Ant Colony Algorithm Compared with Other Data Mining Algorithms for Segmenting Bank Credit Customers

Ali Reza Poor Ebrahimi[1], Hossein Farzad[2]

Abstract

This research aims to segment bank's credit customers in terms of received facilities repayment risk. In this research, using corporate customers' data, which are available in information systems of the studied bank credits department, data mining process is carried out. At first, the required data were collected and pre-processing was carried out on them. Influencing variables in model were identified, and ants' colony algorithm was run on the final data, then the results of this algorithm were compared with some of the other algorithm of the category. The obtained results showed that ants' colony algorithm, compared to traditional trees of CART, CHAID, and QUEST, neural networks, distinctive analysis, logistic regression, and Bayesian networks, has better results in terms of the exactness of customers' separation and segmentation, but it is less accurate compared to tree models of C4.5, and supporting vector machine (SVM). On the other hand, ants' colony technique has better prediction power, compared to prediction performance of the bank's validator experts. Finally, a model which has the highest accuracy in predicting the customers' category was recommended as the proposed algorithm. It is noteworthy that in this research, Ant miner software has been used to run ants' colony algorithm

Key words: Credit customers, Data mining, Credit risk, Ant colony

Available online
www.bmdynamics.com
ISSN: 2047-7031

INTRODUCTION

Due to ever-increasing increase of our business needs, volume of business systems data is rapidly growing. Since the cost of data storage is steadily declining, users tend to store all the information in the database to maintain the information that may be useful in the future. Nowadays, banks adopt methods based on new technologies such as data mining and knowledge discovery in database to analyze the customers' needs and behaviors. In the past, some researches have also been done on the customers' credit rating. Among these, we can refer to "Fisher" study (1936), which is based on credit scoring method and is the first evaluation system of credit application [9]. Among other studies, we can refer to the article "Bever" (1967) in the field of companies' success and failure using some of the financial indicators [5]. Also, Altman (1968), as one of the pioneers of validation, attempted to find a significant relationship between accounting variables of a company and likely inability of the company to pay debts in the future, and offered a relationship known as Z-Score. This method was based on linear discrimination analysis between good and bad companies [3].Deakin (1972) has proceeded on using linear discrimination analysis method to evaluate companies' failure factors with the use of 14 financial ratios as independent variables [7], and to evaluate companies' performance using the same model (1989). Fuzzy systems and techniques, and artificial intelligence have been used in data mining [15] [13] [12] [6].

So far, ants' colony has also been used in researches but none of them have not used of ants' colony for segmenting bank credit customers. Raphael Parpineli et al's studies have dealt with the use of ants' colony algorithm in data mining. This research aims to extract classification rules from data sets. In another research by Lio et al, a new kind of algorithm related to ants' colony algorithm has been created which shows better performance than previous algorithm in the two standard data sets [11]. In a study, Nicholas Holden and Freitas have discovered classification rules using ants' colony algorithm in web mining, and the results showed that ants' colony have had higher performance in the two data sets of Yahoo and BBC sites than C5 tree [14].

[1] *Assistant Professor, Department of Information Technology, Science and Research branch , Islamic Azad University ,Tehran ,Iran.* poorebrahimi@gmail.com

[2] *Graduated of MA , Department of Executive Management, Science and Research branch , Islamic Azad University ,Tehran ,Iran.* h_farzad62@yahoo.com

The aim of this research is to identify factors influencing the credit behavior of customers applying for bank credit facilities given the existing data bases in Saman Bank, and to search existing patterns among data of customers who previously received credit facilities using the classification method of applicants for bank credit facilities with the use of ants' colony and other data mining algorithms. The present research attempts to use this algorithm, and compare its prediction results with data mining algorithms concerning credit customers of an Iranian bank. Therefore it can be considered as the innovative aspect of the plan. In what follows, first, review of the literature will be explained, second in the research method section, the steps of this research will be clarified, third in the theoretical framework section, research data will be stated, and fourth in the research findings, model accuracy, evaluation of model credit, model function in relation to other models, and in comparison with the bank's validator experts performance will be described.

