CREDIT SCORING MODEL BASED ON THE AFFINITY SET

Jerzy Michnik
Department of Operations Research, The Karol Adamiecki University of Economics in Katowice
ul. Bogucicka 14, 40-226 Katowice, Poland

Anna Michnik
Department of Medical Physics, University of Silesia, ul. Uniwersytecka 4, 40-007 Katowice, Poland

Berenika Pietuch
ul. Polna 26/32, 41-600 Swietochlowice, Poland

Keywords: Affinity set, credit score, data mining, logistic regression, rough sets.

Abstract: The significant development of credit industry led to growing interest in sophisticated methods which can support making more accurate and more rapid credit decisions. The parametric statistical methods such as linear discriminant analysis and logistic regression were soon followed up by nonparametrical methods and other techniques: neural networks, decision trees, and genetic algorithms.
This paper investigates the affinity set – a new concept in data mining field. The affinity set model was applied to credit applications database from Poland. The results are compared to those received by Rosetta (the rough sets and genetic algorithm procedure) and logistic regression.

1 INTRODUCTION

The affinity set (Chen and Larbani, 2006; Larbani and Chen, 2006) is inspired from the vague interaction between people in social sciences and developed as the data mining tool to classify, analyze, and build the relationship between observed outcomes (consequences) and possible incomes (causes) of an information system.

For several last decades the growth of the credit industry is observed. The large loan portfolios motivated the industry to develop the more accurate scoring models. It is expected that good scoring model can significantly improve the future cash flows of credit institution. The main goal of credit scoring models is to classify the objects (loan applications) into the two classes: granted and denied.

At the beginning simple parametric statistical models like linear discriminant analysis were employed. Linear discriminant analysis is still very popular, however has been questioned because the credit data do not fulfill the strict assumptions underlying this method . Logistic regression has appeared to be a powerful parametric statistical method when applied to credit databases. Several other techniques like *k* nearest neighbor, classification trees and neural network models have been developed.

Five neural network architectures has been investigated by West: the traditional MLP network, mixture of experts (MOE), radial basis function (RBF), learning vector quantization (LVQ), and fuzzy adaptive resonance (FAR) (West, 2000). 10-fold cross-validation method of testing was performed with two real world data sets: Australian and German. Logistic regression has been found to be the most accurate of the traditional methods. Comparable accuracy have been achieved by two neural network models: the mixture-of-experts and radial basis function. The results of Lee at al. revealed that classification and regression tree (CART) and multivariate adaptive regression splines (MARS) outperformed traditional discriminant analysis, logistic regression, neural networks, and support vector machine (SVM) approaches in terms of credit scoring accuracy (Lee et al., 2006).

To improve the performance of credit scoring, the mixed methods have been also proposed. Integrating the backpropagation neural networks with traditional discriminant analysis approach allowed to simplify the network structure and improved the credit scoring

Michnik J., Michnik A. and Pietuch B. (2008).
CREDIT SCORING MODEL BASED ON THE AFFINITY SET.
In *Proceedings of the Tenth International Conference on Enterprise Information Systems*, pages 287-290
DOI: 10.5220/0001723902870290

accuracy (Lee et al., 2002). Lee and Chen reported that a two-stage hybrid modeling procedure with artificial neural networks and MARS accomplished better results then single method such as discriminant analysis, logistic regression, artificial neural networks or MARS (Lee and Chen, 2005).

A wide study of different approaches was presented recently (Xiao et al., 2006). The authors compared the efficiency of the classical methods (e.g. linear discriminant analysis, logistic regression, neural networks and k-nearest neighbor) and some recently developed advanced data mining techniques such as SVM, CART, and MARS. They tested all the chosen methods on German, US and Australian credit data sets. The experiment results show that SVM, MARS, logistic regression and neural networks yield a very good performance. However, CART and MARS's explanatory capability outperforms the other methods.

