Revenue Recovering with Insolvency Prevention on a Brazilian Telecom Operator

Carlos André R. Pinheiro
Brasil Telecom
SIA Sul ASP Lote D Bloco F
71.215-000 Brasília, DF, Brazil

andrep@brasiltelecom.com.br

Alexandre G. Evsukoff
COPPE/UFRJ
P.O. Box 68506, 21941-972
Rio de Janeiro, RJ, Brazil

evsukoff@coc.ufrj.br

Nelson F. F. Ebecken
COPPE/UFRJ
P.O. Box 68506, 21941-972
Rio de Janeiro, RJ, Brazil

nelson@ntt.ufrj.br

ABSTRACT

This paper deals with a real application on a brazilian telephone company. Data mining analysis, based on neural networks, was performed on the customer base in order to understand and to prevent bad debt events. This paper describes the knowledge discovering process and focuses on two main products: the cluster analysis of the customer base and a bad debt event classification model. The segmentation of the customer base has provided a better understanding of several groups of customers' behavior. Distinct actions are taken, depending on the segment a given client was put in, according to strategic directions of the company. The classification of insolvent customers is used as a tool to help the company to take preventing actions in order to avoid main losses and taxes leakage. The results of the project's implantation show that investment on information technology infra-structure for data mining is highly profitable.

Keywords

Insolvency detection, revenue recovering, customer relationship management, neural networks.

1. INTRODUCTION

Bad debt affects companies in several ways, changing operational processes, and affecting the companies' investments and profitability. In the telecom business, bad debt has specific and significant characteristics, especially in the fix phone Companies. According to regulation, even if the customer becomes insolvent, the service should be continued for a specific period, a fact which increases the losses.

Data mining have been widely recognized as powerful tool for fraud detection and direct marketing applications in several domains [1][2][3]. Nevertheless, one of the important issues in data mining, which is roughly discussed in the literature, is data mining deployment actions, i.e. the results of feedback actions taken from the knowledge extracted from data. Moreover, a clear business understanding and the analysis of goals by the data mining team is crucial for a project's success.

This paper presents a case study of the development of patterns of recognition models to the insolvent customers' behavior, with focus on the previous identification of non-payment events. The purpose of this set of models is to avoid revenue losses due to clients' insolvency [4]. The paper presents a discussion on the importance of business goals understanding in order to direct the choice of algorithms and to offer a better interpretation of the results. The whole knowledge discovering process is described

with a focus on an estimative of revenue recovery from deploying of actions based on the data mining results.

On this case study, the knowledge about the customers came from two distinct data mining models: a segmentation model and a classification model. The segmentation model was based on customer behavior by Kohonen self organizing maps [6]. The clusters found have allowed business intelligence analysts to understand the different ways through which the customers become insolvent and then, to create distinct actions to intelligently collect them. The classification model was based on a bagging of feed forward neural networks. The classification model allows managing the insolvency risks in real time. The company can thus decide, based on customers' behavior prediction, the most appropriate monitoring action, which could be, for instance, not to collect insolvent customers, if they are highly risky. By doing such, losses due to unfaithfully actions are avoided, as so as taxes' leakage [8].

This environment, including all models described in this paper, was implemented and has been deployed since 2005 at *Brasil Telecom*, a Brazilian telephone company. Since the implantation, the accuracy of the predictive models has allowed important refunds that greatly overpasses the initial investment on the hardware/software infrastructure, as well as all the cost of the model development.

The paper is organized as follows. Next section presents the application details, and dataset selection, cleaning and transformation. Section three presents the cluster analysis for the segmentation of the customer base. Section four discusses the implementation of the classification model. Section five presents some estimates of the revenue recovering with collection notices and tax avoiding. Finally, the conclusions are highlighted.

2. APPLICATION DATA AND TOOLS

This application was developed at *Brasil Telecom*, which operates on the central and west states in Brazil, as well as surrounding countries. By the end of 2004, *Brasil Telecom* Group counted 10.7 million fixed lines, 622.3 thousand mobile accesses, and 535.5 thousand ADSL accesses in service, which granted the group an outstanding position in the Brazilian telecommunication sector.

