Indian Journal of Fundamental and Applied Life Sciences ISSN: 2231– 6345 (Online)
An Open Access, Online International Journal Available at www.cibtech.org/sp.ed/jls/2015/01/jls.htm
2015 Vol.5 (S1), pp. 2939-2947/Moradi
Research Article

PROVIDE A NEW METHOD FOR VALIDATION CUSTOMERS BY PARTICLE SWARM ALGORITHMS AND NEURAL NETWORKS COMBINED

***Elaheh Moradi**
Department of Accounting, Khomein Branch, Islamic Azad University, Khomein, Iran
**Author for Correspondence*

ABSTRACT
Improvement in the application and data collection, using computer and also public uses of web and internet as world informing system, introduce data mining as a useful instrument to help solving complicated problems. In recent years, validation is changed to an important issue for financial institutes such as banks and insurance companies. The validation issue poses as a classification problem and is aimed at extraction suitable and efficient model which categorize customers according to their characteristics into two categories: good and bad. In the past in order to create validation issues, classical methods of data mining like linear regression and the analysis of linear point was used. But recently instruments of intelligence emotion and machine learning are noticed by researchers of validation issues. Different models, abstract and hybrid, is posed for this problem in order to create a high performance model. In this regards efforts are going on. The purpose of this thesis is identifying the factors (variables) influencing the behavior of costumers applying for post banks credit facility and provide a model for customers' validation, using neural network and particle swarm algorithm. Datasets used by post bank of Tehran, for each of these methods measure the error rate and the execution time will be calculated and compared. The results indicate the superiority of the combined neural network model and Particle Swarm Algorithm and Matlab software is used for simulation. Finally, conclusions and suggestions for future research in this thesis have ended.

Keywords: *Data Mining, Validation, Neural Network, Genetic Algorithm, Particle Swarm Algorithm*

INTRODUCTION
The advent of information and communication technology, and overflow of information about customers on one hand, daily development of credit facilities market, and also increased rate of delayed or irreversible demands, have led the credit institutions to exploit some new models in order to evaluate credit of their customers.
In recent years, the field of data mining has grabbed the attention of scientific and industrial communities of validation. In fact, since methods of information gathering and saving them in today's world is simple and cheap, it has led various businesses to face a large volume of data and seek some methods and techniques, so that they would be able to use these collected information (Allen, 1995).
Among tasks of data mining, we can refer to clustering, classification, and data visualization, analysis of deviation and rules of necessity. Validation is a method for identifying a group or different groups in a population that demands loan or credit card. According to this, validation is in the category of issues of classification and clustering and since classification and clustering are considered as tasks of data mining, thus we can use data mining for validation (Ang *et al.*, 1979).
Stewart (1991) believes that the profit, earnings per share and earnings growth, are misleading measure of corporate performance and economic added value is the best measure of performance evaluation. After this claim so many studies have been conducted to evaluate this issue. In the studies the relationship between economic added value with market added value and its relationship with the traditional criteria for performance evaluation were compared. Financial institutions, banks and insurance companies, with having a proper validation system, can increase satisfaction of loan or credit card applicants on one hand and on the other hand, by attracting good customers, they can decrease the risk and consequently increase their profit rather than following the goal of putting bad customers aside.

Those customers are considered as good ones that do their commitments in the appointed time after receiving credit facilities and the customer who does not do his/her commitments towards facilities is referred to as a bad customer.

Statement of Problem

In recent years, improper management have enforced many damages on financial institutions such as banks and insurance companies specially in the USA and Europe and even it has driven some of them to the extent of bankruptcy.

Only in the USA, since the year 1997, exporters of credit cards have reported 19.27 million dollars debt as damage and this number has increased to about 91.40 million dollars in the year 2004 (Zhang and Bhatta, 2004). Due to the nature of their activities, the banks have faced all kinds of risks from the beginning such as credit risk, market risk, assets risk, operational risk, risk of liquidity and risk of rate and although incoherently, they have tried to identify and manage these risks.

