Fraud Analysis and Prevention in e-Commerce Transactions

Evandro Caldeira
Federal Center of Technological Education of Minas Gerais (CEFET-MG)
Computing Department
Belo Horizonte, MG, Brazil
Email: evandrocaldeira@gmail.com

Gabriel Brandão
Federal Center of Technological Education of Minas Gerais (CEFET-MG)
Computing Department
Belo Horizonte, MG, Brazil
Email: gabrielbrandao@decom.cefetmg.br

Adriano C. M. Pereira
Federal University of Minas Gerais (UFMG)
Dept. of Computer Science
Belo Horizonte, MG, Brazil
Email: adrianoc@dcc.ufmg.br

***Abstract*—The volume of electronic transactions has raised significantly in last years, mainly due to the popularization of electronic commerce (e-commerce), such as online retailers (e.g., Amazon.com, eBay, AliExpress.com). 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 and prevention, which motivates our research. This work aims to apply and evaluate computational intelligence techniques (e.g., data mining and machine learning) to identify fraud in electronic transactions, more specifically in credit card operations performed by Web payment gateways. In order to evaluate the techniques, we apply and evaluate them in an actual dataset of the most popular Brazilian electronic payment service. Our results show good performance in fraud detection, presenting gains up to 43 percent of an economic metric, when compared to the actual scenario of the company.**

***Keywords*-Fraud Prevention; e-Commerce; e-Business; e-Payment; Machine Learning;**

I. Introduction

Recently we have observed a significant increase in the volume of electronic transactions, mainly due to the popularization of World Wide Web and electronic commerce, such as online retailers (e.g., www.ebay.com, www.walmart.com, www.amazon.com). We also testify a huge increase in the number of online frauds, 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.

Bhatla et al [1] said that the rate at which Internet credit card fraud occurs is 12 to 15 times higher than face-to-face transactions. The 12th annual online fraud report by CyberSource [2] shows that, for most of the current decade, merchant online fraud losses continued to increase, reaching a peak of $4 billion in 2008. According to Siddhartha Bhattacharyya et al. [3] with the growth in credit card transactions, as a share of the payment system, there has also been an increase in credit card fraud, and 70% of U.S. consumers are noted to be significantly concerned about identity fraud.

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 UK. The Royal Bank of Scotland, which has the largest credit card market in Europe, has more than one billion transactions per year [4]. 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 this challenge 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 define a concept of economic efficiency and apply them in an actual dataset of the most popular Brazilian electronic payment service. Our results can be used to create systems to assist fraud analysts in their jobs. The performance in fraud detection, compared with the actual scenario, is up to 43% improvement in the financial gain, that is, using the economic efficiency metric that will be later explained.

The remainder of this paper is organized as follows. Section II describes some related work. Section III presents a brief description about the computational intelligence techniques that we adopt in this work: Bayesian networks, logistic regression, neural networks and random forest. Section IV describes our case study, using a representative sample of actual data, where we present a dataset overview, the experimental methodology and results. Finally, section V presents the conclusions and future work.

II. Related Work

Due to the importance of the fraud detection problem, we may distinguish several works that discuss this subject[3], [5], [6], [7]. 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 a 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. For them, this methodology can be useful as a tool for awareness and education, and can also

978-1-4799-6953-1/14 $31.00 © 2014 IEEE
DOI 10.1109/LAWeb.2014.23

help those responsible for combating frauds associated with IT to design and implement policies to reduce risks. 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. Methodologies that characterize fraud are essential for the first phase of the process, since they are the starting point to create a model of the problem and define the best technique for its solution.

There are several researches that develop methods to detect fraud [11], [5], [12] and we can realize that these methodologies can differ significantly due to the peculiarities of each fraud type. However, what can be noticed is 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. Phua et al. [13] conducted an exploratory study of numerous articles related to fraud detection using data mining and explained these methods and techniques.These algorithms are based on some approaches such as supervised strategy with labeled data, unsupervised strategy with unlabeled data and hybrid approach.

