A Bank Customer Credit Evaluation Based on the Decision Tree and the Simulated Annealing Algorithm

Yi Jiang[1], Yan Chen[2], Zhimin Zeng[1] and Xiangjian He[2]
[1]*Department of Computer Science*
Xiamen University
Xiamen, 361005
P.R.China
[2]*Faculty of Information Technology*
University of Technology, Sydney
PO Box 123
Broadway 2007, NSW
Australia

Abstract

C4.5 is a learning algorithm that adopts local search strategy, and it cannot obtain the best decision rules. On the other hand, the simulated annealing algorithm is a globally optimized algorithm and it avoids the drawbacks of C4.5. This paper proposes a new credit evaluation method based on decision tree and simulated annealing algorithm. The experimental results demonstrate that the proposed method is effective.

1. Introduction

Commercial banks are critically profit driven. Loan is one of the most important sources of profits while it is a business with high risks. To reduce risk and increase profits, it is important for banks to set up a reliable customer evaluation system which minimizes loan risks.

Customer credit evaluation systems classify customers into two categories based on their payment records. One category is “good” customers who pay back loan principals and interests in time, the other category is “bad” customers who fail to pay back loan principals and interests in time. Several features including client’s age, income, assets and so on are commonly used to find out the characters of “good” customers and “bad” customers. Classification rules are drawn and mathematical model is built to predict whether a customer is good or bad and provide support to decision making [1].

Because Simulated Annealing Algorithm (SAA) introduces new random factors in its Metropolis, it is possible for SAA to avoid the trap of local optimization. Moreover, SAA is efficient, powerful, generic and flexible. However, to the best of our knowledge, SAA has not been applied in the area of credit evaluation. This paper presents a new credit evaluation method based on decision tree and SAA. The experimental results show the proposed method is effective.

2. Decision Tree Learning

Decision tree has been widely used in the field of classification, prediction and sampling since 1960s. It is becoming more important in machine learning and knowledge discovery since Quilan introduced ID3 in 1986 [3].

The Decision Tree learning adopts a top-down recursive method. It compares the value of attribute in an inner node; then generates the branch that follows the node; finally derive the result from the leaf node. Therefore, a path from the root to a leaf node corresponds to a conjunction rule while the whole decision tree corresponds to a group of alternative expression rules. The advantage of a decision tree based learning algorithm is that it does not require users to know much background knowledge. As long as the training case can be expressed by attribute conclusion fashion, the algorithm can be applied.

3. Decision Tree Algorithm

C4.5 Decision Tree Algorithm [3] was proposed by Quinlan by extending and improving the ID3 algorithm. Besides the functions provided by ID3, C4.5 is also able to deal with continuous attributes and default attributes. In addition, unbalanced trees are avoided via pruning technology while cross certification is enabled. C4.5's simplicity, efficiency and reliability have made C4.5 become the most important algorithm in machine learning and classification. However, C4.5 is not perfect. The divide-and-conquer approach makes it achieve not global optimization but local optimization through only local search strategy. Moreover, it is difficult to restructure or make further improvement on a constructed tree because C4.5 evaluates a decision tree while building it.

4. Simulated Annealing Algorithm

Decision tree cannot achieve global optimization solution because of its local search strategy. While SAA is a simulation of solid annealing process, it follows the Metropolis rules and is capable of searching for the best possible solution from multinomial. In the mean time, SAA is reliable. This paper proposes a credit evaluation system which synthesizes SAA and decision tree . During the training stage, the input of SAA is the evaluation results from decision tree algorithm. After the process by SAA, the correct rate of decision rules is improved while the complexity of the decision rules is decreased.. SAA is mainly comprised of four parts, decoding of solutions, generating of new solutions, target function and cooling procedure.