The aim of our research is to test an ability of the concept of affinity set to become the tool for data mining and compare its results with the method basing on the concept of rough sets and logistic regression. We applied those selected methods to the database which contained the individual credit applications from Poland.

2 AFFINITY CONCEPT IN DATA MINING

2.1 Affinity as a Performance Measure

The starting point of our analysis is database which has a form of several input and output data. We assume that there are k input data and each input data x_i can have several values which belong to the set X_i, consequently $x = (x_1, \ldots, x_k)$ belongs to Cartesian product $X_1 \times \ldots \times X_k = X$. Similarly, there are l output data y_j with values from Y_j, and $(y_1, \ldots, y_l) = y \in Y = Y_1 \times \ldots \times Y_l$. Each case $c = (x,y)$ from database is a member of Cartesian product $X \times Y$.

Any rule, which can explain the underlying relationships in analyzed database can be considered as a relation in the above defined Cartesian product: $R_p \subseteq X \times Y$, where the index p enumerates the rules. We also can separate the input and output part of the rule and write $R_p = (R_p^x, R_p^y)$, so that $R_p^x \subseteq X$ and $R_p^y \subseteq Y$.

We assume that the output values are mutually excluding. It means that some of the rules are contradictory and can not explain the underlying relationships simultaneously; e. g. for $k = 2$ and $l = 1$, the rules (x_1, x_2, y_1) and (x_1, x_2, y_2) are contradictory. This observation leads to the conclusion that an explanatory power of the rule depends not only on that how many times the rule explains correctly cases in database but also on the difference of its frequency and frequency of all contradictory rules[1].

When we want to measure an explanatory power of the rule, we need to take into account both above properties. Consequently, we define the affinity of the rule as

$$A(R_p) = \frac{\max\left\{ f(R_p) - \sum_{q \in C_p} f(R_q), 0 \right\}}{N}, \quad (1)$$

where C_p contains the indices of all rules contradictory to the rule R_p; N is the number of records in database. $f(R_p) = |\{c_i | c_i \in R_p\}|, \quad i = 1 \ldots, N$; $f(R_p)$ is an integer number which measure how many cases match the rule R_p ($|\quad|$ – denotes the cardinality of the set). With the function 'max' in the above definition, only the rule which is matched by the maximal number of cases in the group of competitive rules has positive value of affinity.

Let assume that the output consists of only single binary variable $Y = \{y_1, y_2\}$ and the set of rules S have been selected. We consider the definite input $\bar{x}$. Let $S_{\bar{x}}$ denotes all rules from S which match with input, i.e. $\bar{x} \in R_p^x$. $f(R_p(\bar{x}, y_i))$ represents the number of cases in training set which support output y_i for $i = 1, 2$. We calculate the difference:

$$\Delta f = \sum_{S_{\bar{x}}} f(R_p(\bar{x}, y_1)) - \sum_{S_{\bar{x}}} f(R_p(\bar{x}, y_2)) \quad (2)$$

and decide the output:

$$\Delta f \geq 0 \quad \Rightarrow y_1, \quad (3)$$

$$\Delta f \leq 0 \quad \Rightarrow y_2, \quad (4)$$

$$\Delta f = 0 \quad \text{undecided.} \quad (5)$$

To get the performance of the set of rules S we calculate the ratios of matched, unmatched and undecided outputs against actual outputs from the testing set.

2.2 Testing the Set of Rules

The beginning step is to generate all possible rules from training set. Depending on particular requirement this usually large set can be reduced using the concept of the $a - core(A)$. The $a - core(A)$ represents all elements with affinity greater then or equal to a, so its definition is: $a - core(A) = \{R_p | A(R_p) \geq a\}$. It seems practical to take for further analysis a to be around 1%.

[1]This reasoning is parallel to the concepts of coverage and accuracy of the rule discussed in the information systems theory.