The company recently implemented new IT systems to meet business need in several strategic targets. In this project, the business target was loss reduction by preventing and minimizing the effects of bad debt events within the residential fixed lines. Bad debts must be distinguished from fraud events, which have a separate treatment. Bad debts are more related to insolvency, i.e.

customers that are delayed on their billing, but do not have (or do not seem to have) the intention to fraud the company.

This project focused on residential, fixed line customers. A first selection from the company's data warehouse generated the database for the study, with more than 5 million clients. Although the difference among insolvent and fraudulent customers is quite subtle, all customers whose services were blocked based on bad used or financial reasons were eliminated from the original database. The original dataset for model building was obtained from a random selection of 5% of the original database. After model building, the models were applied against the whole customer base.

Variable selection is a very important issue in model developing and, in this project, was performed by business analysis. The following variables were used in the unsupervised model study: branch, location, billing average, insolvency average, total days of delay, number of debts and traffic usage by type of call, in terms of minutes and amount of calls. From a total of 152 variables, of different types as shown in Table 1, 47 variables were selected for the study. Additionally, other variables have been created, such as the indicators and average billing, insolvency, number of products and services and traffic usage values. Derived variables were created, by normalization, systematically using logarithm functions and by ranges of values in some cases.

Table 1: Number and type of variables used in the models.

Type of information	# total	# selected
Personal information	8	1
Product using	12	4
Billing behavior	4	2
Insolvency behavior	8	4
Arrears behavior	12	3
Block actions	20	20
Traffic using	88	13
Total	**152**	**47**

A commercial data mining suite was used to develop the models described in the following sections. The study was run on a Fujitsu, with 4 CPU's, 8 GB of RAM hardware with nearly 500 GB of disk array.

3. INSOLVENCY SEGMENTATION

Since the insolvent customers' behavior was unknown, the first step of the project was to build an unsupervised model for the identification of the most significant characteristics of the insolvent customers. Thus, from the modeling dataset, the unsupervised model was built with the records of insolvent customers only, considered as the customers with any payment delay greater than three days.

The unsupervised model was developed using self-organizing Kohonen maps. Several models, considering different variables sets and model parameters, were necessary to improve accuracy and a better understanding of the results [3][4].

The clustering analysis identified five distinct groups, named G1 to G5, according to different insolvent profiles. The relation among population, billing and insolvency for each group is shown in Figure 1. It can be seen that group G1 represents only 13.7% of the population, but responds to 34.7% of the total of billing and is responsible for 26.0% of insolvency.
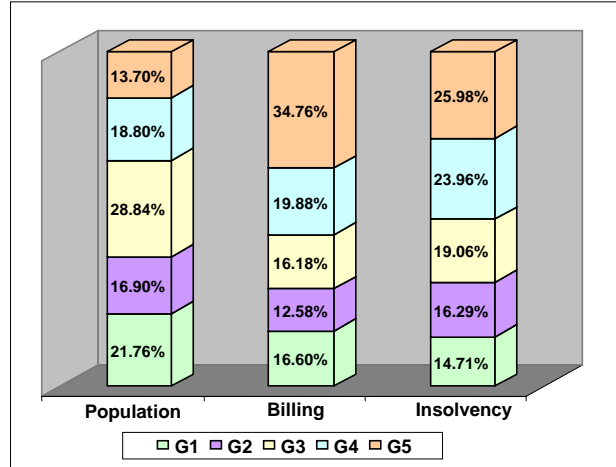

Figure 1: Distribution of billing, bad debt and population based on the clusters identified.

