Assessment of Financial Risk Prediction Models with Multi-Criteria Decision Making Methods

J. Salvador Sánchez[1], Vicente García[1], and Ana I. Marqués[2]

[1] Dep. Computer Languages and Systems – Institute of New Imaging Technologies
[2] Dep. Business Administration and Marketing
Universitat Jaume I
Av. Vicent Sos Baynat s/n, 12071 Castelló de la Plana (Spain)

Abstract. A wide range of classification models have been explored for financial risk prediction, but conclusions on which technique behaves better may vary when different performance evaluation measures are employed. Accordingly, this paper proposes the use of multiple criteria decision making tools in order to give a ranking of algorithms. More specifically, the selection of the most appropriate credit risk prediction method is here modeled as a multi-criteria decision making problem that involves a number of performance measures (criteria) and classification techniques (alternatives). An empirical study is carried out to evaluate the performance of ten algorithms over six real-life credit risk data sets. The results reveal that the use of a unique performance measure may lead to unreliable conclusions, whereas this situation can be overcome by the application of multi-criteria decision making techniques.

Key-Words: Data mining; Multi-criteria decision making; Credit risk prediction

1 Introduction

The recent international financial crisis has aroused increasing attention of financial institutions on credit and operational risk assessment, converting this into a key task because of the heavy losses associated with wrong decisions. One major risk for banks and financial institutions comes from the difficulty to distinguish the creditworthy applicants from those who will probably default on repayments. The decision to grant credit to an applicant was traditionally based upon subjective judgments made by human experts, using past experiences and some guiding principles.

In this context, credit scoring and behavioral management have emerged as more formal and accurate methods to assess credit risk, improve cash flow, reduce possible financial risks and make managerial decisions [15]. Credit scoring is essentially a set of objective risk assessment tools that help lenders discriminate between “good” and “bad” loan applicants, depending on how likely they are to default with their repayments.

The most classical approaches to credit scoring are based upon statistical and operations research models, such as logistic regression, probit analysis and discriminant analysis. However, the problem with applying statistical techniques is that some assumptions, such as the multivariate normality assumptions for independent variables, are frequently violated in practice, what makes them theoretically invalid for finite samples [8]. During the last decades, efforts have focused on the deployment of soft computing techniques to design and implement credit scoring solutions. In contrast with statistical models, soft computing methods do not assume any specific prior knowledge, but automatically extract information from the training examples available.

From a practical point of view, the credit scoring problem basically lies in the domain of binary classification where a new input sample (the credit applicant) must be categorized into one of the predefined classes based on a number of observed variables related to that sample. The input consists of a variety of information that describes socio-demographic characteristics and economic conditions of the applicant, and then the classifier has to produce the output in terms of the applicant creditworthiness. In its most usual form, credit risk prediction aims at assigning credit applicants to either good (those who are liable to reimburse the financial obligation) or bad (those who should be denied credit because of the high probability of defaulting on repayments). The credit risk prediction problem can be formally described as follows. Given a data set of m past applicants $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, where each applicant x_i is characterized by D attributes, $x_{i1}, x_{i2}, \ldots x_{iD}$, and y_i denotes the type of applicant (for example, good or bad), then the task of a credit risk prediction model f is to predict the value y for a new applicant $\mathbf{x}$, that is, $f(\mathbf{x}) = y$.

Many researchers have conducted comparative studies of credit risk prediction models, but their conclusions may vary depending on which performance measure they have used. For example, Desai et al. [6] concluded that customized neural networks perform better than linear models when measuring the percentage of bad applicants correctly classified, whereas logistic regression yields better results in terms of percentage of good and bad applicants correctly classified. Yobas et al. [17] found that linear discriminant analysis outperforms neural networks, genetic algorithms and decision trees in the proportion of samples correctly classified. Baesens et al. [3] showed that the support vector machines achieve the highest accuracy rate, while the neural networks perform the best in terms of the area under the ROC curve. Bensic et al. [4] suggested that the accuracy of probabilistic neural network is superior to that of logistic regression, CART decision trees, radial basis function, multi-layer perceptron and learning vector quantization. Antonakis and Sfakianakis [2] evaluated the performance of k-nearest neighbors decision rule, multi-layer perceptron, decision trees, logistic regression, linear discriminant analysis and naïve Bayes, showing that the k-nearest neighbors rule achieved the highest accuracy and the neural network was the best method in terms of the Gini coefficient. Wang [16] found that stacking and bagging using a decision tree as base classifier achieve the best performance in terms of accuracy, type-I error and type-II error.

