Credit Scoring Data for Information Asset Analysis

Fábio Silva[1], Cesar Analide, Paulo Novais

Abstract Risk assessment is an important topic for financial institution nowadays, especially in the context of loan applications. Some of these institutions have already implemented their own credit scoring mechanisms to evaluate their clients' risk and decide based in this indicator. In fact, the information gathered by financial institutions constitutes a valuable source of data for the creation of information assets from which credit scoring mechanisms can be developed. The purpose of this paper is to, from information assets, create a decision mechanism that is able to evaluate a client's risk. Furthermore, upon this decision mechanism, a suggestive algorithm is presented to better explain and give insights on how the decision mechanism values attributes.

1 Introduction

In current times, situations where people are unable to assess the amount of a loan that is affordable to them and, as such, incur in bad behavior regarding their monthly installments. Economical and social conjunctures are aggravating this problem and, as so, financial institutions are becoming concerned on how to develop new credit scoring systems to classify their clients according to some degree of risk that are updatable on almost real time.

The client history provides an excellent source of information for predicting the behavior of future clients. In fact, some rules and patterns can be identified in this data that may be relevant to decide where a future client should have its loan application accepted or not. From the perspective of information as an asset, this clients' data history usage creates valuable assets to an organization. In fact the information gathered from these sources is considered to be one of the six types of assets, namely it falls into the category of IT information asset [2].

Nowadays, statistical analysis and deterministic system are still the most common means of credit scoring and classification on financial institutions and their applications. This presents an opportunity to develop alternative systems based on techniques from artificial intelligence and data mining in order to extract valid knowledge and information from the data, creating valuable information assets to

[1] Fábio Silva, Cesar Analide, Paulo Novais
Department of Informatics, University of Minho, Braga, Portugal
e-mail: f.aandree@gmail.com, {analide, pjon}@di.uminho.pt

these institutions. These new techniques may also provide the means to develop semi-autonomous or even completely autonomous decision mechanisms that learn new trends as new data becomes available and also update their decision system accordingly to account for the new information.
Only in present times are these instructions conducting studies in order to evaluate how techniques from artificial intelligence and data mining can be used to predict client behavior [5, 6]. This paper is particularly aimed at credit scoring systems using previous records from old clients to predict and avoid those classified as bad client in terms of debt repayment.

2 Related Work

2.1 Models

In order to build decision models there are some considerations to which attention should be devoted. These models should take in consideration legal issues, granting this way that any decisions produced will not be declared illegal. Client discrimination based in attributes such race or gender is generally illegal in most countries and may justify legal suites to those who ignore these considerations.
In the literature, different types of models and approaches can be found regarding credit scoring and risk assessment.
Most financial institutions use statistical pattern recognition models to build their own decision mechanism. In Czech and Slovak Republics' financial institutions, the most used technique is the Logit Analysis which is an improvement upon the Linear Discriminant analysis technique [4]. In a Jordanian bank, studies were conducted to evaluate the benefit of using Multi-Layer Feed Forward Neural Networks [7]. Their study led to the conclusion that these structures are, in fact, good classifiers achieving up to 95% correct evaluations in their tests. Improvements on standard neural networks classifications using genetic algorithms were also proposed. In this case genetic algorithms are used to optimize the weight calculation in neural networks [8]. Neural networks are also used for the detection of anomalous behaviors for intrusion detection [1]. Another classification models found use case based reasoning. These systems use data from past events characterized by a set of attributes. Similarities between past cases and present cases are calculated using an appropriated functions and the final classification is made based on the most similar case [3].
Different approaches make use of financial liquidity to forecast a client's ability to pay a future installment. From an historical set of clients' financial liquidity and the comparison with the financial liquidity of a present client, its risk is calculated and appropriate actions are taken before transgression happens if necessary [11].