Analyzing each group individually allows business analysts to get an extremely important knowledge of different behaviors in each customer segment. One of the most important features of the segmented groups is the relation between the billing and the insolvency values according to the delay of payment. Those variables provide a clear understanding of the clients' capacity of paying on each cluster and, therefore, allow us to quantify the insolvency risk. Another important analysis is the relation between the insolvency values and the average payment delay of a certain group. Most part of the large telephone companies, work with an extremely tight cash flow, with no gross margin and small financial reserves. As a consequence, for a reasonable period of delay associated with high insolvency values, it is necessary to obtain financial resources in the market. Since the cost of capital in the market is bigger than the fines charged for delay, the company has revenue losses, i.e., even if the company receives the fines, there is a cost to cover the cash flow.

Some characteristics of the insolvent customers, based on their average value of bill, bad debt, local and long distance traffic usage, and products and services subscribed are summarized in Table 2.

Table 2: Comparison of average group values.

Variables	G1	G2	G3	G4	G5
Billing	R$ 83	R$81	R$ 61	R$115	R$276
Bad debt	R$ 87	R$124	R$ 85	R$164	R$244
Arrears	27 d	30 d	24 d	22 d	27 d
Local calls	157 p	155 p	124 p	220 p	538 p
Long dist. calls	73 m	78 m	35 m	141 m	297 m
Population	21.7%	16.9%	28.8%	18.8%	13.7%

From the values in Table 2, one can conclude that group G1 presents average usage with relative compromise of payment; group G2 presents average usage with low compromise of payment; group G3 presents low usage with good compromise of payment; group G4 presents high usage with compromise of payment; group G5 presents very high usage with relative compromise of payment.

Based on the main characteristics of each group, the resulting labeling by analysis are summarized as:

• Group G1 was labeled as moderate, with medium billing value, medium insolvency value, high period of delay, medium volume of debts and medium value of traffic usage.

• Group G2 was labeled as bad, with medium value of billing, high value of insolvency, medium period of arrears, medium volume of debts and medium value of traffic usage.

• Group G3 was labeled as good, with low value of billing, medium value of insolvency, medium period of arrears, medium volume of debts and low value of traffic usage.

• Group G4 was labeled as very bad, with high value of billing, high value of insolvency, medium period of arrears, medium volume of debts and medium value of traffic usage.

• Group G5 was labeled as offender, with high value of billing, very high value of insolvency, high period of arrears, high volume of debts and very high value of traffic usage.

The segmentation model makes the identification of insolvency behavior possible and helps the company to create specific collecting politics according to the characteristics of each group. The creation of different politics, according to the identified groups, allows one to use a collection of rules with distinct actions for specific periods of delay and according to the category of client [4]. These politics can bring satisfactory results, not only concerning the reduction of expenses with collecting actions but also the recovering of revenue as a whole. Therefore, two distinct areas of the company can be directly benefited by the data mining models developed: Collection, by creating different collecting politics, and Revenue Assurance, by creating insolvency prevention procedures.

Although the segmentation model itself is a powerful tool for understanding the insolvency behavior, it is not adequate to predict insolvency risks in real time. Therefore, it is necessary to develop a classification model of these events and, consequently, avoid them, reducing the revenue loss for the company as described next.

4. INSOLVENCY PREDICTION

The insolvency prediction model was built over the entire modeling dataset, including all customers, insolvent or not. In order to build a supervised learning base, customers were classified as "good" (G) or "bad" (B) customers. According to the business rules of the company, a customer is classified as "bad" if the payment delay is greater than 29 days. The customer is considered "good" otherwise. The overall customer base has 72% of "good" customers and 28% of "bad customers.

The objective of the classifier model is to predict "bad" customers based on their profile, before any payment delay, allowing the

company to decide the best prevention action to take, according to the costumer group identified in the previous segmentation model.

The classifier was built as a bagging of neural network classifiers, by previously dividing the base into a number of groups with similar behavior. This procedure has enhanced the classifier performance. The Kohonen's self-organizing maps was used again to set up ten clusters, which were used to feed the neural network classifier models.