The following procedure for extracting the set of rules S is proposed:

1. Move the rule with greatest affinity from $a-core(A)$ to S. Check the performance of the set S.
2. If the performance of the set S is less then previous one, remove the rule from the set S. If performance is equal or greater then previous, keep the rule in the set S.
3. Unless the $a-core(A)$ is empty: go to the step 1.
4. Stop

It need to be mentioned that the set of rules received by above procedure may not be the minimal one, as there was not any optimization routine built in it.

3 CREDIT SCORE MODEL – CASE STUDY

The individual credit application database consisted initially of over 40 fields and over 12000 records. It covered – among others – employment status, personal information, age, housing, and job. It contained also variables which summarized the process of credit appraisal, such as credit score, result of credit appraisal and credit capability. With the aid of logistic regression method, the 23 variables which were statistically significant and uncorrelated have been selected. The stepwise (forward, conditional) method shown that only 4 variables significantly contributed to the classification power of the model and those were chosen for subsequent analysis (the variable coding is given in parenthesis).

Input Variables

x_1 – Number of Debtors: 1 (1), 2 (2).

x_2 – Credit Score: below 0 (1), 0-200 (2), 200-400 (3), 400-600 (4), 600-800 (5), 800-1000 (6), above 1000 (7).

x_3 – Result of Loan Application Appraisal: Error (1), To Clarify (2), Negative (3), Positive (4), Not Counted (5), Application for Too Large Loan Value (6).

x_4 – Credit Capability: No (1), Yes (2).

Output Variable

y – Credit Decision: Denied (1), Granted (2).

The initial step database contained 12711 randomly ordered records (with no missing data). First 5000 records were used as a training set, the rest as the testing set. For almost 62% cases in the testing set the credit decision was positive. The number of different

Table 1: 22 rules with the greatest affinity ('$*$' means that the variable is not counted in the rule).

No.	x	y	$f(R_p)$	$A(R_p)$
1	**42	2	2668	0,4636
2	1*42	2	2640	0,4592
3	**4*	2	2714	0,3934
4	1*4*	2	2645	0,3822
5	*542	2	1966	0,3398
6	1542	2	1943	0,3364
7	**3*	1	1496	0,2972
8	1*3*	1	1493	0,2966
9	**32	1	1441	0,2862
10	1*32	1	1441	0,2862
11	*54*	2	2007	0,2834
12	154*	2	1945	0,2734
13	*53*	1	1044	0,2078
14	153*	1	1041	0,2072
15	*532	1	1003	0,1996
16	1532	1	1003	0,1996
17	***2	2	2681	0,1718
18	1**2	2	2653	0,1674
19	*5*2	2	1973	0,1362
20	15*2	2	1950	0,1328
21	1**1	1	442	0,0874
22	***1	1	452	0,0812

rules which take into account at least one variable x_i were equal 1006.

The procedure described in Sec. 2.2 have been applied to the $0.01-core(A)$ which contained 62 rules. The rule with the greatest affinity $(**42,2)$ explained correctly 31.40% of cases, incorrectly 4.68%, and left unclassified 63.92%. The second rule did not changed the result. The third rule improved result to: [corr: 31.80%, incorr: 7.21% and unclass: 60,99%]. The next improvement appeared for the set of 7 rules with the result: [corr: 53.75%, incorr: 7.43% and unclass: 38,81%]. This result ha been remaining stable until the rule no. 17 which improved the result to: [corr: 80.78%, incorr: 11.27% and unclass: 7,95%]. The rule no. 21 gave [corr: 85,46%, incorr: 12,26%, unclass: 2,28%]. The best performance [corr: 85.96%, incorr: 14.04%, unclass: 0.00%] was reached for the set of 22 rules. Adding 23 rule lowered the performance so it was removed. Continuing the procedure until $a-core(A)$ became empty did not make any further improvement.

For the sake of comparison we employed also the "Rosetta" system (Øhrn et al., 1994) which is the computer implementation of rough set modeling to knowledge discovery and data mining. The genetic algorithm implemented in Rosetta have been applied to our training set. We have made several runs of the genetic algorithm changing the boundary region thin-

ning. The received set of rules was then applied to the testing set with the aid of 'Batch Classifier' tool. The best result, which we gained (with the set containing 10 rules, was: 63.42% cases classified correctly, 6.34% incorrectly and 30.24% unclassified.

