An exploratory analysis of methods for extracting credit risk rules

Patricia Jimbo Santana[1], Augusto Villa Monte[2], Enzo Rucci[2,3]
Laura Lanzarini[2], Aurelio F. Bariviera[4]

[1] Carrera de Contabilidad y Auditoría, Facultad de Ciencias Administrativas,
UCE Universidad Central del Ecuador, Quito – Ecuador
prjimbo@uce.edu.ec
[2]III-LIDI, Fac. de Informática, Universidad Nacional de la Plata,
50 y 120, La Plata, Buenos Aires, Argentina,
{avillamonte, erucci, laural}@lidi.info.unlp.edu.ar
[3] III-LIDI, CONICET, Fac. de Informática, UNLP, La Plata, Argentina
[4]Departament of Business, Universitat Rovira i Virgili,
Avenida de la Universitat,1 Reus, Spain
aurelio.fernandez@urv.cat

Abstract. This paper performs a comparative analysis of two kind of methods for extracting credit risk rules. On one hand we have a set of methods based on the combination of an optimization technique initialized with a neural network. On the other hand there are partition algorithms, based on trees. We show results obtain on two real databases. The main findings are that the set of rules obtained by the first set of methods give a set of rules with a reduced cardinality, with an acceptable precision regarding classification. This is a desirable property for financial institutions, who want to decide credit approval face to face with customers. Bank employees who daily deal with retail customers can be easily trained for selecting the best customers, by using this kind of solutions.

Keywords: credit scoring, classification rules, Learning Vector Quantization (LVQ), Particle Swarm Optimization (PSO).

1 Introduction

The 21st century presents an increase in the development and consumption of goods. The extension of financial services in emerging economies is particularly important. Financial intermediation provides a solution to immediate consumption of durable goods, helping to defer the payment several months or years. This sort of “democratization” in consumption poses a challenge to financial institutions. Whereas mortgage lending applications, due to its comparatively reduced number of borrowers, can be decided at a slower pace, consumer lending needs faster (even instantaneous online) decision procedures. Borrowers want small credits for buying home equipment, a car, a trip, etc. They are eager of a quick answer. From the point of view of the borrowers, they want to receive a quick positive answer to their applications. On the other side, financial institutions want to find the appropriate rules in order to approve credit application only to good borrowers, *i.e.* those who pay back their financial commitments.

Financial institutions typically ask exhaustive information about the potential client: age, marital status, salary, other debts, job type, etc. This information is gathered in order to be analyzed, using some decision model. The result of this analysis is either to grant or reject the credit.

The increasing number of applicants and data raises the necessity for suitable techniques that deals with the complexity of this multidimensional problem. Timely resolution of credit applications is key element when deciding a credit scoring method. Precisely, the area known as data mining can shed light on this kind of situations.

Data mining comprises a set of techniques that are able to model available information. One of the most important stages in the process is knowledge discovery. It is characterized by obtaining new and useful information without assuming prior hypothesis. One of the preferred techniques by decision makers is the association rule. An example of association rule is an expression: IF condition1 THEN condition2, where both conditions are conjunctions of propositions of the form (attribute = value) and whose solely restriction is that attributes in the antecedent must not be present in the consequent. When a set of association rules presents in the consequent, the same attribute is called a set of classification rules (Witten [15], Hernández & Ramírez [7]).

The aim of this paper is to present several alternatives of scoring methods. We believe that the combination of a method for obtaining classification rules and competitive neural network provides an intuitive solution with acceptable levels of errors. Standard methods based on classification trees provide good benchmark for alternative ones. The main advantage of the proposed alternatives is the production of a reduced set of rules, that improves both the transparency and delay in the decision making process of the financial institutions.

The rest of the paper is structured as follows. Section 2 briefly discuses relevant literature on credit risk. Section 3 describes the neural network, metaheuristics, and the proposed method. Section 4 describes data and presents results of a true empirical application and section 5 draws the main implications of our proposal.