2.2 Algorithms

As noted in section 2.1, there are artificial intelligence techniques that use machine learning, data mining and soft computing to produce results. In this context, several algorithm optimization proposals can be found in the literature.
Improvements in genetic algorithms used for classification can be found in The Two Stage Genetic Programming Algorithm, which, first, produces a set of if-then rules and, then, a function based in genetic programming to classify instances not covered by the if-then rules[9]. Another example uses a combination of decision trees with genetic programming to improve the tree construction, improving the classification accuracy [10].
Neural networks optimization approaches try to make use of feature selection algorithms before the construction of the neural network, regarding some attributes as more relevant to this structure [12]. Feature selection using decision trees may be used to determine a set of attributes, in the upper levels of the tree, to build the subset of attributes considered to be used with the Naïve Bayes classifier [5].
All these algorithm combinations obtain improved results when compared to their original versions leading to the conclusion that combining different algorithms is a good idea for optimization.

2.3 Tools and Frameworks

There are many tools and frameworks currently available to help the data mining process in order to discover patterns and build rules. These tools provide a helping hand when developing autonomous decision mechanisms. In this context, it can be found both proprietary and open source solutions, but more emphasis will be put on open source tools.
Open source tools like RapidMiner [13] or Weka [14] provide a vast list of data mining and machine learning techniques that might be used together with other applications. Those tools provide libraries that can be imported to custom programs and are referenced in credit evaluation research papers. Due to the interest in neural networks in this paper, it is also mentioned Encong [15], a comprehensive framework for neural networks. To evaluate evolution algorithms for data mining based, KEEL, is also mentioned. This tool allows to evaluate different evolution algorithms as well as to integrate them with other software tools [6].
There are, also, commercial frameworks for neural networks as, for instance, NeuralSolution [16], providing a complete framework for neural networks usage.
In this paper, the Weka Toolkit [14] was used to perform the tests and build the system as well as the evaluation of the algorithms proposed. This decision was, mainly, due to the fact that this framework has a large collection of machine learning algorithms for data mining, available in JAVA. Weka has, also, an active support community and their program is released as free open source software.

3 Problem Description

The problem presented in this paper concerns a client classification system where the objective is to improve available classification models based on artificial neural networks.
The system may use data and information from past events to build an updatable decision mechanism in order to learn new trends from new data in an autonomous manner. Moreover a suggestion model must be developed to provide explanation on why loans are accepted or rejected, providing information to help clients improve their current scoring in the system. The main objective of this mechanism is to indicate which characteristics are desired in clients to grant them with a loan application, even when considering that some client attributes may be immutable. The suggestive system may also be used to investigate client types and promote new financial products and services.
In this work a dataset related to credit scoring was chosen from the UCI repository [17]. The choice fell upon a German credit dataset, where each client is characterized by a set of 20 attributes, followed by the classification of each customer.
The dataset itself, presented in table 1, is a combination of personal, social and financial information about past bank clients.

Table 1 Dataset attributes

Number	Attribute	Number	Attribute	Number	Attribute
1	Status	8	Installment rate	15	Housing
2	Duration	9	Personal status	16	Existing credits
3	Credit History	10	Debtors	17	Job
4	Purpose	11	Residence	18	Liable people
5	Credit amount	12	Property	19	Telephone
6	Savings	13	Age	20	Foreign worker
7	Employment duration	14	Installment plans	21	Classification

4 Classification Algorithms

In order to analyze the data in the dataset to build a classification algorithm some tests were conducted, with the help of decision trees and neural networks from the Weka Toolkit [14]. An optimization on these classifiers was made and the results are shown in comparison with our tests and work from previous authors.

