Integrating Genetic Algorithm and Rough Set Theory for Credit Rating Forecasting

LEE Mushang, LIN Sin-Jin
(Chinese Culture University)

Abstract: Constructing an accurate, objective and effective credit rating model is the primary concern of the rating policies for financial institutions. The well-established forecasting model can help decision-makers to avoid enormous financial losses from potential bad debt. However, real-life dataset consists of redundant, incomplete, or inconsistent information which will impede the performance of forecasting model in terms of prediction accuracy. Rough set theory (RST) is a powerful mathematical technique for the automatic classification of datasets in a diverse range of research topics, including machine learning and knowledge generation and it can be used to handle the imprecise, vague and uncertain data. One of the critical problems of RST is reduct generation and it is very time-consuming. Therefore, the study implements the genetic algorithm (GA) to determine the best reduct of RST. The best reduct generated by GA can enhance the forecasting performance, facilitate the calculation speed, and eliminate the computational burden. In addition, the knowledge generated form RST can be expressed in a logical statement which is easy for decision-makers to realize. To examine the effectiveness of the proposed model, the study uses the model to the real-life credit rating dataset in Taiwan. According the research outcomes, the introduced model poses outstanding performance.
Keywords: Rough set theory, Genetic Algorithm, Credit rating, Knowledge generation

1. Introduction

Corporates' credit rating status have been extensively utilized by investors, decision makers, bond issuers, bankers and public governors as a surrogate of riskiness of themself. They are essential determinants of risk premiums and even the marketability of bonds (Huang et al., 2004). The status of corporate's credit ratings are typically very costly to acquire, since they need agencies such as S&P (Standard and Poor's), Moody's and Taiwan Rating Corporation (TRC) to invest considerable human resource and time to implement deep measurement which are grounded on numerous aspects ranging from strategic competitiveness to operational level details. Therefore, large efforts are made in order to mimic the credit rating procedure of the rating agencies through statistical approaches (Ederington, 1985) and artificial intelligence (Wu and Hsu, 2012). The complexity in establishing such architecture lies in the subjectivity of the credit rating process, as the difficult relations among the financial and non-financial attributes are complicate to evaluate. Such a complicated process makes it hard to discriminate rating status through statistical approaches. However, artificial intelligence methods can be used for

modeling such complex relations.

Rough set theory (RST) proposed by Pawlak (1982) is a useful mathematical approach to tackle with vagueness and uncertainty in available information. The outcomes of a RST are usually expressed in the form of a set of comprehensive rules derived form a decision table and it also can determine the useful information in huge datasets (that is, attribute selection process). Attribute selection is the procedure of selecting a subset of attributes from the initial set of attributes forming patterns in a given dataset (Wang et al., 2007). The subset should be necessary and sufficient to illustrate target concepts, sustaining a superior forecasting performance in expressing the initial attributes. The usefulness of attribute selection is to reduce the problem size and resulting search space for learning algorithms.

There are numerous RST algorithms for attribute selection. The most fundamental solution to determining minimal reducts is to yield all possible reducts and choose any with minimal cardinality, which can be done by generating a sort of discernibility function from the dataset and simplifying it (Wang et al., 2007). Skowron and Rauszer (1992) stated that the problem of minimal reduct determination is NP-hard and the problem of generation of all reducts is exponential. Thus, heuristic algorithms have to be conducted.

Genetic algorithm (GA), one kinds of Evolutionary computing, have been extensively used for handling optimization tasks, such as, optimization tasks of searching, scheduling, knowledge generation tasks, and artificial intelligence and it performs an satisfactory jobs in numerous research fields. Thus, GA was performed to determine the best reduct for RST. The best reduct generated by GA can facilitate the forecasting performance and alleviate the computational complexity.

This paper is organized as follows: In section 2, the RST and GA approaches are illustrated. In Section 3, an empirical result was expressed. Section 4 concludes the paper.