In supervised strategy with labeled data, algorithms examine every transaction, previously labeled, to mathematically determine the profile of a fraudulent transaction and estimate your risk. Neural Networks, Support Vector Machines (SVM), Decision Trees and Bayesian Networks are some of the techniques used by this strategy. Maes et al. [14] used the STAGE algorithm for Bayesian networks and "back propagation" algorithm for neural networks to detect fraud in credit card transactions. The results show that Bayesian networks are more accurate and faster training, but are slower when applied to new instances.

In unsupervised strategy with unlabeled data, the 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 [16] proposed an approach of fraud detection for credit card using anomalies detected 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 [17] 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. It is important to note that the choice of which approach to be used depends on the methodology and the available database.

SVM and random forests are sophisticated data mining techniques, which have been noted in recent years to show superior performance across different applications [18], [19] SVMs are statistical learning techniques, with strong theoretical foundation and successful application in a range of problems [20]. They are closely related to neural networks, and through use of kernel functions, can be considered an alternate way to obtain neural network classifiers. Rather than minimizing empirical error on training data, SVMs seek to minimize an upper bound on the generalization error. As compared with techniques like neural networks which are prone to local minima, overfitting and noise, SVMs can obtain global solutions with good generalization error. Appropriate parameter selection is, however, important to obtain good results with SVM. In our application, which has a very unbalanced data, SVM does not provide good results.

There is a very complete work [21] that performs a review of the literature on the application of data mining techniques for the detection of financial fraud. Although financial fraud detection (FFD) is an emerging topic of great importance, a comprehensive literature review of the subject has yet to be carried out. This paper thus represents the first systematic, identifiable and comprehensive academic literature review of the data mining techniques that have been applied to FFD. 49 journal articles on the subject published between 1997 and 2008 were analyzed and classified into four categories of financial fraud (bank fraud, insurance fraud, securities and commodities fraud, and other related financial fraud) and six classes of data mining techniques (classification, regression, clustering, prediction, outlier detection, and visualization). The findings of this review clearly show that data mining techniques have been applied most extensively to the detection of insurance fraud, although corporate fraud and credit card fraud have also attracted a great deal of attention in recent years. The main data mining techniques used for FFD are logistic models, neural networks, the Bayesian belief network, and decision trees, all of which provide primary solutions to the problems inherent in the detection and classification of fraudulent data. This paper also addresses the gaps between FFD and the needs of the industry to encourage additional research on neglected topics, and concludes with several suggestions for further FFD research.

These related works have helped us, indicating promising strategies for detecting and preventing fraud. As the datasets are different, mainly due to the very unbalanced data of our scenario, it is not possible to directly compare the results, but they provide an idea of the efficiency of these approaches.

Moreover, as in our case the main goal is to rank the transactions to block the ones that have high probability of being fraud (chargeback), we are going to define a more precise quality indicator to measure the economic gain of each computational model.

III. Fundamentals

This section describes the techniques we apply and evaluate in this work: Bayesian networks (Section III-A), logistic regression (Section III-B), neural networks (Section III-C), and random forest (Section III-D).

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 affects directly the event D that if affected directly by event B, and so on. And *e* is an independent event.

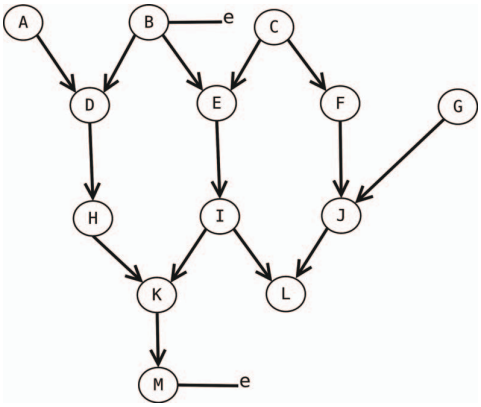

Figure 1. Bayesian Network - Description.

The mathematical definition for BN is derived of Bayes theorem, which shows that conditional probability of a event A_i given a event B, can be calculated by Equation 1, where $P(A_i|B)$ is the probability of A when B occurs.