Logistic regression model developed on the training set and applied to test set gained 86.47% of correctness. This result coincide with other investigations which report that logistic regression belongs to the most efficient methods in the field (West, 2000; Xiao et al., 2006).

Let denote as p_{12} the percentage of actually denied applications misclassified into the granted group and as p_{21} the percentage of actually granted applications misclassified into denied group. It can be observed that in all three models p_{12} is much greater then p_{21}. The Rosetta's total misclassified 6.34% cases was divided as follows: $p_{12} = 5.02\%$ and $p_{21} = 1.32\%$. For the affinity set model: $p_{12} = 11.05\%$ and $p_{21} = 3.00\%$. Similarly, the logistic regression shows $p_{12} = 13.16\%$ and $p_{21} = 0.37\%$.

In credit scoring applications, it is generally believed that the costs of granting credit to a bad candidate is significantly greater than the cost of denying credit to a good candidate. As Rosetta model has the lowest p_{12}, it's results might get better score if the pure classification rate has been substituted by a kind of cost analysis. On the other hand, logistic regression has the highest value of p_{12} which can lower its score. The problem is – that in contrast to full classification reached by affinity and logistic regression – Rosetta left the great amount (30.24%) of cases unclassified. This fact makes difficult exact calculations and drawing well-founded conclusions.

4 CONCLUSIONS

Our results shows that the affinity measure defined in eq.(1) and followed by extracting procedure described in Sec. 2.2 is able to gain the promising results in data mining. The method is rather simple in comparison to other approaches. It is also low demanding (no preliminary assumptions and low demand for computing resources). We received much higher classification rate then rough sets and genetic algorithm implemented in Rosetta software. The performance of our concept was very close to the level gained by logistic regression model which was reported as one of the best in the field of credit scoring.

To confirm the promising efficiency of the proposed method further studies should be carried out. First of all, there is a need for:

- comparison with the neural network models and other highly efficient modern methods of data mining,
- checking the results with other credit databases.

In this paper the outcomes predicted by the model were confronted with actual credit decisions. It would be also interesting to check the model predictions with actual credit performance.

Our model was tested on credit application database but can be applied equally well to medical, marketing, managerial and other databases. The model offers meaningful adaptability and several experiments with various technical modifications can be made.

REFERENCES

Chen, Y. and Larbani, M. (2006). Developing the affinity set and its applications. In *Proceeding of the Distinguished Scholar Workshop by National Science Council, Jul. 14-18, 2006, Taiwan.* National Science Council, Taiwan.

Larbani, M. and Chen, Y. (2006). Affinity set and its applications. In *Proceeding of the International Workshop on Multiple Criteria Decision Making, Apr. 14-18, 2007, Poland*. Publisher of The Karol Adamiecki University of Economics in Katowice.

Lee, T.-S. and Chen, I.-F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, 28:743752.

Lee, T.-S., Chiu, C.-C., Chou, Y.-C., and Lu, C.-J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*, 50:11131130.

Lee, T.-S., Chiu, C.-C., Lu, C.-J., and Chen, I.-F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, 23:245254.

Øhrn, A., Komorowski, J., Skowron, A., and Synak, P. (1994). The design and implementation of a knowledge discovery toolkit based on rough sets: The rosetta system. In Polkowski, L. and Skowron, A., editors, *Rough Sets in Knowledge Discovery 1: Methodology and Applications*, volume 18 of *Studies in Fuzziness and Soft Computing*, chapter 19, page 376. Physica-Verlag, Heidelberg, Germany.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, 27:1131–1152.

Xiao, W., Zhao, X., and Fei, Q. (2006). A comparative study of data mining methods in consumer loans credit scoring management. *J. Syst. Sci. Syst. Eng.*, 15(4):419–435.