2. Methodologies

2.1 Rough set theory: RST

Rough set theory adopts information systems to represent knowledge and deal with vague data. An information system containing condition attributes and decision attributes is depicted as follow:

$$IS = (U, \Omega, V, g)$$

(1)

where U denotes a nonempty finite set with $n$objects $\{p_1, \ldots, p_n\}$, Ω is a nonempty finite set with m attributes $\{q_1, \ldots, q_m\}$. V is called the range of $U \times \Omega$, and $f : U \times \Omega \rightarrow V$is an information function where $g(p,q) \in V$for every $p \in U, q \in \Omega$. In addition, let $Q \subseteq \Omega$and $(x, y) \in U \times U$. At present, x and y are two objects.

Indiscernibility arises from an inability to distinguish among objects in a distinct set, and

results in identical information derived from different observations. The indiscernibility relation of x and y in terms of Q is defined as follows:

$$IND(Q) = \{(X,Y) \in U \times U : g(x,q) = g(y,g) \forall q \in Q\}$$

(2)

The indiscernibility relation partitions the universe U into a family of equivalence classes. The equivalence classes of the relation, $IND(Q)$, are called the Q-elementary sets in IS, and $[x]_{IND(Q)}$ represents the Q-elementary set containing the objective $x \in U$. In RST, knowledge of objects is presented in a decision table.

Lower and upper approximation is the second essential concept of RST. Let $Q \subseteq \Omega$, and $X \subseteq U$. The Q-lower approximation of $X(Q_L)$ and the Q-upper approximation of $X(Q_U)$ are then defined as follows respectively:

$$X(Q_L) = \{x \in U : [x]_{IND(Q)} \subseteq X\}$$

(3)

$$X(Q_U) = \{x \in U : [x]_{IND(Q)} \cap X \neq \phi\}$$

(4)

Each object in the lower approximation set of X must be in X. If an objective is in the upper approximation set of X, then it might or might not be in X.

Two essential RST concepts in knowledge reduction are the reduct, represented by $RED(Q)$, and core, depicted as $CORE(Q)$. A reduct is a basic component of an information table, of which the core is the intersection of all reducts. The relation between reducts and the core can be represented as follows:

$$CORE(Q) = \cap RED(Q)$$

(5)

Reducts can be derived through the discernibility matrix and Boolean operations. The discernibility matrix is a set that can identified between two objects or sets.

The rule generation from decision table to classify new objects is one of the most significant functions of RST. Rules are derived from the condition attributes based on the decision table. The prediction of the new objective is performed by matching its description to one of the rules.

2.2 Genetic algorithm (GA)

Genetic algorithms (GA), inspired by evolution, are search algorithm which were advanced by Holland (1975) and expanded by Goldberg (1989). It has been successfully performed in numerous economic and financial prediction domains (Allen and Karalainen, 1999). The basic illustration of GA is expressed as follows (Goldberg, 1989):

(1) Phase 1 (Initialization): This phase generates the initial population containing G_p chromosomes, which are executed to determine global optimum initial seeds, where K_p is the number of individuals in each generation. The probability of crossover denotes as P_c, the probability of mutation denotes as P_m.

(2) Phase 2 (Evaluation): After the phase 1, each chromosome is measured using a pre-determined fitness function. The fitness value of each string is an index of the problem's design improvement suitability and the probability of survival of reproduction in GA (Cheng et al., 2010).

(3) Phase 3 (Check stopping criteria): After the prior phases, the procedures, from phase 2 to 7, are repeated until the stopping criteria are reached. The stopping criteria were that (1) the maximum number of generation is reached, or (2) the forecasting outcome has not been changed for the present generation.

(4) Phase 4 (Elitism mechanism): To assure the propagation of the elite chromosome, Elitism mechanism was conducted in GA.

(5) Phase 5 (Selection): Selection is the procedure used to choose the suitable chromosome from the parent's population for the next generation. Tournament selection (Goldberg and Deb, 1991) was conducted. The chromosomes with the best fitness values will be selected. This phase is repeated until the number of chromosomes selected and the number of the population are equivalent.

(6) Phase 6 (Crossover): This phase runs by swapping corresponding segments of a string representation of the parents and enlarge the searching field for a new solution. Uniform crossover (Syswerda, 1989) was executed to eliminate the effect of positional bias.