2 Brief literature review

One of the oldest papers on bankruptcy prediction is FitzPatrick [5] who, using 13 accounting ratios calculated for 40 firms during three years. In the 1960s, the development of the capital markets in the United States, showed the necessity for more scientific models to assess economic corporate strength. Consequently, the first z-score model by Altman [2] was developed. At that time, the main concern of banks was to classify corporations according to their credit risk, since they were the main clients. However, in the last decades, there has been an increase in consumer credit. Retail banking became a growing industry. Not only there has been a boom in credit card memberships, especially in emerging economies, but also an increase in small consumption credits.

There are several methods to construct rules in order to evaluate the creditworthiness of credit applicants. Computational intelligent techniques produce, explointing exhaustive credit databases can obtain better results, by capturing subtle characteristics of customers. These techniques, without being exhaustive, include artificial neural networks, fuzzy set theory, decision trees, support vector machines, genetic algorithms, among others. Artificial neural networks is a family of neural networks with different architectures. These architectures include popular models such as back propagation networks, self-organizing maps and learning vector quantization. Decision trees transform data in a tree-shape structure of leaf and decision nodes, and the goal is to test attributes to each branch of the tree, that constitutes a class. Support vector machines search an optimal hyperplane in order to generate a binary classification, maximizing the margin of separation between classes. Genetic algorithms are a set of methods to optimized problems, based on the evolutionary idea of natural selection.

If the goal is to obtain association rules, the a priori method (Agrawal and Srikant [1]) or some of its variants could be used. This method identifies the most common sets of attributes and then combines them to get the rules. There are variants of the a priori method, are usually oriented reduce computation time.

Under the topic classification rules, the literature contains various construction methods based on trees such as C4.5 (Quinlan [14]) or clipped trees as the PART method (Frank and Witten [7]). In both cases, the key is to get a set of rules that covers the examples fulfilling a preset error bound. The methods of construction rules from trees are partitives and are based on different attributes' metrics to assess its ability to cover the error bound.

The original PSO method defined in Kennedy and Eberhart [8] was extended in Lanzarini et al. [10] and [12] in order to obtain classification rules. This extension was later applied to two public databases of credits, in order to obtain a set of rules with low cardinality.

Brown and Mues [4] compared several techniques that can be used for imbalanced credit scoring data sets. Imbalance is a typical feature of credit data sets: in healthy financial institutions the number of defaulting loans is much lower than good performing loans. As a consequence the two classes could be not evenly represented search space. Blanco et al. [3] implements credit scoring modles based on multilayers percepton approach, and benchmark the performance with linear and quadratic discriminant analysis, using a small sample of a microfinance institution in Peru. They find that neural networks based models outperforms classical discriminant methods.

4 Data and Results

We test alternative methods in two real databases and two consumer credit financial data from UCI Machine Learning Repository [13]. One of the real databases comes from an important savings and credit institution (Banco Solidario) of Ecuador with more than 20 yeas of trajectory in the domestic market. This data comprises credit operations between 2011 until August 2014, with the following attributes: status; date of application; branch; province; requested amount; authorized amount; purpose of the credit; cash, bank accounts, investments, other assets, liabilities and salary of the applicant; date of verification of information; date of authorization; approval/denial date; cash, bank accounts, investments, other assets, liabilities and salary of the applicants' partner. In case, the applicant is a small business data requested are revenues and expenses of the business. The 'status' variable correspond to the situation of the credit. Applications can be denied or accepted. In case of being accepted, the status is classified between credits that were duly repaid and those with some delay in the payback. In turn, overdue loans are classified, according to the credit procedures manual between those with less than 90 days overdue, and those with more than 90 days overdue (initiation of legal actions). The other real database is from a mutual savings institution of Ecuador, with the same variables described above, with operations between 2011 and 2015.

