A Hybrid Approach to Credit Scoring Applying Rough Set and Genetic Programming

Chia-Ping Lin, Mu-Chen Chen* and Chih-Ming Hsu

*To whom correspondence should be addressed.

Mu-Chen Chen, Ph.D.
Professor
Department of Business Management
National Taipei University of Technology
No. 1, Section 3, Chung-Hsiao E. Road
Taipei 106, Taiwan, ROC.
Tel: +886-2-2771-2171 ext. 3417
Fax: +886-2-2776-3964
E-mail: bmcchen@ntut.edu.tw
A Hybrid Approach to Credit Scoring Applying Rough Set and Genetic Programming

Chia-Ping Lin
Institute of Commerce
Automation and Management
National Taipei University of Technology

Mu-Chen Chen
Department of Business Management
National Taipei University of Technology

Chih-Ming Hsu
Department of Business Administration
Minghsin University of Science and Technology

ABSTRACT

This paper applies a hybrid classification approach combining rough set and genetic programming (GP) to construct the credit scoring model. Comparing with the previous credit scoring model only based on GP, the hybrid method not only makes an improvement in the average classification accuracy, but also saves the required computational effort.

Keywords: Credit scoring; Classification model; Rough set; Genetic programming.

1. Introduction

In the previous studies of credit scoring, researchers and practitioners have used conventional statistical methods to develop a variety of models for credit scoring, which involve linear discriminant models [21], logistic regression models [10], k-nearest neighbor models [9] and decision tree models [5]. When adopting conventional statistical classification procedures such as linear discriminant analysis and logistic analysis, an underlying probability model must be assumed in order to calculate the posterior probability upon which the classification decision is made. Such assumption restricts the applications of conventional statistical methods in credit classification. However, the data mining techniques such as neural network (NN), rough set and genetic programming (GP) can perform the classification task without this limitation. Further, the nonlinear features of these techniques make them a potential alternative to traditional parametric (e.g., linear discriminant analysis and logistic regression) and non-parametric (e.g., k-nearest neighbor and decision tree) methods. Recently, NN [6, 16, 23] and GP [3] have been successfully utilized to build the credit scoring models.

In the study by Chen et al. [3], they have constructed a GP-based credit scoring model and have made comparisons with several alternative methods involving NN, classification and regression tree (CART) and linear discriminant analysis (LDA). According to their experimental results, GP methodology is very competitive to NN in terms of classification accuracy. Furthermore, GP has the superior capability in discovering discriminative features, which NN is relatively difficult to do. Feature selection is an important issue in building classification systems. It is advantageous to limit the number of input features in a classifier in order to have a good predictive and less computationally intensive model [25].

As mentioned above, recognizing the discriminative features prior to learning in building classification models can decrease the computational requirement, whereas not reduce the accuracy. A recently developed classification technique, rough set theory, was reported to succeed in effectively generating discriminative feature sets (namely reducts). Rough set, a non-parametric technique, was developed by Zdzislaw Pawlak in the early 1980's and it is originated from the mathematical set theory [18, 19]. It has been used for feature selection, feature extraction, data reduction, decision rules...
generation, and pattern extraction, etc [20]. Hashemi et al. [8] and Jelonek et al. [11] have successfully applied rough set to reduce the data input to neural networks. McKee and Lensberg [14] have also developed a hybrid approach that GP coupled with rough set theory, and it performs successfully in modeling bankruptcy.

In this study, we attempt to apply a hybrid method for constructing the credit scoring model, which incorporates rough set theory and genetic programming. Rough set is firstly adopted to select the discriminative features by identifying reducts. Only those discriminative features generated from rough set are then taken as the input features for the GP learning process.

2. Rough set

Rough set at first introduced by Pawlak is a relatively new mathematical method in the area of data mining [18, 19]. Rough set is developed for the analysis and pattern discovery in database, particularly for data that are ambiguous or incomplete [15]. One of the main advantages of rough set is that it dose not need any preliminary or additional information about data, such as probability distribution [19]. Rough set is established in the mathematical set theory. The concepts of rough set theory are discussed in the following subsections.