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

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].

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

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:

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

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:

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

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

IV. Case Study

This section presents our case study where we apply the computational intelligence techniques to detect fraud in electronic transactions, more specifically in credit card in terms of chargeback operations.

A. DATASET OVERVIEW

PagSeguro[1] is a Web service for online payment, owned by the largest Latin America Internet and Web Content Provider, named Universo Online Inc.(UOL)[2], which ensures the safety of those who buy and sell on the web.

In *PagSeguro* each transaction is composed of tens of attributes of the more different types and one of these attributes refers to the status of the transaction, which can result in a valid transaction or chargeback. The purpose of this work is to analyze a set of transactions that occurred in *PagSeguro*, using the attributes that characterize these transactions to apply computational intelligence techniques, such as Bayesian Networks, Logistic Regression, Random Forest and Neural Networks, to detect fraud (chargeback).

Table I shows a short summary of the *PagSeguro* dataset. It embeds a significant sample of valid and chargeback transactions, which has thousands of transactions. Due to a confidentiality agreement, the quantitative information about this dataset cannot be presented.

	Valid	Chargeback
Average Value (US$)	36.33	81.59
Standard deviation (US$)	80.51	122.74
Median (US$)	15.00	40.00
Coefficient Of Variation	2.22	1.50

Table I
PagSeguro Dataset - Summary.

Figure 2 shows the relative quantity of chargeback transactions for each month. Despite this percentage would be considered low, it is very significant, since a chargeback transaction results in a loss of the total transaction value. Moreover, a valid transaction results in a gain of only a small percentage of the transaction value for the payment service company.

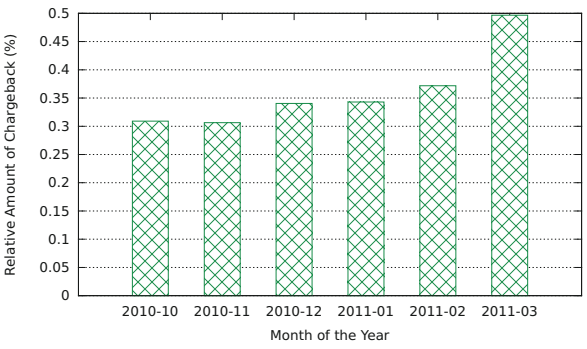

Figure 2. Relative Amount of Chargeback.

Figure 3 shows the cumulative distribution function (CDF) of transaction value. Valid transactions with values lower than US$25 correspond to 66%, and 32% for chargebacks ones. Thus, we can see that in general valid transactions present lower values than chargeback ones.

From the dataset we selected 21 attributes to be used as candidate for the techniques. The most important attributes that we use are described, as follows:

- **Value**: a numeric attribute that represents the value of transaction.
- **Score**: a literal attribute that helps to identify successful, unsuccessful and incomplete transactions.
- **Hour**: a numeric attribute that refers to the transaction create time.

[1]http://pagseguro.uol.com.br
[2]http://www.uol.com.br

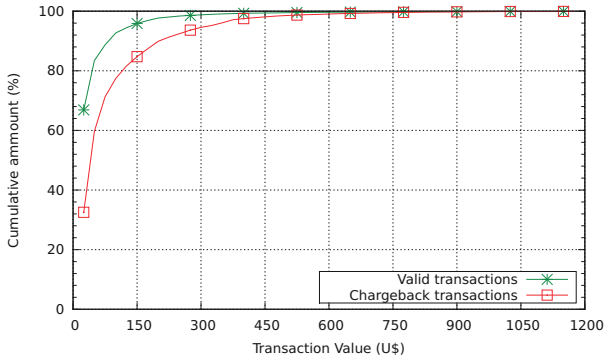

Figure 3. Cumulative Distribution Function (CDF) of Transaction Value.