LITERATURE REVIEW

Data Mining and Ants' Colony Algorithm

Data mining is adaptation or extraction of knowledge of a set of data [8]. In other words, data mining is a process which extracts knowledge of a set of data using intelligent techniques. The data set which are processed, are known as training set. The extracted model is presented in the form of models, patterns or rules. These models, patterns and rules are different forms of presenting extracted knowledge. This knowledge can be a criterion for future decisions, later performances, or necessary changes in system [2]. Classification method and using decision tree algorithms is one of the methods of knowledge discovery which is usually used in data mining. Decision trees can produce understandable rules and even in a big or complicated tree, a path can be traversed easily, and this makes the interpretation of classifications or predictions relatively easy. Different algorithms have been introduced for making decision trees among them we can refer to the methods of AID, SERCH, CHAID, CARD, OC, ID3, C4.5, C5, QUEST, and SAS algorithms [1].

Decision trees attribute symbolic decisions to samples. However, although various methods have been presented for constructing decision trees, and although applying these methods in symbolic domains has been very successful, sometimes, symbolic decision trees will not have good efficiency, for example, when a numerical decision is required, or when numerical decision making can improve further processing [10]; or confusion in the training set including confusion or lack of attributes' value in describing samples [15]; or when fuzzy classification is needed.

Ants' colony algorithm is among collective intelligence algorithms in which people work together to achieve a final goal. Now, working on the development of intelligent systems which is inspired by the nature is among the most popular fields of artificial intelligence. Ants' colony algorithm was firstly presented by Dorigo et al. as a multi-agent factor solution for the issues of optimization problem such as travelling salesman problem (TSP).

Ants' colony algorithm is inspired by the studies and observations on ants' colony. These studies have showed that ants are social insects that live in colonies and their behavior is more toward the survival of the colony rather the survival of one of them. Ant algorithms try to use natural insects' intelligence with the simulation of the facts as a solution to solve problems in artificial intelligence and computer.

Ants use different ways to survive. The way worker ants' access to food and optimization of the path to reach the food is one of the most important rules that has been considered in ant algorithms. The first series of ants are fed by queen's saliva but with increasing population and expanding community need for new food sources is inevitable. When ants go out of their nets, they search for food in all possible ways and as soon as reaching an acceptable food and taking some of it, they return to the nests. Ants identify the closest path to food with the help of their alphabet, which is pheromone, and communication antenna.

The method to find the path is very suitable and practical which has been used in ant algorithm to optimize the path, and with its help, travelling salesman problem has been resolved. Some ants carry food (full) to the nest or their community, or some of them shred food grains and bring them to the food storage area. Some ants are fed with sap of trees or kitchen garden fruits such as melon and cantaloupe, and some specific species pick leaves of trees and transfer them to wet environments so that they feed with fungi that are formed on leaves.

Some features of ants' social life which have been interested by researchers are as follows:

1. To obtain food and supplies
2. Assign responsibilities and tasks among ants
3. Ants' cooperative patterns to solve problems, for example: Picking up a heavy seed by some ants
4. Protection patterns of the queen's life or valuable items of food and small children.

Obtaining these rules and transforming them to specific computer rules for the use of artificial software operators is a great challenge of algorithms.

Some features of using ants' colony algorithm in classification issues are as follows:

- To use discrete variables.
- To avoid local optimum.
- To apply easily.
- To find good solutions quickly.
- To offer several solutions by the best solutions that will be selected.

Credit Rating

Credit rating is a system by which banks and credit institutions evaluate the likelihood of loan repayment by the applicant using present and past information about him/her. In other words, rating means quantifying the likelihood of default in the future [1].

Various scholars have offered relatively similar definitions of credit ratings. Feldman defines credit rating as allocating a qualitative criterion in the form of a unit or score to a potential borrower to provide an estimate of his/her future performance [4]. According to "Standard and Poor" institution, "rating system is a comment on credit value of a debtor based on risk factors". According to Modes institution, rating is "a comment on future ability of debtor and legal obligation of papers issuer for timely principal and interest payments on fixed-income securities" [13].

Given the complexity of their economic environment and activities, credit institutions and banks should select suitable models for evaluating customers' credit rating.

One of the success factors of credit decisions is the accurate, complete, and updated information. The main needed information are information about the history of payments, information about securities, certain credit information about the owner or owners of firms, firm's financial information, and information about economic indicators. But, financial ratios are of the most widely used of this information for the loan and credit office of the bank [11].