Providing bank facilities has faced some problems such as credit risk of customers in the recent years and has made validation necessary for banks.

Table 1: Variables of the research

Row	Name of variable
1	Value of collateral
2	Type of collateral
3	Loan proceeds
4	Loan rate
5	Period of loan
6	Maturity date (year, month, day)
7	Type of facilities
8	Amount of payment
9	Age
10	Housing status
11	Marriage status
12	Paid payments
13	Number of paid payments
14	Belated payments

Credit risk is one of the most important factors of production of risk in banks and financial companies. Default risk happens when the borrower of the loan, due to lack of power or willingness, doesn't keep his/her promises to the lender of the loan at the maturity date. This risk is one of the oldest and most important risks, which specifically influences financial and monetary institutions.

Why default risk of a few numbers of customers can bring so many damages to the organization; therefore, the most important and most effective tools that the banks need for managing and controlling risk is the scoring or validation system.

This system is known as one of the methods of risk evaluation. If the bank faced a problem in identification of customers and giving loans to them, it will face a huge loss.

Pending demands lead blockage of limited bank resources and make their future profitability problematic; thus, credit validation system increases the effectiveness of credit decisions and leads to the reduction of costs and defaults of the borrowers.

This system is one of the ways of quantifying and measuring credit risk. Scoring system based on features and performance of the borrowers predicts performance of future loans and therefore, they are used for proper management of credit risk.

Statistical Population

All of the true customers, who demand post bank facilities of the city Tehran in the years 90, 91 and 92, form the statistical population which is 828 persons.

Methods of Data Collection

Informational data of this research has been provided from the cases of bank customers in the form of data series. This data series includes 15 features and 828 records.

Operational Definition of Variables

Variables that shall be determined and measured in this research are visual and can be understood and measured well, in such way that data associated with these variables has been collected through the bank. These variables have been provided in the table above (table 1).

Variable of age shows the age of the applicant, as for housing, it is associated with the two ownership and rental conditions. Marriage status includes the conditions of being single, married and divorced. The value of collateral is calculated by considering the type of collateral (in the category of installment sales, it is interest free loan) as 15 or 20 percent of principle and interest. The type of rate is the interest that the lender takes from the principle, for example in interest free loans it is 4%.

Loan proceeds: it is in fact the amount of the loan that has been given to the customer by the bank. Paid payment is in fact the amounts that have been paid out of the whole loan. Number of paid payments is the number of payments that the customer has paid until now.

Period of loan: it is the months that the customer shall pay the payments.

Maturity date: it is the date at which the customer shall pay the loan.

Amount of payment: it is the amount that the customer shall pay in each payment.

Belated payment is in fact a field which is specified according to on time-paying and late-paying customers which has two values 0 and 1. When the field of belated payments is 0, it means that the customer has been on time-paying and when the field of belated payments has the value 1, it means that the customer has been a late-paying one.

Questions of the Research

1. How can we estimate the factors which affect the credit of the customers of Tehran's post bank by using data mining techniques?
2. Is the error rate of validation model of neural system with particle swarm algorithm more or the validation model of neural system with genetic algorithm?

Descriptive Information of Data

Firstly, data is processed in the software Excel. This process includes identification and deletion of unrelated data, converting nominal data to numerical data, deleting unknown data and etc. And after doing the above items, we normalize the obtained data serious, which is only numerical, by using the following formula. This means that we put all of the values of data in the ranges [0 1] and [-1 1]. In this formula, the inlet value is shown with x and it puts all data in the range [-1 1] and if we want the data to be put in the range [0 1], we use formula (Huang et. al., 2006).