4.1 Multilayer Perceptron

The Multilayer Perceptron is an algorithm that uses a feed forward neural network with back propagation to classify instances. In this network a variable number of hidden layers can be used with a different number of neurons. Each neuron has a weight attributed to him and uses also a nonlinear activation function which was developed to model the frequency of action potentials of biological neurons in a brain. The most common activation functions are sigmoid and they are used in this algorithm. Another interesting property of this type of neural network is that there no connections between neuron in the same layer, however neurons are fully connected between layers and it often used more than 3 layers in the network. The back propagation learning algorithm changes the weights in each neuron after each instance of a dataset is processed based on the amount of error in the output compared to the expected result

4.2 Feature Selection Algorithm

The proposed feature selection algorithm in this paper uses decision trees and their properties to select some relevant attributes in a given dataset.
The assumption for the basis of this algorithm is that decision trees consider the best set of attributes for the upper branches in a decision tree. From this information two feature selection algorithms are proposed. Both of them use the J48 classifier from the Weka Toolkit [14] with a confidence factor of 0.25 to produce a decision tree from the dataset.
The first algorithm chooses all the attributes presented in such decision tree as important and delivers the set. Not all attributes from a dataset may be presented in a decision tree and those who are not can be considered as less important in the process of classifying instances.
The second algorithm aims to get a reduced list of the most relevant set of attributes from in a dataset. As a consequence, only the attributes that are placed in the upper levels of the decision tree are considered. In this case all attributes presented in the first three levels of a decision tree are selected and returned as the most important algorithms in the given dataset.

4.3 Neural Networks with Feature Selection

From the feature selection proposals in section 4.2 some approaches are now considered to implement feature selection upon neural networks, making them aware of relevant attributes to which special consideration should be given. To accomplish feature selection upon neural network, two approaches will be considered.

The first approach uses the first selection algorithm in section 4.2. The data is then filtered and the attributes not featuring in the feature selection set are eliminated from the dataset. With the new dataset we present it to the neural network and train it with the modified and normalized dataset.
The second approach uses the second feature selection algorithm presented in the section 4.2. With the given attributes from the feature selection algorithm a special dataset normalization is performed. The attributes indicated from the feature selection are normalized within a range from 0 to 2 and all the other attributes are normalized within a range from 0 to 1. Neural networks are very sensible to the input data and normalizing the dataset in different ways will led the network to pay more attention to the values with greater amplitude.

4.4 Analysis

With the dataset used in this project, a number of tests were made using the algorithms detailed above. Table 2 presents a short summary of the results in terms of correct predictions. All tests were made using the dataset presented in section 3.2 and a test split of 66% for training data and 33% to evaluate each algorithm.

Table 2 Comparative list of results.

Algorithm	Correct Evaluation (%)	Error (%)
Multilayer Perceptron	73.5	26.5
Multilayer Perceptron with Feature Selection 1	69.7	30.3
Multilayer Perceptron with Feature Selection 2	76.0	24.0
J48	77.6	22.4
Naïve Bayes	75.6	24.6
OneR	72.4	27.6
Selective Bayesian Classifier	76.0	24.0
Combining Feature Selection and Neural Networks for Solving Classification Problems	75.0	25.0

Multilayer Perceptron with Feature Selection 1 represents the first algorithm proposed while Multilayer Perceptron with Feature Selection 2 represented the second algorithm proposed both in section 4.3.The test shows a decrease in the accuracy of the neural network when the first feature selection algorithm is applied which can be explained with the loss of information introduced by the combination of the feature algorithm in the dataset. From this result it is fair to conclude that reducing the dataset may not improve the client classification. The second approach shows improvement in the accuracy of the neural network. The larger range in the selected attributes induces, in the neural network, special attention to such attributes in relation to others leading to better results than the simple multi-

layer algorithm. This later algorithm also performs almost like the Naïve Bayes and J48 in terms of accuracy, however a neural network is easier to update than the other algorithm which require the analysis of all the past data each time they are updated. Furthermore, other authors in their studies, with same dataset, achieved similar results as Multilayer Perceptron with Feature Selection 2 which can be seen in the table comparing it to the last two algorithms Selective Bayesian Classifier and Combining Feature Selection and Neural Networks for Solving Classification Problems.