(7) Phase 7 (Mutation): In this phase, it selects a member of the population randomly and modifies one randomly selected bit in its string representation.

3. Experimental results

3.1 The financial data

The financial data were gathered form the databases of the publicly websites (i.e., Taiwan Economic Journals: TEJ; Taiwan Stock Exchange: TSE). The selected data contained 900 public electronic corporates from 2011 to 2013. The credit rating status was divided into 10 ranks, from superior to inferior. To prevent the construction of unreliable forecasting model, the credit ranks were further categorized into three levels: lower risk, medium risk and higher risk. The selected informative attributes used to construct the forecasting model were expressed as follows: A1: Working capital/total assets (WCTA); A2: Current assets/current liabilities (CACL); A3: Net income/total assets (NITA); A4: Cash flow/total debt (CFTB); A5: Cash flow/total sales (CFTS); A6: Net income/shareholders' equity (NISE); A7: Total debt/total assets (TDTA); A8:

Inventory/total sales (ITS); A9: Operating income/total sales (OITS); A10: Net income/(total assets-total liability) (NITATL); A11: Long term debt/total assets (LTDTS).

3.2 The assessing criteria and outcomes

Prediction accuracy is core measurement of classifiers in risk evaluation. The study executes the overall accuracy, sensitivity and specificity to construct the performance evaluation.

To evaluate the effectiveness of the proposed model, the study further compared it with three other classifiers (back propagation neural network: BPNN; Bayesian network: BN; and discriminant analysis: DA). The advantage of cross validation is that the impact of data dependency is minimized and the reliability of the outcomes can be improved. A five-fold cross validation was executed in this study. Figures 1-3 depicts the outcomes of the comparison of overall accuracy, sensitivity and the specificity of the four classifiers. According to our research finding, the proposed forecasting mechanism (GA-RST) outperforms than other three classifiers in whole assessing criteria.

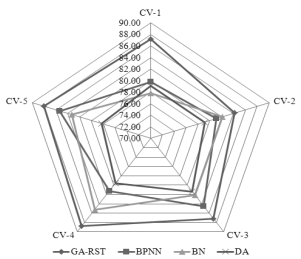

Fig. 1. The assessing outcome (Overall accuracy).

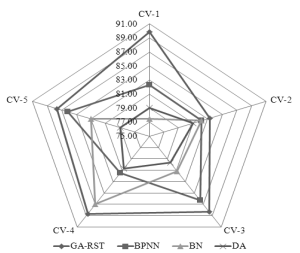

Fig. 2. The assessing outcome (Sensitivity).

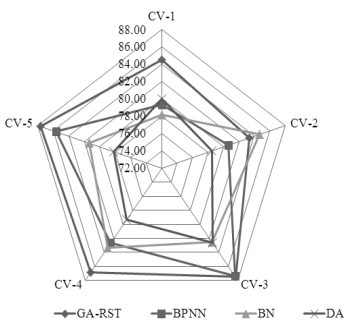

Fig. 3. The assessing outcome (Specifricity).

Furthermore, the decision process of GA-RST can be expressed in logical statement which was realizable for users. The decision rules were expressed in Table 1. The rules can be viewed as a roadmap for users to adjust personnel financial investing policy. For managers, the rules can be used to modify the capital structure to survive in highly competitive environment.

Table 1: The decision rules

Rule expression
Rule 1: If TDTA is between 0.18 and 0.32, and OITS is between -0.11 and -0.07, then the rating condition is "***Higher Risk***"
Rule 2: If WCTA is between 0.02 and 0.07, and NITA is between -0.12 and -0.05, then the rating condition is " ***Higher Risk***"
Rule 3: If WCTA is between 0.11 and 0.19, and NITA is between 0.03 and 0.11, then the rating condition is "***Lower Risk***"
Rule 4: If CACL is between 0.22 and 0.31 and TDTA is between 0.03 and 0.09, then the rating condition is "***Lower Risk***"
Rule 5: If OITS is between 0.01 and 0.05, TDTA is between 0.18 and 0.32, then the rating condition is "***Medium Risk***"