Using the data described above, we compare the performance of several competing methods that combined a fixed and variable population PSO, initialized with two competitive neural networks (LVQ and SOM [9]). We compare these solutions with C4.5 methods defined by Quinlan [14] and PART defined by Frank and Witten [6]. The way of finding classification rules in proposed and control methods is different. On one hand rules discovering is done after searching task. On the other hand, control methods obtain rules, based in a partition strategy. C4.5 is a pruned tree whose branches are mutually exclusive and allow classifying examples. PART gives as a result a list of rules equivalent to those generated by the proposed classification method, but in a deterministic way. PART operation is based on the construction of partial trees. Each tree is created in a similar manner to that proposed for C4.5 but during the process construction errors of each branch are calculated. These errors allow the selection of the most suitable combinations of attributes.

We performed 30 independent runs of each method. For fixed population PSO, we use a competitive network of 30 neurons, whereas for the variable population case, the size begins with 9 neurons. PART method was executed with a confidence factor of 0.3 for the pruned tree. For other parameters default values were used.

Tables 1, 2, 3, and 4 summarize the results obtained by applying each method in each database. In each case was considered not only the accuracy of coverage of the rule set, but also the "transparency" of the obtained model. This "transparency" is reflected in the average number of rules obtained and the average number of terms used to form the antecedent.

The most important feature of our results, is that the combination of a search algorithm with a competitive neural network, gives a set of rules with a significant low cardinality, *vis-à-vis* the partition algorithms. Although partition algorithms provide more accuracy, this is at expense of a much larger number of rules. In fact, the difference in accuracy between both types of methods is within the range of 1 to 3 percentage points. We have to highlight that the accuracy of the classification based on PSO is very good and comparable to the other methods. However, regarding the number of rules is between 10 and 20 times larger in partition methods. Consequently, there is some sort of trade-off between simplicity and accuracy. Given that credit rules should be simple, in order to give customers a quick answer (for example, in consumer online credits), we believe that competitive search based methods are a good alternative to partition methods.

Table 1. Results on Australian database

Method		TRUE +	TRUE -	False +	False -	Precision	#rules	length antecedent
SOM + PSO	**Mean**	**0.4257**	**0.4333**	**0.1097**	**0.0309**	**0.8590**	**3.0167**	**1.3711**
	sd	0.0154	0.0103	0.0069	0.0066	0.0099	0.0461	0.1922
SOM + varPSO	**Mean**	**0.4183**	**0.4391**	**0.1071**	**0.0351**	**0.8574**	**3.0000**	**1.5178**
	sd.	0.0132	0.0158	0.0130	0.0077	0.0104	0.0000	0.1085
LVQ + PSO	**Mean**	**0.4201**	**0.4414**	**0.1079**	**0.0306**	**0.8614**	**3.0000**	**1.2667**
	sd	0.0179	0.0172	0.0093	0.0065	0.0105	0.0000	0.1207
LVQ + varPSO	**Mean**	**0.4199**	**0.4382**	**0.1054**	**0.0363**	**0.8582**	**3.0000**	**1.5578**
	sd	0.0179	0.0172	0.0075	0.0073	0.0092	0.0000	0.1336
C4.5	**Mean**	**0.3910**	**0.4618**	**0.0847**	**0.0625**	**0.8528**	**18.2200**	**4.8394**
	sd	0.0121	0.0063	0.0066	0.0120	0.0124	2.0825	0.2810
PART	**Mean**	**0.3564**	**0.3906**	**0.1562**	**0.0969**	**0.7469**	**33.3433**	**2.4926**
	sd	0.0136	0.0288	0.0289	0.0134	0.0292	1.5793	0.0934