- **Buyer Type**: a literal attribute that helps to identify a type of user buyer type.
- **Buyer Registration Time**: a numeric attribute that display the buyer's registration time.
- **Day**: a numeric attribute that display the day in which the transaction occurs.
- **Buyer's Points**: a numeric attribute that helps to identify users who had successfull transactions in the past.
- **Registered** $_{\text{flg\_registered}}$: A Flag attribute. In PagSeguro, users unregistered also can buy, this flag helps to identify who are the registered buyers.
- **Seller Registration Time**: a numeric attribute that display the seller's registration time.
- **Stores Main Category**: a numeric attribute that is related to the main category of the store. The category refers to the main products type sold by the store.
- **Credit Card Operator**: a numeric attribute that identify credit card operator.
- **Credit Card Owner Age**: a numeric attribute that represents, in years, how old is the credit card owner.
- **Quantity of installments** $_{\text{num\_installment\_qty}}$: A numeric attribute that it says the quantity of installments used in the purchase.
- **Status at the *Serasa***[3] $_{\text{idt\_serasa\_status}}$: A literal attribute that it shows the status of the buyer at the Serasa .
- **Had response from *Serasa*** $_{\text{flg\_has\_answer\_sr}}$: A flag attribute that it shows if the consults at the Serasa returns an answer about the buyer.
- ***CPF***[4] $_{\text{flg\_cpf}}$: A flag attribute. This attributte tells that CPF of the buyer is the same of the CPF of credit card owner.

[3]The Serasa is a private company that owns one of largest data base in the world and devotes its activity to the provision of services of general interest. The institution is recognized by the code of consumer protection as a entity of public nature.

[4]A document that identify the individual taxpayer in face of the Federal Revenue Secretariat of Brazil (FRSB). The CPF holds the registration information provided by the individual taxpayer that the other data systems of the FRSB.

- **DDD**[3]: a flag attribute that compares if the DDD of the registered user in the *PagSeguro*is consistent with the DDD of the credit card owner.
- **Federation Unit**: a nominal attribute that refers to the Federation Unit provided by the user.

B. Methodology

We used the same methodology for all techniques, starting with a characterization of our dataset, which allowed us to remove items with lower significance and categorize some numeric variables. We made a selection of the most relevant attributes to fraud detection, using "Forward Stepwise Regression", which is based in the verisimilitude concept [32]. We also use InfoGain, which shows the relative gain of each variable, and this was made in Weka[4]. Weka is a free software, under GPL License and it has many data mining and classification algorithms in its toolbox.

Following this process we define the training and test sets to evaluate the algorithms. We use the first 3 weeks of the month for training and remaining for test. This reproduces a real scenario situation and guarantees the model generality. We also use the technique of "K-fold-Cross-Validation" to validate the quality of our experiments. In order to perform this, we define the sub-samples number (K) to 5.

To evaluate the fraud detection techniques we use different environments, each of them has its own parameters. The fine tuning of these parameters was made using an exhaustive search testing different values for each technique. Next, we describe some details about the techniques and experiments, such as the parameters used for each technique.

- We use the software R[5] to build the LR model. To binary LR we use GLM package with the parameters: "FORMULA" where we set the response variable to chargeback and independent variables the others. "FAMILY" is defined as binomial and "LINK" as logit.
- For the BN we use Weka. We use the parameter **Q** to algorithm Hill Climbing to search for the network topology, as subcategory of it we have the parameters: "-P" as the maximum number of father nodes set to 1 because this the number of our response variable and "-S" to define the score that is used to mount the Bayes probability table.
- For the NN we use MATLAB Neural Network Toolbox. The network is a MLP with one hidden layer consisting of ten neurons and with the output layer of one neuron. The activation function of the hidden layer is tangent sigmoidal and linear for the output layer. For the training stage, we use the Levenberg-Marquardt algorithm.
- For the RF we use Weka. This implementation only permits the manipulation of the "Max Depth" of the

[3]Long Distance Call

[4]http://www.cs.waikato.ac.nz/ml/weka/

[5]http://www.r-project.org/

trees, the "Number of Features" to be used in the random selection and the "Number of Trees". We set "Max Depth" to unlimited as well as the "Number of Features" and "Number of Trees" to 10.