2.1 Information table

The real world information can be presented in the form of information table (or called decision table), such as shown in Table 1. Table 1 illustrates a simple information table of credit scoring of seven objects. The scenario is weather a financial institution should accept the applications or not. The first two columns, age and income, represent the condition attributes and the last column, admission, the decision attribute. Each application is called an object. The simple example shown in Table 1 is used to explain the concepts of rough set as follows.

<table>
<thead>
<tr>
<th>Case</th>
<th>Age</th>
<th>Income (thousands)</th>
<th>Admission</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>16-30</td>
<td>50</td>
<td>Accepted</td>
</tr>
<tr>
<td>x2</td>
<td>16-30</td>
<td>0</td>
<td>Rejected</td>
</tr>
<tr>
<td>x3</td>
<td>31-45</td>
<td>1-25</td>
<td>Rejected</td>
</tr>
<tr>
<td>x4</td>
<td>31-45</td>
<td>1-25</td>
<td>Accepted</td>
</tr>
<tr>
<td>x5</td>
<td>46-60</td>
<td>26-49</td>
<td>Rejected</td>
</tr>
<tr>
<td>x6</td>
<td>16-30</td>
<td>26-49</td>
<td>Accepted</td>
</tr>
<tr>
<td>x7</td>
<td>46-60</td>
<td>26-49</td>
<td>Rejected</td>
</tr>
</tbody>
</table>

2.2 Discernibility relation

Let $\mathcal{IT} = (\mathcal{U}, \mathcal{A})$ be an information table, where $\mathcal{U}$ is a non-empty finite set of objects, $\mathcal{A}$ is a non-empty finite set of attributes; $\mathcal{B}$ is a subset of $\mathcal{A}$ and $a$ is an attribute of $\mathcal{A}$. With any $\mathcal{B} \subseteq \mathcal{A}$, it exists an equivalence relation as below:

$$\text{IND} (\mathcal{B}) = \{(x, x') \in \mathcal{U}^2 | \forall a \in \mathcal{B}, a(x) = a(x')\}$$

$\text{IND}(\mathcal{B})$ is the $\mathcal{B}$-indiscernibility relation. Taking the $\mathcal{IS}$ in Table 1 as an illustrative example, the non-empty subsets of the condition attributes are $\{\text{age}\}$, $\{\text{income}\}$, and $\{\text{age, income}\}$. The discernibility relations are stated with attribute sets $\{\text{age}\}$,
\{\text{income}\}$, and $\{\text{age, income}\}$ as follows:

\begin{align*}
\text{IND}(\text{age}) &= \{(x_1, x_2, x_3, x_4, x_5, x_6, x_7)\} \\
\text{IND}(\text{income}) &= \{(x_1, x_2, x_3, x_4, x_5, x_6, x_7)\} \\
\text{IND}(\text{age, income}) &= \{(x_1, x_2, x_3, x_4, x_5, x_6, x_7)\}
\end{align*}

2.3 Set approximation

The $B$-lower and $B$-upper approximations can be defined as follows:

\begin{align*}
BX &= \{x | x \subseteq X\} &: \text{the lower approximation}, \\
\overline{BX} &= \{x | x \cap X \neq \emptyset\} &: \text{the upper approximation},
\end{align*}

where $X$ is a subset of $U$, $X \subseteq U$. Two terminologies of rough set, boundary region and outside region, can be defined with respect to the concept of lower and upper approximations. Boundary region is the difference between the upper and the lower approximations. It can be expressed as:

$$BN_B(X) = \overline{BX} - BX$$

The objects in the difference region can not be clearly classified into $X$ in $B$. Outside region consists of those objects that can be with certainty classified as not belonging to $X$. It takes the form as:

$$OR_B = U - \overline{BX}$$

We also use the information system shown in Table 1 to illustrate the concepts of approximations. Let $CS = \{x | \text{admission}(x) = \text{accepted}\}$. Hence,

- the lower approximation: $\overline{ACS} = \{x_1, x_6\}$,
- the upper approximation: $ASC = \{x_1, x_3, x_4, x_6\}$,
- the boundary region: $BN_A(CS) = \{x_3, x_4\}$,
- the outside region: $U - ACS = \{x_2, x_5, x_7\}$.

2.4 Reduct

We can reduce data through only keeping the attributes that are required to preserve the indiscernibility relation. Let $B$ be a non-empty subset of $A$, the set of condition attributes. $B$ is a reduct of $A$ if $B$ is a maximal independent set of condition attributes [19]. There is no redundant attribute in $B$ and all attributes in $B$ are indispensable.