The predicting model based on the segmented database has allowed the classification models to deal with more homogeneous data sets, since the clients classified in a specific segment have similar characteristics. The base segmented in ten different categories of clients was used as the input data set for the construction of ten different classification models as shown in Figure 2.
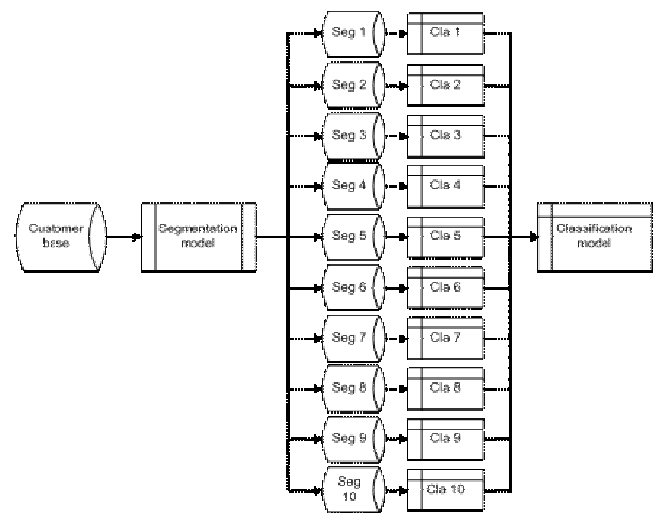

Figure 2: Bagging of neural network classifier.

Feed forward neural network models were used in each one of the sub-classifiers. The parameters of each sub-classifier were optimized for their respective data subset. Individual model results of each model, computed by ten-fold cross validation are presented in Table 3. The results are presented in terms of the sensitivity ratio for each class, i.e the ratio between the correctly and the total of instances of each class in the data base and the number of records assigned to each group.

Table 3: Sensitivity of individual classifiers

Group	# records	Sensitivity (%)	
		Good	Bad
G1	51,173	84.96	96.45
G2	22,051	82.89	96.89
G3	17,116	85.78	92.14
G4	23,276	83.87	89.04
G5	171,789	81.59	88.36
G6	9,902	80.92	88.57
G7	23,827	82.09	89.92
G8	48,068	86.44	97.08
G9	19,194	88.69	91.32
G10	14,868	87.23	93.98
Average		84.45	92.38

By developing the classifier model as a bagging of ten classifiers, each of it using segmented samples of clients, it is possible to adequate and specify each model individually. This procedure has allowed developing more accurate predicting models. For comparison reasons, a simple neural network model was developed previously using the same basic configuration, but using the entire database, i.e. without segmentation, in 10-fold cross validation classification analysis. This model reached a sensitivity ratio of 83.95% for class "good", and 81.25% for class "bad".

Comparing the classification model based on a single data sample with the bagging approach, the benefit to the class "good" sensitivity ratio was small, only 0.5%. However, the sensitivity ration achieved for class "bad" with the bagging approach was more than 11% better than the one obtained with the single base, as shown in Figure 3.
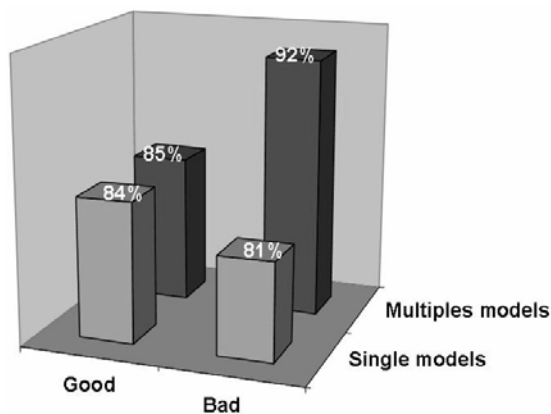

Figure 3: Comparison of predictive models accuracy based on different approaches.

Since the billing and collection actions are focused on the class of "bad" clients, the gain in the classification ratio of good clients was not very significant, but the gain in the bad clients' classification ratio was extremely important, since it helps directing actions more precisely, with less risks of errors and, consequently, less possibility of revenue loss in different billing and collection politics.