2 Evaluation Criteria in Credit Risk Prediction Problems

Standard performance evaluation criteria in the field of credit soring include accuracy, Gini coefficient, Kolmogorov-Smirnov statistic, root mean squared error, area under the ROC curve, geometric mean of accuracies, and type-I and type-II errors [1, 7, 15]. For a two-class problem, most of these metrics can be easily derived from a 2×2 confusion matrix, where each entry (i, j) contains the number of correct/incorrect predictions.

Many credit risk systems often use the accuracy to evaluate the performance of the prediction models. It represents the proportion of the correctly predicted cases on a particular data set. However, empirical and theoretical evidences show that this measure is strongly biased with respect to data imbalance and proportions of correct and incorrect predictions. Because credit data are commonly imbalanced, the area under the ROC curve (AUC) has been suggested as an appropriate evaluator without regard to class distribution or misclassification costs [3, 11]. The AUC for a binary problem can be defined as the arithmetic average of the mean predictions for each class [14]:

$$AUC = \frac{sensitivity + specificity}{2} \tag{1}$$

where the sensitivity or true positive rate (TPrate) measures the percentage of good applicants that have been predicted correctly, and the specificity or true negative rate (TNrate) corresponds to the percentage of bad applicants predicted as bad.

Another measure often used in skewed domains, such as credit risk prediction, is the geometric mean of accuracies. The idea behind this metric is to maximize the accuracy on each class while keeping them balanced. It punishes those models that produce big disparities between both accuracies.

$$Gmean = \sqrt{sensitivity \cdot specificity} \tag{2}$$

The root mean squared error (RMSE) is another common performance measure used in credit risk prediction problems. Let $p_1, p_2, \ldots, p_m$ and $a_1, a_2, \ldots, a_m$ be the predicted and actual outputs on the test samples, respectively. The RMSE allows to quantify the difference between the predictions and the true labels, measuring the deviation of the classification model from the target value.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(p_i - a_i)^2} \tag{3}$$

3 Multiple Criteria Decision Making

Assessing the performance of classifiers means to take more than one criterion of interest into account, usually weighting this against the gains of other complementary criteria. Under this setting, classifier selection can be modeled as a multiple criteria decision making (MCDM) problem. MCDM tools are analytic methods to evaluate the pros and cons of a set of alternatives based on several criteria, with the aim of making a reliable decision [10]. In the context of financial risk prediction, MCDM methods should allow decision makers to choose the model that achieves an optimal trade-off of the assessment criteria of interest. An MCDM problem can be expressed in the form of a $M \times N$ decision matrix as that given in Table 1.

Various MCDM methods have been proposed in the literature, each one having its own characteristics. Popular examples are TOPSIS and PROMETHEE, among many others. Next, these two methods used in the present paper will be described.

Table 1. Decision matrix for an MCDM problem

	C_1	C_1	$\cdots$	C_N
A_1	z_{11}	z_{12}	$\cdots$	z_{1N}
A_2	z_{21}	z_{22}	$\cdots$	z_{2N}
A_M	z_{M1}	z_{M2}	$\cdots$	z_{MN}

3.1 TOPSIS Method

The basic principle behind TOPSIS is to find the best alternative by minimizing the distance to the ideal solution and maximizing the distance to the negative-ideal solution [9]. The ideal solution is formed as a composite of the best performance values exhibited by any alternative for each criterion, whereas the negative-ideal solution is the composite of the worst performance values.