4.1 Decoding Solutions

Based upon fixed item expressions, this paper proposes flexible fixed item expressions. There are two parts in an item expression: conditional part and prediction part. The prediction part is denoted by 0 or 1. The conditional part is composed of all attribute items of samples. The possible value of an item is represented by real numbers. For example, if a discrete-attribute item *A* has 4 possible values, these values will be assigned to 1, 2, 3 and 4 respectively. Similarly, if continuous attribute item *B* has 3 intervals, they will be assigned to 1, 2 and 3 respectively. If the value of the attribute item can be any number, then 0 is assigned to it. For example, if a record has three attributes *x1*, *x2* and *x3*, where *x1* and *x2* are discrete while *x3* is continuous, so the decision rule is:

$$(x_1 = A) \wedge (B < x_3 < C) \rightarrow \mathit{breaking} \qquad (1)$$

Suppose the sequence of value A is 1, the sequence of interval (*B*, *C*) is 3. Denote breaking by 0 and denote the attributes by *x1*, *x2* and *x3*. Then the solution will be 1_0_3_0 (each value is separated by an underscore). The merit of using real number to denote sequence is that it is easy to recognize and process continuous attributes.

4.2 Generating New Solutions

For each attribute, there is a set of possible values and a corresponding sequence set R. R consists of the sequence numbers of corresponding attribute values and special number 0. Here 0 can represent any value. When generating a new solution, a random two point replacing method is adopted, i.e., select two attributes and a sequence number from the corresponding sequence number set randomly and replace the current sequence number.

4.3 Target Function

There are two potential errors in credit evaluation systems. One is classifying "good" customers to "bad" customers. The other is classifying "bad" customers to "good" customers. If the rate of first type error is too high, it means that the rule is too strict and potential customers are losing. If the rate of second type error is too high, it means that the bank is carrying more bad loans riskily. Therefore, the smaller the target functions *f(x),* the lower error rate the credit evaluation has.

The parameters used in target functions are defined as follows:

(1) *p*: the sample set which matches the conditional part of evaluation process.
(2) *a*: an accurate prediction of good clients with *p*;
(3) *b*: an inaccurate prediction of good clients with *p*;
(4) *c*: an accurate prediction of bad clients with *p*;
(5) *d*: an inaccurate prediction of bad clients with *p*;

The target function *f(x)* is critical for the quality of solutions. This paper compares 3 target functions by experiments.

(a) the first target function:

$$f(x) = \alpha \cdot \frac{b}{a+b} + \beta \cdot \frac{d}{c+d} \qquad (2)$$

where α and β are two coefficients. The best values of α and β are determined by experiments. Generally, the cost of the second error is higher then the first error. Therefore, it is better to set β at a higher value.

(b) the second target function:

$$f(x) = \left(\frac{b}{a+b} - \alpha\right)^2 + \left(\frac{c}{c+d} - \beta\right)^2 \qquad （3）$$

where α and β are constant, and they represent the tolerance of the first type error and the tolerance of the second type error respectively.

(c) the third target function:

$$f(x) = \alpha \times \frac{b}{a+1} + \beta \times \frac{c}{d+1} \qquad (4)$$

where α and β are two coefficients.

4.4 Cooling Procedure

There are 4 steps in cooling procedure:

(1). Let the original temperature be 10 Centigrade.

(2). Let the Metropolis recursion process be stable, i.e., using a fixed number of iterations:

$$L_k = \beta \times n (k = 0,1,2...) \qquad (5)$$

where $\beta \geq 1$ and β is an integer, n is the number of attributes of each sample.

(3). Lower the temperature by

$$T_k +1 = T_k \times 0.9 (k = 0,1,2...) \qquad (6)$$

(4). Outer iteration uses the controlling method based on none improvement rule, i.e., stopping the algorithm when there is no further improvement for the current local optimized solution under given temperature and number of iteration. Here the number of iteration is 8.

5. Mixed Algorithm

5.1 procedures

(1) Original decision rule sets are generated. C4.5 algorithm is widely used for decision tree learning algorithm because of its high quality of solution. This paper uses C4.5 algorithm to obtain original decision rule sets, and the sets are used as the primary solution of SAA which ensures the high quality data source.

(2) New decision rule sets are produced. SAA is used to optimize the rule sets obtained by the above steps. With lower target function *f(x)*, SAA makes a better solution decision rules. The process can be modified or abandoned when the decision rules by SAA is too complicated.