For credit rating, different criteria are used such as 5C criterion which includes character, capacity, capital, collateral, and credits and facilities' terms and conditions, and LAPP criterion which includes liquidity, activity, profitability and potential ability, or 5P criterion which includes people, product, support, and payments, and overall future overview [9].

Among the benefits of credit rating for customers we can refer to much easier credit process, response in a shorter time frame, reduction of the amount of information needed, and faster and easier access to credit when customers need it, and among the benefits of credit rating for banks we can refer to reducing the costs of evaluating loans, providing a standard loan granted in bank, and increasing the efficiency of loan granted.

METHODOLOGY

Present research has been based on knowledge discovery from databases of the studied banks, so universal standard of CRISP-DM has been used to do research process which consists of understanding business issue, understanding data, preparing data, modeling, evaluating results, and applying the model.

In this research, after collecting data of the former customers of the bank from the respective databases, and refining data, influencing variables in customer rating were identified through interviewing with the related experts and scientific documentation. After that, given the definition of good pay or bad pay customer, a class label with the same definition has been intended for all final good customers. In the next step, customers were classified based on their features using the technique of ants' colony and the respective soft wares. Then, classification was done with other algorithms and its accuracy was compared with the classification of ants' colony algorithm. Finally, rules and patterns in the data of the customers were found based on the classes defined and presented as a framework for predicting the credit of the new applicants so putting applicants in the defined classes can be predicted.

Implementation stages of this research can be summarized as follows:

1. Collecting data from the available databases;
2. Identifying the influencing factors in credit behavior of customers which are available in the studied databases;
3. Determining indicators for defining the classes of good customers (good pay), middle, and bad customers (bad pay);
4. Preparing data format for putting them in the respective soft wares;
5. Dividing sample data to two sets of educational and test data;
6. Creating rules using experimental data with ants' colony algorithms and data mining;
7. Model test with test data set;
8. Presenting discovered pattern from customers' classification;
9. Evaluating model's credit and comparing it with the classification models.

Research Methodology

Data was collected by both "observation" and "interview" methods. In a way that the required information has been observed and recorded in check list to access the information of the bank's customers from different available databases such as files and computer systems. Also, open interview with experts of the bank has been used to identify the variables affecting customer's credit behavior and verbal variables.

Statistical population of this research consists of 418 cases of the bank's corporate customers which received credit facilities during fiscal years of 1387 and 1388.

Given that only these data were available to the researcher and access to all data of the bank's corporate customers for sampling was not possible and according to remarks of the bank's credit section managers, all branches across the country have the rest of data as scattered and non-coherent, this population has not been sampled, and after refining their data, 120 cases has been used to build the final model.

These data have been stored in a database in Excel software. Customers' features, which are independent variables of the research, and customer's classification, which is dependent variable of the research, form fields of this database. Titles of this field consist of 19 customer's features including company’s age, company’s type, the amount of capital, the field of activity, director's education, director's age, type of loan agreement, loan's interest rate, loan amount, loan term, type of loan, repayment type, the profitability ratio, the number of employees, work experience, collateral power, the current ratio, debt ratio, and claims' collection period, and nominal variables with respective arms are as table 1.

Insert table-1 here

Following the process of preparing, two important operations of reducing data and applying changes in the data form were carried out on the relational database for cleansing and preprocessing data.

Data cleansing has been carried out in three major parts: 1) correcting user's errors, 2) homogenizing data, 3) nominalizing variables. Users' errors in data entry have been identified with observing unacceptable data, matching with other data, and comparing common data in different databases. For integrating the data of some fields and making some data processable in the model, data should be uniformed. From among fields that were changed in this stage were the field of "loan term" which was in month and year and all of them were changed to day; or the field of the years of company's activity which were in year and changed to the number of activity's years.

Since ants' colony algorithm needs nominal variables for classification, quantitative and continuous variables should change into discrete groups. According to experts' comments, these groups were classified into high, low and middle groups in table 2. Regarding the studies which were mentioned in the reports of the bank's experts (they are available in archived records) and being informed of the experts' work, we came to this conclusion that the used categorization for classes can be interpreted as good, average, and bad, and for independent variables (features) as high, middle, and low. In the next stage, experts were asked to express the spectra of customers and variables classes (features) which consider in their decisions with these 3 conditions.

