.52	10.80	1.52
	MS$_{WP}$	4.25	6.38	2.19	0.22
LR	RS$_{WP}$	6.40	4.60	9.21	1.27
	MS$_{WP}$	3.89	5.02	2.96	0.37
RF	RS$_{WP}$	**16.41**	7.36	**21.29**	1.84
	MS$_{WP}$	**21.14**	7.91	**26.44**	2.12
RBF	RS$_{WP}$	3.30	3.94	5.48	0.88
	MS$_{WP}$	4.76	7.26	2.10	0.18
SMO	RS$_{WP}$	5.01	6.54	7.08	0.61
	MS$_{WP}$	1.28	14.29	2.14	0.10
SVM	RS$_{WP}$	1.91	5.31	8.85	0.94
	MS$_{WP}$	0.67	12.31	1.71	0.09

(b) *Results With Penalty*

Table 3: Summary results for all techniques in both cases, EE^{NP} and EE^{WP}

[4] E. Caldeira, G. Brandão, and A. C. M. Pereira, "Characterizing and preventing chargebacks in next generation web payments services." in *CASoN 2012 - Fourth International Conference on Computational Aspects of Social Networks (part of the Eighth International Conference on Next Generation Web Services Practices (NWeSP 2012))*. IEEE, 2012, pp. 333–338. [Online]. Available: http://dx.doi.org/10.1109/CASoN.2012.6412424

[5] T. Fawcett and F. Provost, "Adaptive fraud detection," *Data Mining and Knowledge Discovery*, vol. 1, pp. 291–316, 1997. [Online]. Available: http://www.bibsonomy.org/bibtex/2c69214e294a91a9a6216b58f438c8db8/jamesh

[6] R. Maranzato, A. Pereira, M. Neubert, and A. P. do Lago, "Fraud detection in reputation systems in e-markets using logistic regression and stepwise optimization," *SIGAPP Appl. Comput. Rev.*, vol. 11, pp. 14–26, June 2010. [Online]. Available: http://doi.acm.org/10.1145/1869687.1869689

[7] E. L. Barse, H. Kvarnström, and E. Jonsson, "Synthesizing Test Data for Fraud Detection Systems," in *19th Annual Computer Security Applications Conference*, ser. ACSAC '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 384–. [Online]. Available: http://portal.acm.org/citation.cfm?id=956415.956464

[8] B. Thomas, J. Clergue, A. Schaad, and M. Dacier, "A comparison of conventional and online fraud," in *CRIS'04, 2nd International Conference on Critical Infrastructures, October 25-27, 2004 - Grenoble, France*, 10 2004.

[9] L. Vasiu and I. Vasiu, "Dissecting Computer Fraud: From Definitional Issues to a Taxonomy," in *37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 7 - Volume 7*, ser. HICSS '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 70 170.3–.

[10] D. H. Chau, P. Shashank, and C. Faloutsos, "Detecting fraudulent personalities in networks of online auctioneers," in *In Proc. ECML/PKDD*, 2006, pp. 103–114.

[11] C. Phua, V. C. S. Lee, K. Smith-Miles, and R. W. Gayler, "A Comprehensive Survey of Data Mining-based Fraud Detection Research," *CoRR*, vol. abs/1009.6119, 2010. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1009.html#abs-1009-6119

[12] S. Maes, K. Tuyls, B. Vanschoenwinkel, and B. Manderick, "Credit Card Fraud Detection Using Bayesian and Neural Networks," in *In: Maciunas RJ, editor. Interactive image-guided neurosurgery. American Association Neurological Surgeons*, 1993, pp. 261–270.

[13] R. J. Bolton and D. J. Hand, "Unsupervised Profiling Methods for Fraud Detection," *Statistical Science*, vol. 17, no. 3, pp. 235–255, 2002.

[14] K. Chaudhary, J. Yadav, and B. Mallick, "A review of fraud detection techniques: Credit card," *International Journal of Computer Applications*, vol. 4, no. 1, pp. 39–44, May 2012, published by Foundation of Computer Science, New York, USA.

[15] Netmap, "Fraud and Crime Example Brochure," Tech. Rep., 2004.

[16] G. J. Williams and Z. Huang, "Mining the Knowledge Mine: The Hot Spots Methodology for Mining Large Real World Databases," *Lecture Notes in Computer Science*, vol. 1342, pp. 340–348, 1997.