Assuming a problem with M alternatives (prediction models) and N criteria (performance measures), the procedure of TOPSIS can be summarized as follows:

1. Calculate the normalized decision matrix, where the normalized value n_{ij} is calculated as
   $n_{ij} = z_{ij} / \sqrt{\sum_{i=1}^{M} z_{ij}^2} \quad i = 1, \ldots, M \quad j = 1, \ldots, N$
2. Calculate the weighted normalized values $v_{ij} = w_j z_{ij}$, where w_j is the weight of the criterion C_j and $\sum_{j=1}^{N} w_j = 1$.
3. Determine the ideal solution A^+ and the negative-ideal solution A^-
   $A^+ = \{v_1^+, \ldots, v_N^+\} = \{(\max_j v_{ij} | i \in I), (\min_j v_{ij} | i \in J)\}$
   $A^- = \{v_1^-, \ldots, v_N^-\} = \{(\min_j v_{ij} | i \in I), (\max_j v_{ij} | i \in J)\}$
   where I and J are associated with benefit and cost criteria, respectively.
4. Calculate the separation of each alternative from the ideal solution and that from the non-ideal solution using the N-dimensional Euclidean distance
   $d_j^+ = \sqrt{\sum_{j=1}^{N} (v_{ij} - v_j^+)^2}$ and $d_j^- = \sqrt{\sum_{j=1}^{N} (v_{ij} - v_j^-)^2} \quad i = 1, \ldots, M$
5. Calculate the relative closeness to the ideal solution. The relative closeness of the alternative A_i with respect to A^+ is defined as
   $R_i^+ = d_i^- / (d_i^+ + d_i^-) \quad i = 1, \ldots, M$
6. Rank alternatives according to the index R_i^+.

3.2 PROMETHEE Method

The aim of the PROMETHEE method [5] is to rank alternatives based on their values over different criteria. As an outranking technique, it quantifies a ranking through the pairwise comparisons (differences) between the criterion values describing the alternatives.

This MCDM method uses the concept of preference flow: the positive preference flow indicates how an alternative is outranking all the other alternatives, whereas the negative preference flow indicates how an alternative is outranked by the remaining alternatives. The procedure of PROMETHEE can be expressed in a series of steps:

1. For each pair (A_i, A_j) of a finite set of alternatives $A = \{A_1, A_2, \ldots, A_M\}$, calculate aggregated preference indices as follows:
   $\pi(A_i, A_j) = \sum\limits_{k=1}^{N} P_k(A_i, A_j) w_k$
   $\pi(A_j, A_i) = \sum\limits_{k=1}^{N} P_k(A_j, A_i) w_k$
   where w_k is the normalized weight of the criterion C_k. $\pi(A_i, A_j)$ indicates how A_i is preferred to A_j and $\pi(A_j, A_i)$ indicates how A_j is preferred to A_i. $P_k(A_i, A_j)$ and $P_k(A_j, A_i)$ are the preference functions for alternatives A_i and A_j, respectively.
2. Define the positive and the negative preference flows as
   $\phi^+(A_i) = \frac{1}{M-1} \sum\limits_{a \in A} \pi(A_i, a)$
   $\phi^-(A_i) = \frac{1}{M-1} \sum\limits_{a \in A} \pi(a, A_i)$
3. Compute the net preference flow for each alternative as follows:
   $\phi(A_i) = \phi^+(A_i) - \phi^-(A_i)$

When $\phi(A_i) > 0$, A_i is more outranking all the alternatives on all the evaluation criteria. Conversely, when $\phi(A_i) < 0$, the alternative A_i is more outranked.

4 Experimental Setup

The aim of the experiments is to evaluate the performance of credit risk prediction models by means of MCDM methods, demonstrating that it is necessary more than a unique measure to draw accurate conclusions about the most suitable prediction technique. The classifiers here used are a Bayesian network (BNet), the naïve Bayes classifier (NBC), logistic regression (logR), a multi-layer perceptron (MLP), a radial basis function (RBF), the nearest neighbor rule (1-NN), the RIPPER propositional rule learner, and two decision trees (C4.5 and CART).

Experiments have been carried out on seven real-life financial data sets, whose main characteristics are reported in Table 2. The Australian, German and Japanese data sets are taken from the UCI Machine Learning Database Repository (`http://archive.ics.uci.edu/ml/`). The Iranian data set comes from a modification to a corporate client database of a small private bank in Iran [13]. The Polish data set contains bankruptcy information of 120 companies recorded over a two-year period [12]. The Thomas data set, which comes with the book by Thomas et al. [15], describes applicants for a credit product. The UCSD data set corresponds to a subset with samples randomly chosen from the database used in the 2007 Data Mining Contest organized by the University of California San Diego and Fair Isaac Corporation.