Insert table-2&3 here

Given the amount of "delay in repayment of loan installments", a level has been considered for each class (good, bad, average) as the basis in each record for each customer.

5. Research Findings

Since the method presented in each research should be evaluated in terms of validity, in this research, given that the research method is "data-oriented", the validation method is in a way that data are divided into two sets of training data and test data. The aim of this work is that selection algorithm will gain knowledge with the number of training data. But, the amount of validity of the results should be tested by the results of new data and the prediction power of algorithm about the data it has not been encountered with. In this way, test data are given to algorithm as observer data, and the results can evaluate the amount of model's accuracy.

The criterion of model's validity and accuracy evaluation is the accuracy of classification or separation of test data in classes. In this research, "cross- validation with 10 repetitions" has been used. This validation method divides data set into 10 parts, and each time selects 90% of data as training data set and 10% as test data set, finally evaluates the amount of classification accuracy. This process is repeated 10 times, and as a result, all levels of accuracy are averaged and presented as the model's final accuracy. Using this method, there will be no concern about random selection of training and test data sets.

Given the number of tested repetitions (times algorithm is applied on data), the amount of algorithm's accuracy (in terms of correct classification of customers) is as table 4.

Insert table-4 here

The rule can be extracted according to the number of routes from the root (top node) to a leaf.

To assess the performance of the model presented by ants' colony algorithm, the accuracy of the results of this model has been compared with the algorithm of another classification. These algorithms are as follows:

Decision tree C4.5, decision tree CART, neural network MLP, neural network RBF, SVM, Bayesian network, K nearest neighbor, and logistic regression that the results of their classification have been shown in table 5.

Insert table-5 here

As it is clear, the resulted rule of C4.5 algorithm with the accuracy of 68.3% has the highest amount of accuracy, and the resulted rule of KNN algorithm with the accuracy of 45.83% has the lowest amount of accuracy. The tree resulted from modeling is considered as the extracted knowledge of the research. Now, regarding the comparisons made, the technique which predict more accurately than the others are suggested for the use of the bank's efficients.

CONCLUSION

The present research is an applied one which has been studied with a case study in a bank and modeling of the research process and its results can be useful for all banks and other financial and credit institutions. In this research, data mining and ants' colony algorithm have been used for the classification of the bank's customers.

The direct results of the research implementation can be summarized as follows:

- Ants' colony algorithm classified customers in three predetermined classes with the accuracy of 65.8%.
- C4.5 decision tree has the higher accuracy than ants' colony algorithm and other data mining classification algorithms used in this research.

- Comparing with decision tree and SVM, Ants' colony algorithm has the lower efficiency in terms of the accuracy of customers' separation (in three classes) but it has the higher accuracy than the other studied algorithms.
- The use of ants' colony algorithm achieved far better results than the performance of validation experts of the studied banks. And this shows high efficiency of the model used in this research in comparison with the efficiency of the banks validation experts who predict about customers based on the experience and judgment.

Also, other results of this research are the reduction of human errors and prevention of the experts' judgment process by using the discovered patterns in an intelligent system and faster response to applicants for credit facilities. Among the limitations of the present research are the lack of coherent information in a computerized information system, the lack of unity in the record of customers' data and data record in a given time series, incomplete and incorrect data in the studied data which face data with too much noise.

REFERENCE

[1]Eslami Nosratabadi.H, Pourdarab. S,Nadali.A, “A New Approach for Labeling the Class of Bank Credit’s Customers in Classification Method with Data Mining Approach”, International Journal of Information and Education Technology, Vol.1, No.2, June, 2011.

[2]Nadali.A, Pourdarab.S, Eslami Nosratabadi.H,”A hybrid Method for Credit Risk Assessment of Bank Customers” International Journal of Trade Economics and Finance(IJTEF),Vol.2, No. 2, April, 2011.

[3]Altman E.I, “Financial ratios discriminate analysis and the prediction of corporate Bankruptcy”, The Journal of finance 23, 1968.

[4] Banasik, J., Crook, J. and Thomas,L., “ Sample selection bias in credit scoring models ”,Journal of the Operational Research Society ,vol.54,2003, pp. 822–832.

[5] Beaver W.H., “Financial ratios as Predictors of Failure” , Journal Of Accounting Reserch, 1967.