Formula 1 $$X_{Normalize} = (X - MinX)/(Max\,X - MinX) * 2 - 1$$

Formula 2 $$X_{Normalize} = (X - MinX)/(Max\,X - MinX) * 2$$

Selection of Features

There are 15 features in this issue. They are given to the neural network as input and the output of the program has two modes, it is either 0 or 1. These 15 features are the characteristics of the bank's customers and the output 0 means bad customer and the output 1 means 1. By using the data that we have received from the bank, we educate the artificial neural network, in such way that after education, by receiving 15 features of each customer, it would be able to print the goodness or badness of the customer as the output. However, it is clear that some of these 15 features have more impact on the outlet and also some others don't have an accurate and regulated effect on the output. So among these 15 features, we shall choose those that have more impact on the output and by using them we shall educate the neural network and delete those features that are not useful. In order to choose proper features, we use genetic algorithm and with particle swarm algorithm and we will compare the results with one another. In order to do this, we consider a linear vector with 15 members that each of its members correspond with one of the customers' features. We name this vector x and each member of this vector can be 0 or 1. 0 means that

this feature shall not be chosen and 1 indicates that it shall be chosen. For example, if the third element of vector x was equal to 1, it would mean that feature number 3 is among selective features and for instance, if 14th element was equal to 0, it would mean that we shall not select 14th feature. So the issue of finding distinctive features turns to finding the proper x vector.

Choosing Features with Genetic Algorithm

In order to find the proper vector x that reduces the error of neural network in categorizing bank's customers, first we use genetic algorithm. First, we produce a 30-member population of random x vectors. As it was mentioned before, this vector has 15 binary members (0 or 1). After that, by using genetic operators, mutation and crossover of the second generation are created. In this issue, the first 6 members of each population (which are 6 members that create the lowest value of cost function) directly go to the next generation and 20 members are created by the crossover operator and also 4 others by creating mutation in them. This process continues until 40 generations. In order to specify the rate of cost function for each x vector, it is necessary to educate the network with the chosen features in it and then evaluate the educated network with education and test data and consider the error rate of neural network in categorizations of the customers as the rate of cost function of that x vector. Selection of features was explained above. And in propagation networks, one of the default functions for educating the network is **Levenberg-Marquardt** function. In this research, we used propagation network with two layers, in such way that in the first hidden layer the nonlinear Tansig function with 5 neurons and in the output layer the linear Hardlim function with one neuron were used and we educated the network and the results have been shown in Table. 2. In this method, in the best mode we achieved the 13.14% education error and 14.45% test error and 13.40% total error with the 636.8 seconds of operation time. The features that were effective in this mode of classification model were: type of collateral, loan rate, amount of payment, housing status, paid payment and number of paid payments.

Table 2: Choosing features with genetic algorithm

Number of population	Number: [Mutation Crossover]	Number of generation production	Minimum of number of features	Selective features	Education error	Test error	Total error	Operation time (second)
30	[4 20]	40	1	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15	94.10%	9.96%	93.47%	782.3
30	[10 10]	40	1	2,4,10,12,14,15	13.14%	14.45%	13.40%	636.8
30	[10 10]	40	7	2,3,4,6,7,8,9,10,11,12,13,14,15	24.51%	21.68%	25.22%	178.8

Choosing Features with Genetic Algorithm and Education by Using Particle Swarm Algorithm

In this combined model, first selection of features is done by genetic algorithm (was explained above) and after building propagation network with two layers with 5 neurons in the hidden layer and 1 neuron in the output layer, transmission functions Tansig and hardlim are created (Olafsson and Wu, 2008).

First, the effective features are determined through genetic algorithm and in order to educate the network (determining weights and biases) the particle swarm algorithm is used. The results have been provided in Table. 3. In the best mode in this model of particle swarm algorithm, out of a population with 20 members with a number of 10 repetitions, with the features loan rate, loan's maturity year, amount of payment, marriage status, paid payment and operation time in this algorithm is 3983.6 seconds.

In comparison of this model with the model made above (choosing features with genetic algorithm and educating the network by the (Levenberg-Marquardt) function, we conclude that the obtained error has halved.

It means that when the education of network by is done an optimizing algorithm, we reach higher accuracy (less error); but the above model obtains answer in the operation time of 636.8 seconds which is far less compared to the operation time of this model.