Table 2. Results on German database

Method		TRUE BAD	TRUE GOOD	False BAD	False GOOD	Precision	#rules	length antecedent
SOM + PSO	**Mean**	**0.1026**	**0.5993**	**0.0984**	**0.1994**	**0.7019**	**8.4400**	**2.1619**
	sd	0.0123	0.0183	0.0149	0.0183	0.0153	0.6009	0.1415
SOM + varPSO	**Mean**	**0.1046**	**0.5954**	**0.1034**	**0.1965**	**0.7000**	**8.0233**	**2.0464**
	sd	0.0115	0.0135	0.0110	0.0134	0.0162	0.6745	0.1030
LVQ + PSO	**Mean**	**0.0999**	**0.5997**	**0.1017**	**0.1986**	**0.6996**	**8.7767**	**2.1802**
	sd	0.0151	0.0171	0.0136	0.0153	0.0133	0.7224	0.1075
LVQ + varPSO	**Mean**	**0.1089**	**0.5973**	**0.1057**	**0.1880**	**0.7063**	**8.8867**	**2.0884**
	sd	0.0100	0.0129	0.0103	0.0116	0.0109	0.4918	0.0960
C4.5	**Mean**	**0.1219**	**0.5894**	**0.1106**	**0.1781**	**0.7113**	**86.4600**	**5.6267**
	sd	0.0069	0.0070	0.0070	0.0069	0.0079	4.0788	0.1382
PART	**Mean**	0.1404	0.4385	0.1687	0.2258	0.6967	70.9133	3.0138
	sd	0.0120	0.0091	0.0135	0.0170	0.0139	2.1575	0.0561

Table 3. Results on Cooperativa de Crédito database

Method		TRUE N	TRUE O	False N	False O	Precision	#rules	length antecedent
SOM + PSO	**Mean**	**0.6242**	**0.1601**	**0.1253**	**0.0898**	**0.7844**	**3.7867**	**1.6375**
	sd	0.0069	0.0062	0.0057	0.0059	0.0059	0.2980	0.2151
SOM + varPSO	**Mean**	**0.6014**	**0.1914**	**0.0947**	**0.1125**	**0.7928**	**4.1533**	**1.6953**
	sd	0.0052	0.0059	0.0057	0.0047	0.0030	0.2801	0.0867
LVQ + PSO	**Mean**	**0.6227**	**0.1671**	**0.1191**	**0.0910**	**0.7899**	**3.2933**	**1.4021**
	sd	0.0048	0.0055	0.0051	0.0039	0.0031	0.1837	0.1066
LVQ + varPSO	**Mean**	**0.6029**	**0.1902**	**0.0956**	**0.1114**	**0.7930**	**4.3733**	**1.6553**
	sd	0.0056	0.0055	0.0054	0.0053	0.0025	0.2625	0.0567
C4.5	**Mean**	**0.6320**	**0.1786**	**0.1075**	**0.0819**	**0.8106**	**114.2600**	**9.6762**
	sd	0.0014	0.0013	0.0013	0.0013	0.0011	6.0543	0.1144
PART	**Mean**	**0.6229**	**0.1825**	**0.1036**	**0.0910**	**0.8054**	**42.3567**	**4.6956**
	sd	0.0065	0.0064	0.0064	0.0065	0.0023	2.1661	0.0880

Table 4. Results on Solidario database

Method		TRUE N	TRUE O	False N	False O	Precision	#rules	length antecedent
SOM + PSO	**Mean**	**0.0457**	**0.8863**	**0.0228**	**0.0451**	**0.9320**	**4.3967**	**5.4962**
	sd	0.0058	0.0047	0.0045	0.0056	0.0050	0.4895	0.4221
SOM + varPSO	**Mean**	**0.0609**	**0.8870**	**0.0210**	**0.0300**	**0.9480**	**3.8600**	**2.8940**
	sd	0.0029	0.0059	0.0051	0.0030	0.0063	0.2415	0.3532
LVQ + PSO	**Mean**	**0.0482**	**0.8870**	**0.0219**	**0.0428**	**0.9352**	**4.6067**	**5.9166**
	sd	0.0052	0.0056	0.0055	0.0049	0.0054	0.4193	0.2771
LVQ + varPSO	**Mean**	**0.0565**	**0.8882**	**0.0198**	**0.0346**	**0.9447**	**3.8533**	**3.1013**
	sd	0.0043	0.0062	0.0056	0.0043	0.0056	0.2921	0.3395
C4.5	**Mean**	**0.0762**	**0.9017**	**0.0073**	**0.0148**	**0.9779**	**153.5733**	**11.2349**
	sd	0.0003	0.0003	0.0003	0.0003	0.0003	5.1687	0.1565
PART	**Mean**	**0.0749**	**0.9013**	**0.0077**	**0.0161**	**0.9762**	**80.9400**	**4.7650**
	sd	0.0010	0.0014	0.0014	0.0010	0.0008	2.2034	0.0688