[6] Chiang I.J., and Hsu J.Y., “Fuzzy Classification Trees for Data Analysis,” Fuzzy Sets and Systems, vol.130, no.1, pp.87-99,2002.

[7] Deakin E.B, “A Discriminate analysis of predictors of business failure”,journal Of Accounting Research 10(1), 1972.

[8] Edward F.R , Mishkin F.S., “the decline of traditional banking: implication for financial stability and regulatory policy”, Federal reserve bank of New York policy Review, 1995, pp.27-45.

[9]Fisher R.A, “ The Use-of multiple measurement in Taxonomic problem”, Annals of Eugenics, 1936

[10] Janikow C.Z., “Fuzzy Decision Trees: Issues and Methods,” IEEE Trans. on Systems, Man, and Cybernetics, vol.28, no.1, pp.1-14,1998.

[11]Lio,Bo and et al,(2004),” Classification Rule Discovery with Ant Colony Optimization”, IEEE Computational Intelligence Bulletin, Vol.3 No.1,p31-35.

[12] Mikut R., Jens J., “Interpretability in Data-based Learning of Fuzzy Systems”, Fuzzy Set and Systems, Elsevier, 2004.

[13] Moon, C. G. and Stotsky, J. G., “Testing the Differences Between the Determinants of Moody's and Standard & Poor's Ratings: An Application of Smooth Simulated Maximum Likelihood Estimation” , Journal of Applied Econometrics, Vol. 8, No. 1, 1993, pp. 51-69.

[14]Nicholas, Holden ; Freitas A. Alex,(2005)” Web Page Classification with an Ant Colony Algorithm”, Computing Laboratory, University of Kent , UK.p1-10.

[15] Olaru C. and Wehenkel L., “A Complete Fuzzy Decision Tree Technique,” Fuzzy Sets and Systems, vol.138, no.2, pp.221-254,2003.

[16] Parpinelli,s.Rafael and et al,(2002),” Data Mining With an Ant Colony Optimization Algorithm”,IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTING, VOL. 6, NO. 4, p331-332.

Table 1. Arms of Nominal Variables

Respective arms	Nominal variables
Commercial, service, manufacturing	Field of activity
Cooperative, public joint stock, private joint stock, LTD	Type of company

Limited partnership, civil partnership, general limited partnership, reward, forward, debts purchase, installment purchase	Type of loan agreement
Low (diploma and under diploma), Middle (associate degree , and bachelor), high (master and above master)	Director's education
Domestic commercial, production - industry, importation	Type of loan
Together, Installment	Type of repayment

Table 2. Spectrum of variables in verbal quantities

	low	Middle	High
Company's age	1--5	6--14	15--54
The amount of registered capital	1 – 10,000	10,000,--100,000	10,000 – 1,000,000
Director's age	20--35	36--45	46--82
loan's interest rate	4--14	15--19	20--28
Approved amount of loan	0,1 – 1,000	1000 – 10,000	10,000 – 200,000
Loan term	18--179	180-365	366-1460
collateral power	0.5 – 2.9	3 – 3.9	4--5
work experience	0--1	2--4	5 – 7
Current ratio	0 – 1.2	1.2 – 1.4	1.4 – 5.7
Debt ratio	65.00--	0.65 – 0.8	33.1 – 8.0
Claims' collection period	09--0	0150--9	1463--150
Profitability ratio	0 – 0.1	0.10 – 0.2	0.2--064
Number of employees	4--49	50--199	200--1269

Table 3. Spectrum of classes in verbal quantities

	Class of good customer	Class of average customer	Class of bad customer
Delay in loan installment repayment (day)	0--60	60--180	180--700

Table 4. The amount of ants' colony algorithm accuracy

Number of repetition	*30*	*50*	*100*	*150*
Accuracy	*58.2%*	*55.5%*	*65.8%*	*61.3%*

Table 5. Comparison of the results of algorithm's accuracy

Algorithm	*Ant Colony*	*C4.5*	*CART*	*MLP*	*RBF*
Accuracy	*65.8%*	*68.3%*	*54.2%*	*58.3%*	*53.3%*
Algorithm	*KNN*	*SVM*	*BayesNet*	*NaiveBaves*	*Logistic-R*
Accuracy	*45.83%*	*66.7%*	*51.67%*	*50.83%*	*46.67%*