Table 3: Choosing features with genetic algorithm and education by using particle swarm algorithm

Number of population	Number: [Mutation Crossover]	Number of generation production	Minimum of number of features	Parameters of population repetition [PSO]	Selective features	Education error	Test error	Total error	Operation time (second)
30	[10 10]	40	1	[10 5]	2,7,9,12,13,14	6.49%	6.62 %	6.52 %	1411.5
30	[10 10]	40	1	[20 8]	3,6,10,13,14,15	6.04%	6.62 %	6.15 %	3983.6
30	[10 10]	10	7	[4 10]	2,3,4,5,6,8,10,11,12,13	10.57%	13.85 %	11.23 %	356.21

Choosing Features with Genetic Algorithm and also Education of Network with Genetic Algorithm

In this combined model, first selection of features is done by genetic algorithm (was explained above) and after building propagation network with two layers with 5 neurons in the hidden layer and 1 neuron in the output layer, and transmission functions Tansig and hardlim are used.

Table 5: Choosing features with genetic algorithm and also educating network with genetic algorithm

Number of population	Parameters of GA selective features: [Mutation Crossover]	Number of generation production	Minimum of number of features	Parameters of GA education with [population Max-Gen Mutation Crossover	Selective features	Education error	Test error	Total error	Operation time (second)
30	[10 10]	40	7	[10 10 4 2]	2,6,8,9,10,11,14,15	5.89%	9.63%	6.64%	536.7
30	[10 10]	40	7	[30 20 4 2]	2,4,5,6,7,10,12,13,14,15	5.58%	10.24 %	6.52%	4018.8
30	[10 10]	10	7	[30 4 15 5]	1,3,5,6,7,8,10,11,12,13,14	6.64%	6.02%	6.52%	941.75

This network is educated by the genetic algorithm for doing the act of classifying (Zhang and Bhattacharyya, 2004).

In fact, in order to determine the weights and biases, the genetic algorithm is used so that the best weight and bias would be selected for the act of classification and the results have been provided in Table. 5.

As it is obvious in table, in the best mode of this model, the genetic algorithm is for an education with a preliminary population of 30 and the maximum number of 4 generations and 15 crossovers and 5 mutations which is formed by features value of collateral, amount of payment, period of loan, year and month and day of loan's maturity, amount of paid payments of the loan, age, housing status, marriage status and paid payment of the loan; the education error is 6.64, and test error 6.02 and total error 6.52 and its operation time is 941.75 seconds.

In comparison with the two made models with same conditions but genetic and particle swarm algorithm for educating neural network, we come to this conclusion that models that have used particle swarm algorithm (genetic algorithm) for educating the network have less error rate (more error) but longer time of operation (shorter operation time).

In order to find the proper x vector that would reduce the error of neural network in classification of bank's customers and identify the effective features in this model, in this phase we use particle swarm algorithm. At first, we produce a 30-member population of random x vectors. As it was explained in the section of selection of features, this vector has 15 binary members (0 or 1).

Then, all particles are dispersed in the searching space and each particle shows its own position and speed based on its best spatial and general position found in the searching space of the issue. In order to specify the rate of cost function of each x vector, it is necessary to educate the network with the chosen features in it and then evaluate the educated network with test and education data. Error rate of neural network in classification of customers shall be considered as the value of cost function of that x vector.

Choosing Features with Particle Swarm Algorithm and Education with the Default Function of Propagation Networks

In this combined model, first selection of features is done by particle swarm algorithm (was explained above) and after building propagation network with two layers with 5 neurons in the hidden layer and 1 neuron in the output layer, transmission functions Tansig and hardlim are created.

And this network is educated with the default function Levenberg-Marquardt for doing the act of classification.

Table 6: Choosing features with particle swarm algorithm

Number of population of algorithm in PSO of selection features	Maximum of repetition in PSO of algorithm of selection features	Minimum of number of features	Selective features	12.53 %	Test error	Total error	Operation time (second)
30	5	1	1,2,3,4,5,7,9,10,11,12,13,14,15	31.26 %	12.04 %	12.43%	43.7
30	10	1	1-15	82.47 %	24.09 %	29.83%	70.7
30	10	7	2,3,5,6,9,10,11,13,14	12.53 %	81.32 %	82.2%	79.5

In fact, in order to determine weights and biases, this function is used so that the best weight and bias would be selected for the act of classification. The results have been provided in Table. 6.