In Table 3 we see the behavior of some classification algorithms when presented with the full dataset for both training and classification. As the test shows when a case that was initially handled in the financial institution and given for learning the second algorithm proposed in section 4.3 shows a better performance than Naïve Bayes and the J48 algorithms. This leads to the conclusion the proposed algorithm retains information about past cases better than other models and also is less likely to repeat errors when evaluating known client types.

Table 3 Results when all instances in the dataset are used for training and classification

Algorithm	Correct Evaluation (%)	Error (%)
Multilayer Perceptron with Feature Selection 2	97.2	2.8
J48	90.2	9.8
Naïve Bayes	75.6	24.6

5 Suggestive System

5.1 Case Study

Normally, a client will test different scenarios to see which one is more likely to help him have his loan application accepted. A suggestion model might be useful to the client and may also help the financial institution advise their clients on actions they can take to improve their risk assessment by the decision mechanism. Imagining a client to whom a loan application was refused using the present classification model, with a suggestive algorithm he may find a solution for his problem. He would give the system an incomplete set of information of a predetermined set of attributes he cannot change and the system would calculate how changes in the not specified attributes increase his chances to be granted with the loan. These changes could be increasing his credit amount available in the financial in a different account or reduce the amount of the loan by a percentage. Moreover this suggestion mechanism may also help financial institution understand how the decision mechanism classifies clients.

5.2 Proposed algorithm

The classification model used in the suggestive algorithm is Multilayer Perceptron with Feature Selection 2. As it is derived from a neural network the process of building a suggestive mechanism becomes more difficult, since neural networks do not provide any explanation on the results given. The idea is to use genetic algorithms to perform a search in the global space of possible solutions and deliver the positive answers to the client. The algorithm used to search such responses is a set of steps here explained:

- Select each missing client attribute as an gene in a chromosome;
- If not created, randomly create the initial population of chromosomes; otherwise, select the best clients from the set created earlier;
- Apply the selection operator and, in selected pairs of chromosomes, calculate a split point to exchange genes between chromosomes;
- Apply the mutation operator and assign a random value to the gene;
- Join the gene information with the known immutable client attributes and use the multilayer perception with feature selection 2 as the objective function;
- If the maximum time of calculation not exceeded, if there are still negative client classifications or if the number of desired alternatives is not met start, from the beginning; otherwise, the algorithm ends here.

In the credit data system, each individual in the population will be the set of attributes that were not specified by a client. Those attributes are then generated randomly between the space of possible solutions for each attribute type. After the selection and mutation operators the attributes are joined with the immutable client attributes and a classification of each pseudo-client is done, retaining the raw classification value as the client score to select the chromosome population for the next iteration and chose the best classified clients from the possible set. The classification algorithm used in this algorithm, multilayer perceptron with feature selection 2, is supposed to be already trained and to have an initial filter that normalizes the client set of attributes according to the rules created in the training step of the classification algorithm.
When the algorithm reaches the end of a stage, the population selected for the next iteration is the set of chromosomes that achieved better classification from the previous generation or the present modified generation that have a different combination of attributes. This last step assures that the answers to the initial problem are all different.

5.3 Results

Some interesting results came to light when investigating the proprieties of certain types of clients. Simulating an unemployed person, who wants a loan for a new

car valued up to 50000€, it could be seen that in a certain set of conditions a loan application could be accepted by the decision mechanism. These conditions must be, according to the suggestive system, a person with up to 38 years old, with a rented house, the the loan only takes up to 27% of his unemployment allowance, no bad history in previous credits, full payback in up to 74 months, with savings or property in his name and with no liable people. This proves the utility of this algorithm in the client perspective. From the financial institution perspective, it was also possible to understand from the attributes present in the accepted simulations what attributes are more important to accept a loan application from an unemployed person with low risk for the financial institution. Those attribute values are identified through their repetition in the set of client attributes generate by the suggestive algorithm. Although the client attributes generated are different from each other some value of some attributes might not be and here is the information needed to help understand the decision process applied by the classification algorithm. For instance the suggestive system shows that only people with no bad history and savings or property in his name are fixed while other may vary. This simulation proves the usefulness of the developed algorithm and provides answers presented in the initial case study.