4. Conclusion

Due to the dramatic increase in the delinquency rate, many baking institutions and corporate have had to set aside large amount of money to compensate for bad debt. This will result in liquidity problem and profit elimination. Most proportion of banking institutions making final decision relies on rating status. Unfortunately, an evaluation of the rating status yielded by professional organizations is not accessible right away. Therefore, many artificial intelligence techniques have been executed for early assessment of credit risk. The study implements GA-RST model to analyze the status of credit rating has numerous advantages. The GA was performed to determine the best reduct for RST. It would eliminate the computational complexity and facilitate the operational speed. After determining the best reduct, the RST can yield comprehensive decision rule for users. The rules can be used to modify the capital structure and investing strategies. Moreover, the study advanced compared the proposed model with other classifiers to examine the usefulness. The results indicated that the GA-RST is a promising forecasting mechanism.

Reference

[1]. F. Allen, R. Karalainen, Using genetic algorithms to find technical trading rules, Journal of Financial Economics, 1999, 51, 245–271.

[2]. C. H. Cheng, T. L. Chen, L. Y. Wei, A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting, Information Sciences, 2010, 180(9), 1610-1629

[3]. L. H. Ederington, Classification models and bond ratings, Financial Reviews, 1985, 20 (4), 237-262.

[4]. D. E. Goldberg, K. Deb, A comparative analysis of selection schemes used in genetic algorithm, in: G. Rawlins (Ed.), Foundation of Genetic algorithms, Morgan Kaufmann, 1991, 69–93.

[5]. D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison Wesley, Reading, MA, 1989.

[6]. J. H. Holland, Adaptation in Nature and Artificial Systems, University of Michigan Press, 1975.

[7]. Z. Huang, H. Chen, C. J. Hsu, W. H. Chen, S. Wu, Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision Support Systems, 2004, 37(4), 543-558.

[8]. Z. Pawlak, Rough set. International Journal of Information and Computer Sciences, 1982, 11, 341-356.

[9]. Z. Pawlak, Rough sets and intelligent data analysis, Information Sciences, 2002, 147, 1-12.

[10]. A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems. In: Slowinski, R. (Ed.), Intelligent Decision Support—Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publishers, Dordrecht, 1992, 311–362.

[11]. G. Syswerda, Uniform crossover in genetic algorithms, in: Proceeding of the 3rd International Conference on Genetic Algorithms, 1989, 2-9.

[12]. X. Wang, J. Yang, X. Teng, W. Xia, R. Jensen, Feature selection based on rough sets and particle swarm optimization, Pattern Recognition Letters, 2007, 28(4), 459-471.

[13]. T. C. Wu, M. F. Hsu, Credit risk assessment and decision making by a fusion approach. Knowledge-Based Systems, 2012, 35, 102-110.

整合基因演算法與約略集合於信用評等之預測

李慕萱　林欣瑾
（中國文化大學）

摘要：構建一個準確，客觀與有效的信用風險模型是許多金融機構的共同目標，構建完善的預測模型可以降低潛在的違約風險與損失，然而，現今的資訊充滿著不完整，多餘和不一致，而這些資訊嚴重影響預測模型的診斷能力。約略集合是一個相當具有效力的數學工具，它可以用來處理不完整，模糊與不精確的資料型態，此技術亦廣泛應用於許多機器學習與知識擷取之研究中，且都有相當優越的表現。如何產生最佳的子集合是約略集合被詬病的缺陷，因為產生最佳子集是相當耗費時間，因此，本研究將採用基因演算法用於決定其最佳子集合，此步驟不僅大大降低其運算時間並提升其模型的預測能力。此外，約略集合可以歸納出有用的資訊並以邏輯概念呈現，易於決策者所理解與使用，為了進一步驗證此模型的效力，本文將此模型應用於台灣信用評等之預測上，經研究結果比較顯示，此模型具有相當優越的鑑別力。

關鍵字：約略集合，基因演算法，信用評等，知識產生