After the execution of the techniques we construct a ranking by the degree of reliability to fraud assigned to each transaction. On the top of the ranking would be the transactions with the higher probability to be fraud. After this, we apply Equation 7 in many ranking ranges to obtain the best result.

Beyond the precision we measured the recall that is the ability to find all the existent frauds. We also defined a fitness function called Economic Efficiency (EE) that can be seen on Equation 5. The Gain (G) represents the financial value of true positive transactions, rate (r) is a percentage that the company gains in a successful transaction and the Lost (L) is the financial value of false negative transactions. Applying this formula in the ranking we find the position that maximizes the profit for a given algorithm.

$$EE_{Technique} = \sum_{j=1}^{n} G_j \times r - L_j \times (1 - r) \quad (5)$$

Equation 6 is a simplification of the Company profit. The **r** value has the same meaning as described before in Equation 5, **NF** represents the financial value of Non-Fraud transactions and **F** is the financial value of fraud transactions.

$$EE_{Real} = \sum_{j=1}^{n} NF_j \times r - F_j \times (1 - r) \quad (6)$$

Equation 7 gives a relative gain where 100% represents the maximum gain and 0% is the actual scenario without the use of any technique. The EE_{Max} is the maximum gain that the company could have when no fraud occurs. We will use this equation in section IV-C to compare all techniques.

$$EE = \frac{EE_{Technique} - EE_{Real}}{EE_{Max} - EE_{Real}} \quad (7)$$

We are not using precision rate to measure how efficient is a technique due to the unbalanced dataset. A random algorithm model would get a very low precision for chargeback, less than 0.5%. This is the reason why we use the EE that is the most relevant factor in our scenario. Using this concept we also avoid the misunderstand of a high precision when classifying all transactions as valid and none as chargeback.

C. Results

Table II summarizes the results for techniques previously described in Section III. The best result between all techniques was the **NN** in March, with 43.66% of EE. Except BN, October is the worst month for all techniques. It is important to emphasize the most important measurement is the Economic Efficiency (EE), which is represented by *Rank*. We inform about precision and recall, which are traditional classification metrics, however in our problem we want to rank transitions according to a fraud score ranking, thus it is not a typical classification problem.

		BN	**LR**	**NN**	**RF**
Oct.	Prec.	7.05	4.10	7.00	10.17
	Rec.	18.93	27.52	9.00	11.47
	Rank.	0.79	1.98	0.36	0.33
	EE	**25.28**	12.03	11.69	8.13
Nov.	Prec.	14.70	8.33	5.00	19.02
	Rec.	32.38	36.67	39.00	27.01
	Rank.	0.73	1.47	2.57	0.47
	EE	29.70	28.73	**33.64**	22.42
Dec.	Prec.	7.40	3.53	5.00	14.17
	Rec.	21.08	30.20	23.00	14.55
	Rank.	1.16	3.49	1.75	0.42
	EE	16.61	10.64	**20.04**	18.02
Jan.	Prec.	8.78	9.70	6.00	13.11
	Rec.	25.56	21.19	21.00	10.60
	Rank.	1.30	0.98	1.30	0.32
	EE	**16.57**	15.54	11.98	9.90
Feb.	Prec.	7.78	6.06	9.00	7.55
	Rec.	42.96	44.62	19.00	18.36
	Rank.	3.10	4.13	1.03	1.12
	EE	**27.40**	25.75	24.03	12.01
Mar.	Prec.	9.93	5.38	6.00	4.24
	Rec.	43.01	49.94	34.00	32.32
	Rank.	2.22	4.76	3.18	3.91
	EE	35.53	35.61	**43.66**	13.48

Table II
COMPARATIVE RESULTS FOR ALL TECHNIQUES ON THE WHOLE DATASET. ABBREVIATIONS: "PREC." IS PRECISION, "REC." IS RECALL, "RANK." IS RANKING

BN has its lowest gain in October with 14.33% of EE, 7.05% of precision at position 0.79% of the ranking. The best result was achieved in March with 35.53% of EE, 9.93% of precision and ranking coverage with 2.22%. The higher precision value was obtained in November with 32.38% of recall. The higher recall rate is in March with 43.01%.