As it is obvious in table, in the best mode of this model, the particle swarm algorithm is for an education with a preliminary population of 30 and the maximum number of 5 repetitions with features value of collateral, type of collateral, amount of loan, period of loan, month of loan's maturity, rate of loan, amount of paid payments of the loan, age, housing status, marriage status and number of paid payment and amount of the loan is formed; the education error is 12.53, and test error 12.04 and total error 12.43 and its operation time is 43.7 seconds.

In this combined model, first selection of features is done by particle swarm algorithm (was explained above) and after building propagation network with two layers with 5 neurons in the hidden layer and 1 neuron in the output layer, transmission functions Tansig and hardlim are created. And this network is educated with the default function Levenberg-Marquardt for doing the act of classification. In fact, in order to determine weights and biases, this function is used so that the best weight and bias would be selected for the act of classification (Olafsson and Wu, 2008). The results have been provided in Table. 7.

As it is obvious in table, in the best mode of this model, the particle swarm algorithm is for an education with a preliminary population of 30 and the maximum number of 5 repetitions with features value of collateral, type of collateral, amount of loan, month of loan's maturity, rate of loan, amount of paid payments of the loan, marriage status and number of paid payment and paid amount of the loan is formed; the education error is 6.79 and test error 5.42 and total error 6.52 and its operation time is 627.5 seconds. However, when the preliminary population is 25 members and the maximum number of repetition of particle swarm algorithm is in 5 educations, we obtain the same error rate but with the time of 2311.6 seconds; which means that when the preliminary population increases in this algorithm, our error rate won't be reduced but the time of our program's operation will be longer.

Table 7: Education with particle swarm algorithm

Number of population of algorithm in PSO of selection features	Maximum of repetition in PSO of algorithm of selection feature	Number of algorithm's population in PSO education	Maximum repetition of algorithm's population in PSO education	Minimum of number of features	Selective features	12.53%	Test error	Total error	Operation time (second)
30	10	10	4	1	1,2,4,6,7,9,10,13,14,15	6.79 %	5.42 %	6.52 %	627.5
30	10	25	5	7	1,2,4,5,7,8,10,11,12,13,14,15	7.09 %	4.21 %	6.52 %	2316.5

Choosing Features with Particle Swarm Algorithm and Education with Genetic Algorithm

In this combined model, first selection of features is done by particle swarm algorithm (was explained above) and after building propagation network with two layers with 5 neurons in the hidden layer and 1 neuron in the output layer, transmission functions Tansig and hardlim are created (Yueh and Chun, 2006). And this network is educated for doing the act of classification. In fact, in order to determine weights and biases, the genetic algorithm is used so that the best weight and bias would be selected for the act of classification (Yueh and Chun, 2006). The results have been provided in Table .8.

In the best mode of this model, the particle swarm algorithm is for an education with a preliminary population of 20 and the maximum number of 5 repetitions and genetic algorithm with a preliminary population of 30 members and a maximum number of 4 for the education of the network by selection of

the features value of collateral, period of the loan, month of loan's maturity, age, housing status, marriage status and paid payments is formed and used with the education error of 6.94 and test error 4.21 and total error 6.4 and its operation time is 452 seconds.

Table 8: Choosing Features with Particle Swarm Algorithm and Education with Genetic Algorithm

Number of population of PSO algorithm	Maximum of repetition of PSO algorithm	Number of GA population	Maximum repetition of generation production	Minimum of number of features	Selective features	12.53 %	Test error	Total error	Operation time (second)
20	5	10	4	1	1,5,6,10, 11,12,13, 15	6.94%	4.21%	6.4%	452.3
30	5	25	5	7	1-15	6.79%	5.42%	6.52%	649

Conclusion

Computational results show that using particle swarm algorithm in selecting features and also in education of neural network has a better result than genetic algorithm. It means that in fact, a model that is made by using particle swarm algorithm and neural network has higher accuracy than the model created by neural network and genetic algorithm. By comparing the two following tables, it is clearly obvious. Comparing error rate and operation time of the made models has been presented in the following two tables. In table number 1, selection of features has been done through genetic algorithm, and in table number 2, selection of features has been done by particle swarm algorithm.