Taking into account that data are too limited, a five-fold cross-validation method has been adopted to assess the credit risk prediction models. Ten repetitions have been run

Table 2. Some characteristics of the credit data sets

	Australian	German	Iranian	Japanese	Polish	Thomas	UCSD5000
#Attributes	14	24	27	15	30	12	38
#Good	307	700	950	296	128	802	2500
#Bad	383	300	50	357	112	323	2500

for each trial. The results from classifying the test samples have been averaged across the 50 runs and then evaluated with six performance measures and analyzed with two MCDM tools.

5 Results

Table 3 shows the results of six performance measures averaged over the seven data sets. For each measure, the best performing model is underlined. When using the accuracy, logistic regression, SVM, RIPPER and CART are the classifiers with the highest average rates. With the RMSE measure, logistic regression and CART appears to be the best classifiers. Assessment by means of the true positive rate suggests that logistic regression, RBF neural network and CART decision tree are the best performing algorithm. The naïve Bayes classifier has yielded the highest true negative rate. By using AUC, the best model corresponds to logistic regression, whereas the 1-NN decision rule outperforms the remaining classifiers in terms of geometric mean.

Table 3. Performance results averaged over the experimental databases

	Accuracy	RMSE	TPrate	TNrate	AUC	Gmean
BNet	0.80	0.39	0.87	0.48	0.78	0.55
NBC	0.64	0.54	0.76	0.57	0.77	0.62
logR	0.81	0.37	0.88	0.51	0.80	0.61
MLP	0.79	0.40	0.85	0.52	0.77	0.62
SVM	0.81	0.43	0.87	0.50	0.68	0.51
RBF	0.77	0.39	0.88	0.42	0.74	0.49
1-NN	0.77	0.47	0.81	0.56	0.69	0.66
RIPPER	0.81	0.38	0.87	0.51	0.70	0.61
C4.5	0.79	0.40	0.86	0.51	0.71	0.61
CART	0.81	0.37	0.88	0.48	0.71	0.53
Weight	0.04762	0.23810	0.09524	0.14286	0.19048	0.28571

There is no classification algorithm that achieves the best results across all measures and therefore, one might draw different conclusions about the best performing model depending on the performance evaluation measure used. For example, the TNrate indicates the naïve Bayes classifier as the most suitable method, the AUC proposes the logistic regression model as the best algorithm, and the geometric mean of accuracies suggests that the 1-NN rule is the most accurate technique. Clearly, this corresponds to a realistic situation in which multiple criteria should be considered in order to make a more reliable decision.

Although assigning weights to alternatives is nontrivial, here the weight of each performance measure used in TOPSIS and PROMETHEE methods has been set according to its relative importance for the financial risk prediction task. Then the weights have been normalized in the interval $[0, 1]$ such that the sum of all weights is equal to 1 (see the last row in Table 3).

Table 4 summarizes the ranking of models generated by TOPSIS and PROMETHEE multi-criteria decision making methods. The results are straightforward: the higher the ranking, the better the classifier. From the analysis with these two MCDM tools, the logistic regression model appears to be the best performing algorithm, whereas the RIPPER rule learner and the MLP neural network are among the top-three ranked classifiers by both TOPSIS and PROMETHEE methods. These results indicate that TOPSIS and PROMETHEE, which provide similar top-ranked classification algorithms, can be useful to make accurate decisions in financial risk prediction problems.

Table 4. Rankings of the TOPSIS and PROMETHEE methods

Alternative	TOPSIS	Alternative	PROMETHEE
logR	0.91444	logR	0.61905
RIPPER	0.87712	MLP	0.29101
MLP	0.80115	RIPPER	0.19577
BNet	0.78659	BNet	0.06349
CART	0.78476	NBC	0.00529
C4.5	0.76144	CART	−0.04762
RBF	0.72369	1-NN	−0.04762
SVM	0.59485	C4.5	−0.18519
1-NN	0.44631	RBF	−0.31746
NBC	0.20299	SVM	−0.57672

6 Conclusions

This paper advocates the application of MCDM methods to evaluate the performance of credit risk prediction models. It has been demonstrated that the use of single performance evaluation measures may lead to unreliable conclusions about the best performing algorithm, what makes difficult the selection of the most accurate model for solving a particular financial problem.