5.2 Results

A German credit database is used to verify the proposed algorithm. The database consists of 1000 clients' information, of which 700 are "good" customers and 300 are "bad" customers. Each client has 20 attributes, among which the 2nd, 5th, 8th, 10th, 13th, 16th and 18th attributes are continuous and the others are *discrete*. Table 1 shows the attributes of the database:

Table 1: Attribute of German credit database:

Attribute	
1. Current account	11.Current living status
2. Account duration with the bank	12.Asset info.
3. Loan history	13.Age
4. Loan purpose	14.Other install payment
5. Loan amount	15.Housing info.
6. Savings account with the bank	16.Current balance at the bank
7. Current job since when	17.Job info.
8. Loan payment percentage of the monthly income	18.Dependents
9. Personal info. & gender	19.Telephone
10.Other debt & Guarantee	20.Nationality

Let Φ represent the target function to obtain C4.5 decision rule and Ω represent the target function of SAA. Table 2 to Table 4 shows the experimental results with the three different target functions. (the classifying results by C4.5 are *a*=625, *b*=75, *c*=126, *d*=174):

Table 2: Experimental results according to Equation (2)

Φ	α	β	*a*	*b*	*c*	*d*	Ω
0.149	1	0.1	700	0	300	0	0.1
0.527	1	1	547	153	86	214	0.505
0.611	1	1.2	561	139	99	201	0.594
1.367	1	3	0	700	4	296	1.04

As shown in Table 2, when the ratio of β and α is around 1.1, the overall error rate is 24% which is higher than C4.5's error rate (20.1%). However, due to a more stringent target function selected by SAA, it greatly reduces the rate of misidentifying bad clients as good ones. Statistic shows that the cost of misclassifying "bad" customer as "good" customers is 5~20 times of that of misclassifying "good" clients to "bad" clients. When the ration of β and α is approaching to 0 or greater than 2, SAA identifies all the clients either as good ones or as bad ones because of the limitation of the target function.

Table 3: Experiment results according to Equation (3)

Φ	α	β	*a*	*b*	*c*	*d*	Ω
0.1775	0.01	0.01	463	237	75	225	0.1655
0.1024	0.1	0.1	466	234	94	206	0.1004
0.1111	0.2	0.1	466	234	94	206	0.0635
0.0697	0.2	0.2	473	227	94	206	0.0520

As shown in Table 3, when C4.5 is combined with the SAA, the evaluation decision rules are relatively stable. It does not classify all customers into good or bad customers. Even with a stringent target function, the overall error rate for SAA is 34% which shows that the banks lose some good customers.

Table 4: Experiment results according to Equation (4)

Φ	α	β	*a*	*b*	*c*	*d*	Ω
0.8398	1	1	463	237	71	229	0.8194
0.9118	1	1.1	508	192	79	221	0.7686
0.9190	1	1.11	532	168	101	199	0.8757
0.9298	1	1.125	470	230	72	228	0.8420
0.9838	1	1.2	465	235	74	226	0.8954

As shown in Table 4, the performance is better than that in Table 3. However, significant variations exists in values *a* and *c* when the combined reference is slightly modified, which makes the selection of the best reference a little bit difficult.

From Table 2 to Table 4, it can be concluded that mixed algorithm performs much better than that using C4.5 only and the choice of target function has a great influence on the effectiveness of algorithms.

6. Conclusion.

This paper presents a new credit evaluation method based on decision tree and SAA. It applies SAA to credit prediction. It discusses coding, generating new solutions and target function, and then compares 3 sets of target functions by experiments. In this paper, the SAA is not directly involved in decision tree, but employed as the decision rules produced by C4.5. The future work includes combining these two to find a more suitable target function.

7. References

[1] Xiangyang Xu, Jike Ge. "Research on Personal Credit Scoring Model based on Clustering" [J]. *Microcomputer Information*, 2006, (27):229-231

[2] Lishan Kang , Yun Xie, Zuhua Luo, *Non-Numerical Parallel Algorithmic* [M].Beijing: Science Press, 2003

[3] J.R.Quinlan. " Induction of decision trees " [J]. *Machine Learning*, 1986, (1):81-106.

[4] Scott Finnerty, Sandip Sen. "Simulated annealing based classification", *Proc.of the 6th IEEE Int'l Conf on Tools with Artificial Intelligence*[C], 1994: 824-827.

[5] Aihua Shen, Rencheng Tong, Yaochen Deng. "Application of Classification Models on Credit Card Fraud Detection", *Proc. of the 10th International Conference on Service Systems and Service Management*, 2007 pp.1-4.