Table 9: Selection of features by genetic algorithm

Neural network education	Error percentage	Operation time
1-Levenberg-Marquardt function ("Matlab" function)	13.40	636.8
2- genetic algorithm	6.52	941.75
3-particle swarm algorithm	6.15	3983.6

Table 10: Selection of features by particle swarm algorithm

Neural network education	Error percentage	Operation time (Second)
1-Levenberg-Marquardt ("Matlab" function)	12.43	43
2- genetic algorithm	6.52	627.5
3-particle swarm algorithm	6.4	452.3

This point shall be mentioned that we can conclude that one model that works swell on a data series will work well on another data series as well. Also it is not accurate to use similar specifications for any data series of each country. Although choosing some of they might be unavoidable, but this selection of preliminary specifications for creation of such database requires psychological and sociological studies that will require experience of experts of the field of bank and credit affairs as well.

Recommendations

According to the extracted knowledge and patterns from the available data series, new customers are classified in two categories, good or bad, with the presentation of information; but it is not specified that to some extent they belong to these categories. This issue becomes significant when two customers are

simultaneously applicants of a credit rate and both of them are in the good category. How will one of the two customers are selected then so that it will be in the direction of reduction of risk and increase of profitability. In other words, it is expected from the presented model to have the capability of assigning credit score to the customer. Developing probabilistic or phase models that will determine the rate that one customer belong to the good or bad categories can be useful in this respect. If a customer is in the bad category, how can he/she promote him/herself and be in the good category? Or if a customer mentioned this question that what are his/her weaknesses so that he/she would be able to correct them and would be able to use credit facilities by being in the good category. In other words, perhaps by presenting the extracted rules from the data series based on that classification, we would be able to convince the customer of his being bad and about the credit facilities. Various customers want the facilities such as loan with a time period, various loan amounts and different interest percentage; how shall be deal with these customers? Can we develop a model that would determine what loan with what percentage rate and with what time period of paying back shall be given so that the risk of institution will be minimized? How accurate is the information presented to the financial institutions by customers? Such review does not mean a field review or inquiry in person, because such reviews require a lot of time and cost a lot. If wrong information was presented by the customers, even the best models will present wrong answers. There is another challenge in the issue of validation which is the fact that customers' behaviors differ from one country to another given the economic status of a country, citizen's culture and items as such. In order make these challenges go away, using the experience of experts in the field and psychological and sociological studies in this field is required. Various methods have a different power in predicting whether good categories are more harmful or bad categories and how much is this damage.

REFERENCES

Allen JC (1995). A promise of approvals in minutes, not hours. *American Banker* **28**.

Ang JS, Chua JH and Bowling CH (1979). The Profiles of Late-Paying Consumer Loan Borrowers: An Exploratory Study: Note. *Journal of Money, Credit and Banking* **11** 222-226.

Desai VS and Crook J (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research* **95** 24-37.

Huang YM, Hung CM and CH (2006). Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications* **7** 720-747.

Olafsson S, Li X and Wu S (2008). Operations research and data mining. *European Journal of Operational Research* **187** 1429-1448.

Stewart GB (1991). *The Quest for Value: A Guide for Senior Managers* (Harper Business Publisher) New York.

Thomas LC, Edelman DB and Crook JN (2002). Credit scoring customers banking. *Expert Systems with Applications* 135-140.

Yueh MH and Chun MH (2006). Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications.*

Zhang Y and Bhattacharyya S (2004). Genetic programming in classifying large-scale data: an ensemble method. *Information Sciences* **163** 85.