Two popular MCDM techniques, TOPSIS and PROMETHEE, have been tested in the experiments over seven real-life credit data sets, using ten prediction models (alternatives) and six performance measures (criteria). The evaluation of models by means of single performance measures has given contradictory results, in the sense that different measures have proposed different algorithms as the best alternative. This suggests that credit risk prediction is a real-world problem where MCDM tools should be applied to consistently evaluate a set of models. Both TOPSIS and PROMETHEE have indicated that logistic regression, RIPPER and MLP are the best prediction models when the performance is evaluated with a composite of measures.

Acknowledgment

Work partially supported by the Spanish Ministry of Education and Science (CSD2007–00018 and TIN2009–14205), and the Generalitat Valenciana (PROMETEO/2010/028).

References

1. Abdou, H.A., Pointon, J.: Credit scoring, statistical techniques and evaluation criteria: A review of the literature. Intelligent Systems in Accounting, Finance & Management **18**(2–3) (2011) 59–88
2. Antonakis, A.C., Sfakianakis, M.E.: Assessing naïve Bayes as a method for screening credit applicants. Journal of Applied Statistics **36**(5) (2009) 537–545
3. Baesens, B., Gestel, T.V., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J.: Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society **54**(6) (2003) 627–635
4. Bensic, M., Sarlija, N., Zekic-Susac, M.: Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. Intelligent Systems in Accounting, Finance and Management **13**(3) (2005) 133–150
5. Brans, J.P., Vincke, P.H.: A preference ranking organisation method: The PROMETHEE method for multiple criteria decision-making. Management Science **31**(6) (1985) 647–656
6. Desai, V.S., Crook, J.N., Overstreet, G.A.: A comparison of neural networks and linear scoring models in the credit union environment. European Journal of Operational Research **95**(1) (1996) 24–37
7. Hand, D.J.: Good practice in retail credit scorecard assessment. Journal of the Operational Research Society **56**(9) (2005) 1109–1117
8. Huang, Z., Chen, H., Hsu, C.J., Chen, W.H., Wu, S.: Credit rating analysis with support vector machines and neural networks: A market comparative study. Decision Support Systems **37**(4) (2004) 543–558
9. Hwang, C.L., Yoon, K.: Multiple Attribute Decision Making – Methods and Applications. Springer-Verlag, New York (1981)
10. Köksalan, M., Wallenius, J., Zionts, S.: Multiple Criteria Decision Making: From Early History to the 21st Century. World Scientific, Singapore (2011)
11. Lee, J.S., Zhu, D.: When costs are unequal and unknown: A subtree grafting approach for unbalanced data classication. Decision Sciences **42**(4) (2011) 803–829
12. Pietruszkiewicz, W.: Dynamical systems and nonlinear Kalman filtering applied in classification. In: Proceedings of the 7th IEEE International Conference on Cybernetic Intelligent Systems, London, UK (2008) 263–268
13. Sabzevari, H., Soleymani, M., Noorbakhsh, E.: A comparison between statistical and data mining methods for credit scoring in case of limited available data. In: Proceedings of the 3rd CRC Credit Scoring Conference, Edinburgh, UK (2007)
14. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management **45**(4) (2009) 427–437
15. Thomas, L.C., Edelman, D.B., Crook, J.N.: Credit Scoring and Its Applications. SIAM, Philadelphia, PA (2002)
16. Wang, G., Hao, J., Ma, J., Jiang, H.: A comparative assessment of ensemble learning for credit scoring. Expert Systems with Applications **38**(1) (2011) 223–230
17. Yobas, M.B., Crook, J.N., Ross, P.: Credit scoring using neural and evolutionary techniques. IMA Journal of Mathematics Applied in Business and Industry **11**(4) (2000) 111–125