[17] M. F. A. Gadi, X. Wang, and A. P. do Lago, "Credit Card Fraud Detection with Artificial Immune System," in *ICARIS*, ser. Lecture Notes in Computer Science, P. J. Bentley, D. Lee, and S. Jung, Eds., vol. 5132. Springer, 2008, pp. 119–131. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-85072-4_11

[18] N. Wong, P. Ray, G. Stephens, and L. Lewis, "Artificial immune systems for the detection of credit card fraud: an architecture, prototype and preliminary results," *Inf. Syst. J*, vol. 22, no. 1, pp. 53–76, 2012. [Online]. Available: http://dx.doi.org/10.1111/j.1365-2575.2011.00369.x

[19] A. Hens and M. Tiwari, "Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method," *Expert Systems with Applications*, 2012.

[20] A. R. Webb and K. D. Copsey, *Statistical Pattern Recognition* , 3rd ed. Wiley, Nov. 2011.

[21] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, March 2000.

[22] S. Maes, karl Tuyls, B. Vanschoenwinkel, and B. Manderick, "Credit card fraud detection using bayesian and neural networks," *Vrije Universiteir Brussel*, 2001.

[23] G. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, no. 4, pp. 309–347, 1992.

[24] D. W. Hosmer, *Applied Logistic Regression*, 2nd ed. New York: Wiley, 2000.

[25] A. J. Dobson, *An Introduction to Generalized Linear Models*. London:Chapman and Hall, 1990.

[26] W. N. Venables, D. M. Smith, and t. R. D. C. Team, "An introduction to R," www.cran.r-project.org, 2009, acesso em maio de 2010.

[27] K. Gurney, *An introduction to neural networks*. CRC Press, 1997, .ISBN: 978-1857285031.

[28] A. P. Engelbrecht, *Computational Intelligence: An Introduction*, 2nd ed. Wiley, 2007.

[29] A. Konar, *Computational Intelligence: Principles, Techniques and Applications*. Springer-Verlag New York, 2005.

[30] M. I. Lourakis, "A brief description of the Levenberg-Marquardt algorithm implemented by levmar," *Institute of Computer Science, Foundation for Research and Technology*, vol. 11, 2005.

[31] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[32] A. B. Brahim, R. Khanchel, and M. Limam, "Robust ensemble based algorithms for multi-source data classification."

[33] S. Burke, "Understanding the Structure of Scientific Data," *LCGC Online Supplement*, pp. 3–8, 1997. [Online]. Available: http://chromatographyonline.findanalytichem.com/lcgc/article/articleDetail.jsp?id=4489

[34] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," in *Proceedings of the Fourteenth International Conference on Machine Learning*, ser. ICML '97. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, pp. 412–420. [Online]. Available: http://portal.acm.org/citation.cfm?id=657137

[35] M. A. Hall and L. A. Smith, "Practical feature subset selection for machine learning," *Springer*, 1998. [Online]. Available: http://researchcommons.waikato.ac.nz/handle/10289/1512

[36] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," *Discovery*, no. 1999, pp. 1–12, 2004.

[37] Y. Xie, X. Li, E. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5445–5449, 2009.

[38] J. Luengo, A. Fernández, S. García, and F. Herrera, "Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling." *Soft Comput.*, vol. 15, no. 10, pp. 1909–1936, 2011.

Author Biographies

Adriano C. Machado Pereira is an Adjunct Professor in Computer Science Department at Federal University of Minas Gerais (DCC / UFMG), Brazil. He received his bachelor degree in Computer Science at UFMG in 2000, his MSc. in 2002, and his Ph.D. in 2007. He also had performed a Post-Doc research in electronic markets in 2008-2009. His research interests include e-Business, e-Commerce, Workload Characterization, Distributed Systems, Web 2.0, Social Networks, Performance of Computer Systems, Web Technologies, and Business Intelligence. He is also a member of the Brazilian National Institute of Science and Technology for the Web - INWEB (`www.inweb.org.br`).

Evandro Caldeira is an Undergraduated Student in Computer Engineering at Computer Department (DECOM) of Federal Center for Technological Education of Minas Gerais (CEFET-MG). His research interests include e-Business, Data Mining, Distributed Systems, Artificial Intelligence and Big Data.

Gabriel Brandão is an undergraduate student of Computer Engineering at Computer Department (DECOM) of Federal Center for Technological Education of Minas Gerais (CEFET-MG). His research interests include Data Characterization, e-Business, e-Commerce, Big Data, Data Mining and Artificial Intelligence.