Combining the results of the behavior segmentation model, which has split clients into ten different groups according to the subscribers' characteristics and the use of telecom services, with the classification models that predict the class of the clients "good or "bad, we observe that each of the identified groups has a very peculiar distribution of clients between classes, which is, in fact, another group identification characteristic.

The results of both unsupervised and supervised classification models have allowed defining action plans for bad debt prevention, as described next.

5. DEPLOYMENT ACTIONS
One of the important issues in data mining that is rarely discussed in the literature is the feedback actions taken from the knowledge extracted from data. Besides the steps of the process as a whole, such as the identification of the business objectives and the functional requirements, the extraction and transformation of the original data, the data manipulation, the mining modeling and the analysis of the results, there is a very important step for the company, which is the effective implementation of the results in terms of business actions.

The action plans, which are deployed based on the knowledge acquired from the results of the data mining models, are responsible for bringing the real expected financial return when implementing projects of this nature.

Several business areas can be benefited by the actions that can be established. Marketing can establish better relationship channels with the clients and create products that better meet their needs. The Financial Department can develop different politics for cash flow and budget according to the behavior characteristics of the clients concerning payment of bills. Collection can establish revenue assurance actions, involving different collecting plans as well as credit analysis politics.

The use of the results of the models constructed, even the insolvent clients' behavior segmentation or the subscribers' summarized value scoring, as well as the classification and insolvency prediction can bring significant financial benefits for the company. Some simple actions defined based on the knowledge acquired from the results of the models can help recovering revenue, as well as avoid debt evasion due to matters related to regulations.

5.1 Recovering from collection notices
The insolvency segmentation model helps to define specific collecting actions and to recover revenue related to non payment events. This model has identified five groups of characteristics, some of them allowing us to use different collecting actions for each group of customers, avoiding additional expenses with collecting actions.

In *Brasil Telecom*, the collecting actions occur in temporal events, according to the defined insolvency rule. Generally, when the clients are insolvent for 15 days, they receive a collection notice, informing the possibility to have their services cut off. After 30 days of arrears, the customers have their services interrupted partially, becoming unable to make calls. After 60 days, the customers have their services cut off totally, becoming unable to make and receive calls. After 90 days, the clients' names are forwarded to a collecting company which keeps a percentage of the revenue recovered. After 120 days, the insolvent clients' cases are sent to judicial collection. In these cases, besides the percentage given to the collecting company, part of the revenue recovered is used to pay lawyers fees. After 180 days, the clients are definitely considered as revenue lost, and are not accounted as possible future revenue for the company.

Each action associated to this collection rule represents some type of operational cost. For instance, and as part of the goals of this case study, the printing cost of letters sent after 15 days of arrears is approximately $0.05. The average posting cost for each letter is $0.35. These may seem trifling at first view, but when considered number of letters required, the final values can represent a great recovering.

In order to estimate the recovery that could be obtained from collection actions, the prediction model was applied on the customers' database concerned with the collection actions, i.e. customers with at least three days of delay. The idea is to predict those customers that will indeed pay their bill, independently on any collection action, i.e. the "good" customers.

Considering 95% of confidence on the prediction model result, the classifier has identified 340,500 customers that would pay their bills, even with some delay, but less than 29 days, and hence, any type of actions from the company will produce no real effects. Assuming that these customers will pay their bills before the collection letter arrives, the company could not issue the 15th day collect letters, which will mean a significant cost reduction. In order to get an estimate, taking into account that the average printing and posting cost is around $0.40 for each notice, the financial expenses monthly avoided only by inhibiting the collecting notice issuing would be of approximately $136 K. On the other hand, considering the classification error of the prediction model, about 15% of these customers would need to receive the collection letter to accomplish the payment. Thus, a total of 51,075 should be noticed but, according the collection policies, those customers will receive a letter assigned to the 30 days of arrears.

The recovering from collection notice is not an expressive number but the application of a selective notice represent a great progress on the company's relationship with insolvent customers. Moreover, the results of the insolvent customers' segmentation can be used to produce different letters, according to the different groups of insolvent customers.

A greater revenue recovery can be obtained from taxes avoiding, as described next.