LR has its lowest gain in October with 12.03% of EE with 4.10% of precision and its best EE is 35.61% in March.

NN presents its worst results in October and January with 11.69% and 11.98%, respectively. Its best result is in March with 43.66% of EE and 6% of precision.

RF has the worst EE in October with 8.13%. Its best is 22.42% of EE in November with precision of 19.02% at 0.47% of the ranking.

Figure 4 shows the EE until 8% of the ranking in March. **NN** presents the best performance until 5.80% of the ranking, and after that it drops and stays bellow **BN** curve. The **RF** stays bellow the others until to the end.

These results shows that all the four algorithms can bring

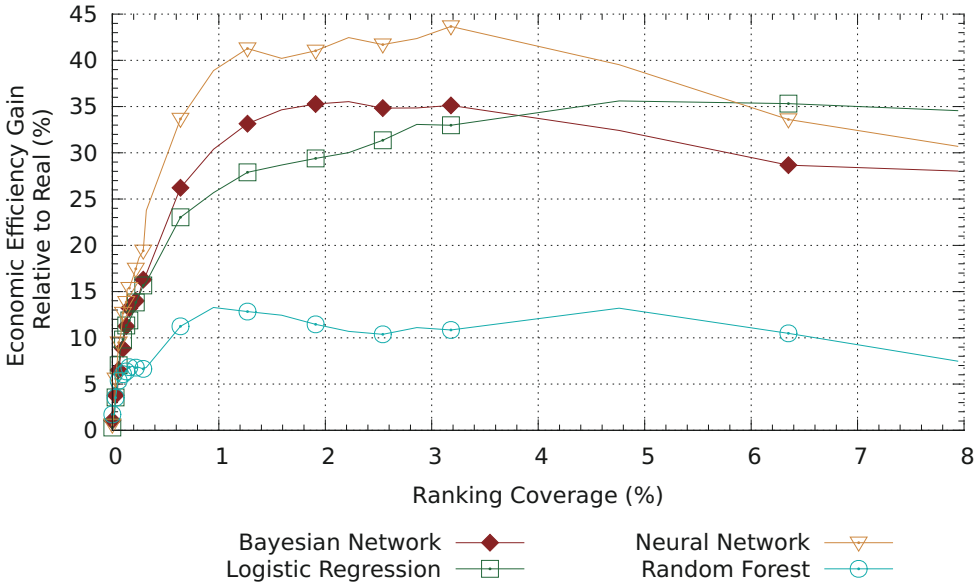

Figure 4. March - EE versus Ranking Position

gains to the company, even the less effective technique reaches at least 8% of Economic Efficiency gain. This methodology of fraud detection can be used by e-commerce companies to reduce the risk in credit card operations. If we compare the techniques to choose the one that would be the best to avoid chargeback, we identify that Bayesian Networks (BN) is the best one, since Neural Networks (NN) presented lower values in some months of the actual dataset. Therefore **BN** has been chosen as the best technique for this scenario, presenting significant gains for all months of data.

V. Conclusion

In this work we build different fraud detection models to predict fraud in online transactions, more specifically credit card operations. We apply and evaluate four different computation intelligence techniques, after choosing them from an initial set of evaluated experiments that adopt several distinct techniques. In order to evaluate the techniques, we apply them in an actual dataset, containing thousands of transactions per day, from the most popular Brazilian electronic payment service, called *PagSeguro*.

We confirm that imbalanced classes, fraud and non-fraud, was a factor that directs impacts on the prediction gains. The achieved results present significant gains when compared to actual scenario of the company, which adopts some fraud detection procedures. In order to compare the techniques, we adopt an Economic Efficiency (EE) function, which describes the financial improvement relative to the actual scenario from the corporation. In the best case, we have achieved a gain of 43.66%.