6 Conclusion

The algorithms described in this paper provided good results in client classification for loan application in a financial institution. The proposed classification algorithm showed improvements when compared with his normal version and the suggestive algorithm also produced good results evaluating alternatives to client situations. In addition, it was also demonstrated the usefulness of the suggestive algorithm from both the client and financial institution perspectives, allowing the clients to have their loan applications accepted without increasing the risk for financial intuitions.
As a reference for future work different datasets could be used to train the classifier and also different classifiers could be improved in order to have a more comprehensive list to compare performances between each algorithm.

Acknowledgements

The work described in this paper is part of TIARAC -Telematics and Artificial Intelligence in Alternative Conflict Resolution Project (PTDC/JUR/71354/2006), research project supported by FCT (Science & Technology Foundation), Portugal.

References

[1] Corchado E., Herrero A.: Neural visualization of network traffic data for intrusion detection, Applied Soft Computing (2010).

[2] Khatri, V, Brown C. V.: Designing Data Governance. Communications of the Acm, Vol. 53 (2010).

[3] Simic, D., Simic, S.: An approach to efficient business intelligent system for financial prediction. Springer-Verlag (2007).

[4] Vojtek, M., Kocenda, E.: Credit Scoring Methods. Czech Journal of Economics and Finance, (2006).

[5] Ratanamahatana, C. A., Gunopulos, D.: Scaling up the Naïve Bayesian Classifier: Using Decision Trees for Feature Selection. In proceedings of Workshop on Data Cleaning and Preprocessing (2002).

[6] Alcalá-Fdez, J., Sánchez, L., García, S., Jesús, M. J., Ventura, S., Guiu, J. M. G., Otero, J., Romero, C., Bacardit, J., Rivas, V. M., Fernández, J. C., Herrera, F.: KEEL: a software tool to assess evolutionary algorithms for data mining problems. Soft Computing 13(3): 307-318 (2009).

[7] Eletter, S. F., Yaseen, S. G., Elrefae, G. A.: Neuro-Based Artificial Intelligence Model for Loan Decisions. American Journal of Economics and Business Administration, 27-34 (2010).

[8] Islam, Md. S., Zhou, L., Li, F.: Application of Artificial Intelligence (Artificial Neural Network) to Assess Credit Risk: A Predictive Model for Credit Card Scoring. Dissertation, School of Management, Blekinge Institute of Technology (2009).

[9] Huang, J. -J., Tzeng, G. -H., Ong, C. –S.: Two-stage genetic programming (2SGP) for the credit scoring model. Elsevier Inc (2005).

[10] Eggermont, J., Kok, J. N., Kosters, W. A.: Genetic Programming for Data Classification:Partitioning the Search Space. Proceedings of the 2004 Symposium on applied computing (2004).

[11] Madeira, S. C., Oliveira, A. L., Conceição, C. S.: A Data-Mining Approach To Credit Risk Evaluation and Behaviour Scoring. Progress in artificial intelligence: 11th Portuguese Conference on Artificial Intelligence (2003).

[12] O'Dea, P., Griffith, J., O'Riordan, C.: Combining Feature Selection and Neural Networks for Solving Classification Problems. Intelligent exploration of the web, Physica-Verlag GmbH, 389-401 (2003).

[13] RapidMiner: http://rapid-i.com/content/view/181/190/ Accessed in 28/6/2010

[14] Weka: http://www.cs.waikato.ac.nz/ml/weka/ Accessed in 28/6/2010.

[15] Encog: http://www.heatonresearch.com/encog Accessed in 28/6/2010.

[16] NeuroSolutions: http://www.neurosolutions.com/ Accessed in 28/6/2010.

[17] Machine Learning Repository: http://archive.ics.uci.edu/ml/ Accessed in 28/6/2010.