5.2 Recovering from tax avoiding
The behavior characteristics of each group identified and the population distribution according to class "good" or "bad" brings new knowledge of the business, which can be used to define new corporate collecting politics.

In Brazil, the Telephone Companies are obliged to pay value-added taxes on sales and services when issuing the telephone bills. Independently if the client will pay the bill or not, when the bills are issued, 25% of the total amount is paid in taxes by the company.

The insolvency classification model can be very efficient to predict the clients that will not pay the bill, as a matter of fact. Therefore, an action inhibiting issuing bills for clients that tend to be highly insolvent can avoid a considerable revenue loss for the company, since it will not be necessary to pay the taxes.

In the classification model developed, the customers' trend to become insolvent or not varies according to the confidence values issued by the model. Taking into account only the range with 95% of confidence value to the "bad", there are a total 481,000 clients with high risk to not pay the bill. Since the monthly average billing of these clients is about $47.50, the total possible value to be identified and related to insolvency events is $22.8 M, which corresponds to $5.7 M in taxes due.

The total accuracy level of the classifier is 92.5% for the bad class observations that are predicted as really belonging to the bad class. This ratio was obtained from the insolvency classification model based on cross-validation. The total clients whose telephone bills issuing would be correctly inhibited for the first segment was 444,000 cases. Based on the same billing average, the value relative to non-payment events is about $21.1 M. Consequently, the total of tax that can be correctly avoided by inhibiting the telephone bill issuing is about $5.3 M.

Taking into consideration the model classification error, the population of incorrect predictions is 36,000 clients, with a billing loss of $1.7 M. The value that would be assigned to the tax must not to be considered as loss, and thus must be subtracted from the total amount of billing loss.

In order to get an estimate of the total of revenue recovering due to tax avoiding, a balance is shown in Table 4. The columns in Table 4 show the number of customers, the total of revenue loss, the taxes due and the possible recovery. The lines show the total of selected "bad" customers, considering 95% of model confidence, the rate of correctly classified (92.5%) and the rate of incorrectly classified (7.5%). In the balance line, it is shown an estimate of the total amount of the recovery, considering the recovery due to not issuing the bill discarding the lack of income due to classification error.

Table 4: Total revenue recovered from the insolvency prediction.

	# Cust.	Loss	Tax Due	Recover
Selected	480,000	$22,800,000	$5,700,000.	
Correct (92.5%)	444,000	$21,090,000	$0.	$5,272,500.
Wrong (7.5%)	36,000	$1,710,000	$427,500.	-$1,282,500.
Balance				**$3,990,000.**

The above estimation was computed on a monthly basis, considering a linear and constant behavior of the insolvent customers, a fact which has actually been happening for the last years. The amount of tax recovery can be estimated as around $ 4 M per month.

Naturally, there is always the risk of not issuing a bill for a client that would make the payment, due to the classification error inherent to the classification model. However, the classifier error in the above analysis was overestimated, since 7.5% is the total classifier error but only the customers with 95% of confidence were considered in the selection. In that case, the actual classification error would be considerably lower. Nevertheless, even considering a greater classifier error, the financial gain is still significantly compensating, overcoming all the investment for data mining infrastructure in hardware, software and human resources.

According to Brazilian laws, the company is allowed to recover a portion of the taxes paid for bills that were not collected, at end of the fiscal year, if it is proved that they are due to fraudulent actions. Insolvency, however, is considered a business risk and do not allow tax recovery. The gains related to insolvency are, in fact, on the cost of capital acquisition in the market that is necessary to cover cash flow deficits occurred due to values not received from insolvent clients.

6. CONCLUSION
This work has shown a real application of data mining techniques for insolvency prevention and revenue recovering in a brazilian telecom operator. Two kinds of models were developed: one unsupervised clustering model for insolvent customers' base segmentation and a supervised classification model for insolvency prediction.