3. Genetic Programming

Recently, genetic programming (GP) [12] has been developed by Koza to extend the more familiar genetic algorithms [7]. Instead of performing the search within a numerical solution space, GP searches within a topological space. This allows GP to generate structures, mathematical expressions and computer programs, which can be easily executed by other application systems.

In genetic algorithms, solutions are represented by fixed length strings called chromosomes. While this representation is efficient, it is not easy to map various problems into such a schema. Koza [12] extended the framework of genetic algorithms by relaxing the limitation of fixed length string presentation. This allows the flexible presentation of solutions as hierarchies of different functions in tree-like structures. In this paper, we attempt to combine rough set and GP to develop a hybrid method for

3
credit scoring. The basic components of GP are described in [4]. The further details are well documented in Koza [12].

3.1 Terminal set and function set

The first part of the design of GP is to determine the elements in terminal set $T$ and function set $F$. GP is an automatic parallel search procedure conducted over a predefined space $\Omega$ composed of computer programs. In GP, there is a standard manner to represent these computer programs, i.e., the LISP S Expression. It can also be depicted as a parse tree.

The set of possible computer programs in GP consists mainly of elements from terminal set $T$ and function set $F$. The selection of function set and terminal set depends on the problem. Terminal set includes elements of input features and constants. The function set may include the following functions:

- Arithmetic operators (+, -, $\times$, ÷)
- Mathematical functions (e.g., SIN, COS, EXP, LOG)
- Boolean operators (e.g., AND, OR, NOT)
- Conditional operators (e.g., IF THEN ELSE)
- User-defined domain-specific functions.

The set of possible computer programs in GP consists mainly of elements from terminal set $T$ and function set $F$. The selection of function set and terminal set depends on the problem. Terminal set includes elements of input features and constants. The function set may include the following functions:

- Arithmetic operators (+, -, $\times$, ÷)
- Mathematical functions (e.g., SIN, COS, EXP, LOG)
- Boolean operators (e.g., AND, OR, NOT)
- Conditional operators (e.g., IF THEN ELSE)
- User-defined domain-specific functions.

![Figure 1. A tree expression for GP.](image)

3.2 Genetic Operations

Initialization scheme: Once $T$ and $F$ are determined, GP begins with a population of randomly generated computer programs. Generally, the population size ($N$) is arbitrarily given and is fixed throughout the complete evolution process.

Fitness measure: Fitness is a numeric measure based on the suitability of each member of the population as a solution, and it is used as the basis for selecting members of the population for evolving. Considering the objective of credit scoring, the fitness function in this study is based on the hit rate of classification.

Selection scheme: A high fitness value implies the capability of a chromosome (computer program) to create offspring more frequently, whereby the genetic material is passed on to the following generation. In this paper, we choose the proportionate selection scheme, also known as the roulette wheel selection scheme.

3.3 Genetic Operators

We now briefly describe three genetic operators.

Reproduction: This operator is used for fast focusing in the direction of hopeful chromosomes (computer programs). The reproduction replaces the worst individual in the population by the offspring. Reproduction makes the copies of individual computer
programs.

**Crossover**: This operator crosses two computer programs in parse tree structure by exchanging subtrees at randomly selected nodes. The crossover operation for the GP paradigm is a sexual operation that starts with two parental computer programs, which are randomly selected from population proportionally to their normalized fitness.

**Mutation**: Mutation operates on only one computer program (chromosome). It allows new computer programs to be created. This operator starts by selecting a computer program expressed by a parse tree from the population based on its normalized fitness.

The basic procedure of a steady-state GP algorithm is illustrated as follows [12]:

1. Initialize the population.
2. Randomly choose a subset of the population to take part in the tournament (the competitors).
3. Evaluate the fitness value of each competitor in the tournament.
4. Select the winner(s) from the competitors in the tournament using the selection scheme.
5. Apply genetic operators: reproduction, crossover and mutation to the winner(s) of the tournament.
6. Replace the losers in the tournament with the results of the application of the genetic operators to the winners of the tournament.
7. Repeat Steps 2-7 until the termination criterion is met.
8. Choose the best individual in the population as the output.