5 Conclusions

We compute several variations of a competing method for credit scoring using a variation of PSO (fixed and variable population), and a neural network. We test our model on two actual credit databases from important retail credit institutions from Ecuador and from two public databases from a repository. Results show that search algorithms allow to reduce significantly the number of rules, required to reach an acceptable and very similar level of classification accuracy.

Future research lines is to explore the incidence of the initial settings such as the speed of growth in population, which helps to determine the antecedent of rules. Results show no significant difference between fixed and variable population PSO. This means that the exploration of the solution space is not satisfactory solved.

Finally, we would like to highlight that the goal of our work is to achieve an intuitive model for credit scoring with a comparable accuracy to popular benchmark models. Our results suggest that the simplification of decision rules generates transparency in credit scoring, which could improve the reputation of financial institutions.

References

1. Agrawal, R.,Srikant, R.,. Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB '94, pp. 487–499. Morgan Kaufmann Publishers Inc., San Francisco (1994).
2. Altman, E.I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), pp.589–609.
3. Blanco, A., Pino-Mejías, R. Lara, J. Rayo, S.: Credit scoring models for the micro-finance industry using neural networks: evidence from Peru, Exp. Syst. Appl. 40 356–364 (2013).
4. Brown, I., Mues, C.: An experimental comparison of classification algorithms for imbalanced credit scoring data sets, Exp. Syst. Appl. 39, 3446–3453 (2012).
5. FitzPatrick, P.J., (1932). A comparison of the ratios of successful industrial enterprises with those of failed companies. The Certified Public Accountant, Oct., Nov., Dec.
6. Frank, E., Witten, I. H., 1998. Generating accurate rule sets without global optimization. In: Proceedings of the Fifteenth International Conference on Machine Learning, ICML '98., pp. 144–151. Morgan Kaufmann Publishers Inc., San Francisco.
7. Hernández Orallo, J., Ramírez Quintana, M.J., Ferri Ramírez, C., 2004. Introducción a la Minería de Datos. 1ra Edición. Pearson.
8. Kennedy, J. & Eberhart, R., 1995. Particle swarm optimization. In, Proceedings of IEEE International Conference on Neural Networks. pp. 1942–1948 vol.4.
9. Kohonen, T. Self-Organizing Maps. Volume 30, Springer Series in Information Sciences. Springer, Heidelberg (2012).
10. Lanzarini, L., Villa Monte, A., Aquino, G., De Giusti, A.: Obtaining classification rules using lvqPSO Advances in Swarm and Computational Intelligence. Lecture Notes in Computer Science. Vol 6433, 183-193. Heidelberg: Springer-Verlag Berlin (2015).
11. Lanzarini, L., Villa Monte, A.; Bariviera, A.F., Jimbo Santana, P.: Obtaining Classification Rules Using LVQ+PSO: An Application to Credit Risk. In J. Gil-Aluja et al., eds. Scientific Methods for the Treatment of Uncertainty in Social Sciences. Advances in Intelligent Systems and Computing. Springer International Publishing, 383–391 (2015).

12. Lanzarini, L., Villa-Monte, A., Ronchetti, F.: SOM+PSO. A Novel Method to Obtain Classification Rules. Journal of Computer Science & Technology (JCS&T), 15(1), 15-22 (2015).
13. Lichman, M.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA. http://archive.ics.uci.edu/ml. Accessed 5 Jan 2015 (2013)
14. Quinlan, J.R.: C4.5: programs for machine learning, Morgan Kaufmann Publishers (1993).
15. Witten, I.H., Eibe, F. & Hall, M.A.: Data Mining Practical Machine Learning Tools and Techniques 3rd. ed., San Francisco, CA: Morgan Kaufmann Publishers Inc. (2011).