The segmentation model has allowed to structure the knowledge about the clients' behavior according to specific characteristics, based on which they can be grouped. The segmentation model allowed the company to define specific relationship actions for each of the identified groups, associating the distinct approaches with some specific characteristics. The analysis of the different groups identifies the main characteristics of each of them according to business perspectives. The knowledge extracted from the segmentation model, which is extremely analytical, helps to define more focused actions against insolvency creating more efficient collecting procedures.

The classification models have allowed us to anticipate specific events, becoming more pro-active and, consequently, more efficient in the business processes. The deployment of the classification results, even considering an overestimated classifier error, has shown to be significantly compensating as tax recovering policies.

The actual application of the main recommendation of not issuing bills for highly risky insolvent customers is not only a technical matter. There are many other issues to be considered other that the financial estimate of gains. At the moment this paper was written, the actual application of recommended actions for tax recovering was yet in study. However, the issuing of selective collecting notices was deployed roughly as described in this paper.

The practical results of this study have largely surpassed the most optimistic estimates. More than the modeling results themselves and the perspective of financial gains, this study has contributed to disseminate a culture of data mining as a tool for efficiency in large corporate business. In this sense, the application and deployment of data mining techniques have spread over several other departments of the company.

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] Berry M. J. A., Linoff G. Data Mining Techniques: for Marketing, Sales, and Customer Support, Wiley Computer Publishing, 1997.

[2] Fayyad U. M., Piatetsky-Shapiro G., Smyth P. and Uthurusamy R.. Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, 1996.

[3] Abbot, D. W., Matkovsky, I. P., Elder IV, J. F., *An Evaluation of High-end Data Mining Tools for Fraud Detection*, IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA, October, pp. 12-14, 1998.

[4] Burge, P., Shawe-Taylor, J., Cooke, C., Moreau, Y., Preneel, B., Stoermann, C., *Fraud Detection and Management in Mobile Telecommunications Networks*, In Proceedings of the European Conference on Security and Detection ECOS 97, London, 1997.

[5] Haykin, S., *Neural Networks – A Comprehensive Foundation*, Macmillian College Publishing Company, 1994.

[6] Kohonen, T., *Self-Organization and Associative Memory*, Springer-Verlag, Berlin, 3rd Edition, 1989.

[7] Lippmann, R.P., *An introduction to computing with neural nets*, IEEE Computer Society, v. 3, pp. 4-22, 1987.

[8] Pinheiro, C. A. R., Ebecken, N. F. F., Evsukoff, A. G., *Identifying Insolvency Profile in a Telephony Operator Database*, Data Mining 2003, pp. 351-366, Rio de Janeiro, Brazil, 2003.

[9] Ritter, H., Martinetz, T., Schulten, K., *Neural Computation and Self-Organizing Maps*, Addison-Wesley New York, 1992.

About the authors:

Carlos Andre R. Pinheiro is IT Manager from *Brasil Telecom*, and also is responsible for BI projects, specially the ones assigned to Data Mining process, as fraud detection, revenue assurance, marketing segmentation, and several types of prediction. He received his PhD degree in Computational Systems in COPPE/UFRJ, the Engineering Graduated Center of Federal University of Rio de Janeiro. Currently, he has joined a Post Doctoral program at IMPA – *Applied and Pure Math Institute*, where his research focus is on neuro-fuzzy systems and genetic algorithms techniques.

Alexandre G. Evsukoff is Professor of Computational Systems at COPPE/UFRJ, the Engineering Graduated Center of Federal University of Rio de Janeiro. He graduated and received MSc degree both in mechanical engineering in UFRJ, respectively in 1989 and 1992. He received his Ph.D. in control theory in 1998 from National Polytechnic Institute of Grenoble, France. His research interests include the application of soft computing tools for data mining, decision support and the modeling of complex systems.

Nelson F. F. Ebecken is Professor of Computational Systems at COPPE/UFRJ, the Engineering Graduated Center of Federal University of Rio de Janeiro. His research focuses on basic methodologies for modeling and extracting knowledge from data and their application across different disciplines. He develops and integrates ideas and computational tools from statistics and information theory with artificial intelligence paradigms.