**4. The Hybrid Credit Scoring Model**

In this study, we propose a hybrid approach combining rough set and genetic programming for credit scoring to achieve better performance. The proposed hybrid approach primarily concentrated on data reduction for modeling. Rough set is applied herein to eliminate the redundant attributes of credit dataset. Instead of using the complete feature set, only the *reduct* (partially discriminative feature set) generated by rough set is adopted as the input feature set of GP. It is expected that the noise of data caused by redundant attributes can be eliminated. Therefore, the computational requirement is considerably decreased, while the classification accuracy of credit scoring approximately maintains.

There are two stages in the hybrid model for credit scoring, and the procedure can be described as follows. In the first stage, the discriminative features are selected by rough set. All attributes are input to the rough set classifier to discover *reducts*. The features in *reducts* are taken as the discriminative attributes. In the second stage, GP is triggered to generate the credit scoring model. Instead of using all attributes in the credit dataset, only the significant features presented in *reducts* obtained from the first stage are fed into the GP learning procedure as the input variables. Figure 2 schematically illustrates the procedure of the hybrid model.
Both the adopted techniques, rough set and GP, are effective tools for classification. It is beneficial to incorporate these two methods in building credit scoring models. One of the advantages of the hybrid method is that rough set can eliminate the irrelevant features, and GP can then focus on the discriminative features to construct the classification model. Such a hybrid approach can improve the classification accuracy and reduce the required learning time. It should be noticed that eliminating an attribute is meant to delete a column in data table. Provided that the objects in data set are tremendous, elimination of any attribute can reduce the data to a large extent, and thereby translate considerably into savings in CPU time.

5. The test example

We use a real world data set, Australian credit, from UCI Repository of Machine Learning Database [17] to evaluate the performance of the proposed hybrid model. The computational results of the hybrid approach are benchmarked to the results of neural network, GP and other conventional methods such as LDA and CART.

The Australian credit data consists of 307 instances of creditworthy applicants and 383 instances where credit is not creditworthy. Each instance contains 15 attributes \{a_1, a_2, \cdots, a_{15}\} and one class attribute (accepted or rejected). This dataset is interesting because there is a good mixture of attributes: continuous, nominal with small numbers of values, and nominal with larger numbers of values. There are also a few missing values. To protect the confidentiality of data, the attributes names and values have been changed to meaningless symbolic data.

In the data pre-processing step, we take out the instances where exist missing attribute values because of not knowing the meaning of the attributes. The instances of creditworthy applicants are 296, and not creditworthy instances 357 after deleting missing values. Since rough set is only able to deal with discrete data, the continuous attributes in the credit dataset are discretized by using an entropy based method. The discretized dataset is randomly partitioned into a training set of 436 instances and a testing set of 217 instances. Rough set is utilized to discover the discriminative features (reducts). The significant features in reducts include \{a_1, a_4, a_6, a_7, a_9, a_{10}, a_{11}, a_{12}, a_{13}, a_{14}, a_{15}\}. Only these 11 attributes are then selected as input features to the GP classifier for generating the credit scoring model. In this case, 26.7\% (4/15) of data are removed from the dataset for GP.

GP can deal with the data with both continuous and discrete types. Therefore, it is unnecessary to discritize the continuous data. In the second stage of the proposed hybrid approach, the credit dataset is randomly partitioned into training and independent test sets using a 5-fold cross validation. Ten repetitions are run for each trial due to the
stochastic nature of GP. The testing set is used to guarantee that our results are valid and can be generalized to predict the class labels of new data. Each of the 5 random partitions performs as an independent holdout testing set for the credit scoring model trained with the rest of four partitions. The benefits of cross validation are that the impact of data dependency is minimized and the reliability of results is improved. In addition, the credit scoring model is developed with a huge portion of the accessible data (80% in this case) and all the data is utilized to test the trained models. The GP specific parameters for 11 inputs and 15 inputs are listed in Table 2.