We realize that the worst results were obtained in the months with a fewer amount of fraud transactions than other ones. Neural Network and Bayesian Networks have performed the best results. The Logistic Regression approach reached its better result in March with 35.61% of EE, slightly better than Bayesian Networks and worst than Neural Networks, with 43.66% of EE. The worst technique was Random Forest with gains in the range of 8.13% to 22.42%.

One of the challenges of this research is the nature of data, since they are much unbalanced with the minor class with less than 1%. As a future work we intend to use techniques to deal with this data imbalance, preserving the generality of the model. One possible solution would be consider weights to assign to classes, where the minor class receives the larger weight [33]. Moreover, Xie et al. [34] have proposed an improvement to Random Forest technique, combining the balanced random forest and the weighted random forest. Thus, an idea is to optimize the computational techniques and another proposal to improve the gains is to use hybrid models that can be composed by ensemble of techniques.

Acknowledgment

This research was supported by the Brazilian National Institute of Science and Technology for the Web (CNPq grant numbers 573871/2008-6 and 477709/2012-5), CAPES, CNPq, Finep, and Fapemig.

References

[1] V. P. Tej Paul Bhatla and A. Dua, *Understanding Credit Card Frauds*, 2003.

[2] C. Mindware Research Group, *2011 Online Fraud Report*, 12th ed., 2011. [Online]. Available: http://www.cybersource.com

[3] S. Bhattacharyya, S. Jha, K. Tharakunnel, Westland, and J. Christopher, "Data mining for credit card fraud: A comparative study," *Decis. Support Syst.*, vol. 50, pp. 602–613, February 2011.

[4] R. J. Bolton and D. J. H, "Statistical fraud detection: A review," p. 2002, 2002.

[5] 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

[6] P. Ravisankar, V. Ravi, G. R. Rao, and I. Bose, "Detection of financial statement fraud and feature selection using data mining techniques," *Decision Support Systems*, vol. 50, no. 2, pp. 491–500, 2011.

[7] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems*, vol. 50, no. 3, pp. 559–569, 2011.

[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 *Proceedings of the Proceedings of the 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–. [Online]. Available: http://portal.acm.org/citation.cfm?id=962755.963148

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

[11] T. Fawcett and F. Provost, "Adaptive fraud detection. data mining and knowledge discovery," 1997.

[12] E. L. Barse, H. Kvarnström, and E. Jonsson, "Synthesizing test data for fraud detection systems," in *Proceedings of the 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

[13] C. Phua, V. Lee, K. Smith-Miles, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," 2005.

[14] 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.

[15] Netmap, "Fraud and crime example brochure," 2004.

[16] R. J. Bolton and D. J. Hand, "Unsupervised Profiling Methods for Fraud Detection," *Statistical Science*, vol. 17, no. 3, pp. 235–255, 2002. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.24.5743

[17] G. J. Williams and Z. Huang, "Mining the knowledge mine: The hot spots methodology for mining large real world databases," 1997.

[18] B. Larivière and D. Van den Poel, "Predicting customer retention and profitability by using random forests and regression forests techniques," *Expert Syst. Appl.*, vol. 29, no. 2, pp. 472–484, Aug. 2005. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2005.04.043

[19] A. R. Statnikov, L. Wang, and C. F. Aliferis, "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification." *BMC Bioinformatics*, vol. 9, 2008. [Online]. Available: http://dblp.uni-trier.de/db/journals/bmcbi/bmcbi9.html#StatnikovWA08

[20] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011. [Online]. Available: http://doi.acm.org/10.1145/1961189.1961199

[21] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decis. Support Syst.*, vol. 50, no. 3, pp. 559–569, Feb. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.dss.2010.08.006

[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 the R Development Core Team, "An introduction to r," http://www.cran.r-project.org, [Online; Accessed: July 20, 2014].

[27] K. Gurney and K. Gurney, *An introduction to neural networks*. CRC Press, 1997.

[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. A. Lourakis, "A brief description of the levenberg-marquardt algorithm," vol. 3, p. 2, 2005.

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

[32] T. R. D. C. T. Version, "R: A language and environment for statistical computing," http://www.r-project.org, [Online; Accessed: July 20, 2014].

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

[34] 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.