Table 2. GP parameters.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>11 inputs (this study)</th>
<th>15 inputs (previous study) [1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Maximum number of generation</td>
<td>3600</td>
<td>3600</td>
</tr>
<tr>
<td>Crossover rate</td>
<td>0.8, 0.9</td>
<td>0.8, 0.9</td>
</tr>
<tr>
<td>Mutation rate</td>
<td>0.1, 0.2</td>
<td>0.1, 0.2</td>
</tr>
</tbody>
</table>

A threshold value of 0.5 is used to distinguish between credit groups, good credit and bad credit. The results for Australian credit dataset by using GP and hybrid method are summarized in Tables 3 and 4, respectively. The results are averages of accuracy rate obtained from each of the 5 independent holdout dataset partitions (testing accuracy) used in the 5-cross validation methods. From the results shown in Tables 3 and 4, the hybrid method makes an improvement against to GP in classification accuracy of about 3%. Although the hit rate of acceptance declines slightly (about 0.6%), the hit rate of rejection rises about 4%. The improvement in the hit rate of rejection is more beneficial than that in the hit rate of acceptance since the cost of granting a loan to a defaulter is much larger than that of rejecting a good applicant [7].

Table 3. The results of GP with 15 inputs.

<table>
<thead>
<tr>
<th></th>
<th>Mutation rate =0.1</th>
<th>Mutation rate =0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept hit rate (%)</td>
<td>90.91</td>
<td>89.26</td>
</tr>
<tr>
<td>Crossover Reject hit rate</td>
<td>87.34</td>
<td>88.18</td>
</tr>
<tr>
<td>Overall hit rate (%)</td>
<td>88.99</td>
<td>88.69</td>
</tr>
<tr>
<td>Training time (seconds)*</td>
<td>208</td>
<td>172</td>
</tr>
</tbody>
</table>

Table 4. The results of hybrid method with 11 inputs

<table>
<thead>
<tr>
<th></th>
<th>Mutation rate =0.1</th>
<th>Mutation rate =0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept hit rate (%)</td>
<td>86.93</td>
<td>91.93</td>
</tr>
<tr>
<td>Crossover Reject hit rate</td>
<td>91.62</td>
<td>90.16</td>
</tr>
<tr>
<td>Overall hit rate (%)</td>
<td>89.61</td>
<td>90.98</td>
</tr>
<tr>
<td>Training time (seconds)*</td>
<td>153</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>Hybrid method</td>
<td>GP</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------</td>
<td>------</td>
</tr>
<tr>
<td>Accept hit rate (%)</td>
<td>90.34</td>
<td>90.99</td>
</tr>
<tr>
<td>Reject hit rate (%)</td>
<td>90.31</td>
<td>86.42</td>
</tr>
<tr>
<td>Overall hit rate (%)</td>
<td>90.37</td>
<td>87.26</td>
</tr>
<tr>
<td>Training time (seconds)*</td>
<td>154</td>
<td>194</td>
</tr>
</tbody>
</table>

*The CPU time is based on an IBM compatible PC with a Pentium IV 800 MHz processor.

Moreover, the hybrid method has a better performance of classification than the previous methods such as NN, LDA and CART. The results of adopting NN, LDA and CART are summarized in Table 5. The better performance indicates that selecting discriminative features as inputs can reduce the volume of data and improve the classification accuracy.

In this case, the CPU time of generating *reducts* by rough set is short and within one second. Comparing the results of the hybrid approach to that of GP, the hybrid credit scoring model can save about 20% of computational requirement. It indicates that rough set can reduce the superfluous attributes successfully and thus considerably decrease the size of dataset. Without the insignificant features, GP can perform with better efficiency and be more concise only with the discriminative features in the model building process. Generally, the hybrid approach improves the GP-based method in credit scoring.

6. Conclusion

Credit scoring is a widely used technique that helps banks to decide whether or not to grant credits to consumers who submit an application. With the rapid growth in the credit industry, it needs more accurate and quick-responded models to help credit officers making decisions. In order to improve the prediction accuracy, this paper applies a hybrid classification approach combining rough set and genetic programming to construct the credit scoring models. In this hybrid approach, rough set is used to discover discriminative attributes, which then are taken as the input for GP to reduce the learning effort. The present approach has two major advantages: (1) GP can focus on discriminative features without the irrelative features in modeling to improve the classification accuracy; (2) the computational requirement is lessened due to the reduction of data by using rough set.

Acknowledgement

This research was partially supported by the National Science Council, Taiwan, ROC under grant NSC 90-2218-E-027-008.
References


