Fuzzy-Genetic Classifier algorithm for bank's customers

Elawady R.M.[1], Asim S.A. [2], and Sweidan S.M.[3]

[1]Department of Communication, Faculty of engineering, Mansoura University, Mansoura, Egypt

[2]Department of information system, faculty of computers &information system, Mansoura University, Mansoura, Egypt

[3]Department of information system, faculty of computers &information system, Mansoura University,

Abstract

Modern finical banks are running in complex and dynamic environment which may bring high uncertainty and risk to them. So the ability to intelligently collect, mange, and analyze information about customers is a key source of competitive advantage for an E-business. But the data base for any bank is too large, complex and incomprehensible to determine if the customer risk or default. This paper presents a new algorithm for extracting accurate and comprehensible rules from database via fuzzy genetic classifier by two methodologies fuzzy system and genetic algorithms in one algorithm. Proposed evolved system exhibits two important characteristics; first, each rule is obtained through an efficient genetic rule extraction method which adapts the parameters of the fuzzy sets in the premise space and determines the required features of the rule, further improve the interpretability of the obtained model. Second, evolve the obtained rule base through genetic algorithm. The cooperation system increases the classification performance and reach to max classification ratio in the earlier generations.

Keywords: *fuzzy system, genetic algorithm, rule extraction, E-business*

1. Introduction

Bank customer classification plays an important role for commercial banks to keep away from default risks in customer loan market. A complete customer profile has two parts; factual and behavioral. The factual profile contains information such as name, gender, date of birth that personalization system obtained from the customer's factual data. The behavioral profile models the customer's actions and is usually derived from transactional data. Personalization begins with collecting customer data from various sources, web purchasing, and browsing activities [1]. After the data is collected, it must be stored in the data warehouse. Extracting rule from a given database for cluster customers data is important, there are several algorithms proposed by several researchers,[2,7] used computational intelligence, neural network, genetic algorithms, swarm intelligence (PSO), fuzzy system, rough sets for extracting accurate rules that solve the problem of classification customers with large and incomprehensible database.[5] employed BP neural network to classify customers into 5 groups according to the actual need, [6]use genetic algorithm to predict customer purchasing behavior.

In practical customer classification, there are two problems that can influence the accuracy. On the one hand, in credit scoring areas, we usually cannot label one customer as absolutely good who is sure to repay in time, or absolutely bad who will default certainly. On the other hand, there usually exist many irrelevant variables in the sample data. These redundant irrelevant variables spoil the classification, and increase many unwanted calculations and decrease the accuracy of customer classification. A good computerized classification tool should possess two characteristics, which are often in conflict. First, the tool must attain the highest possible performance, i.e. classify the presented cases correctly as being either *normal* or *risk*. Moreover, it would be highly desirable to be in possession of a so-called *degree of confidence*: the system not only provides a binary classification (*normal* or *risk*), but also outputs a numeric value that represents the degree to which the system is confident about its response. Second, it would be highly beneficial for such a classification system to be human-friendly, exhibiting so-called *interpretability*. The

proposed method combines two methodologies fuzzy systems and genetic algorithms which exhibit two important characteristics; first, each rule is obtained through an efficient genetic rule extraction method which adapts the parameters of the fuzzy sets in the premise space and determines the required features of the rule, further improve the interpretability of the obtained model. Second, evolve the obtained rule base through genetic algorithm by enabling the automatic production of fuzzy systems based on a database of training cases (fitness) for extract good rules. The cooperation among the fuzzy system and genetic algorithm increase the classification performance and reach to max classification ratio in the earlier generations.

2. Preliminaries

2.1 Fuzzy systems

Fuzzy logic is a computational paradigm that provides a mathematical tool for representing and manipulating information in a way that resembles human communication and reasoning processes. A fuzzy variables (also called a *linguistic variable*) is characterized by its name tag, a set of *fuzzy values* (also known as *linguistic values* or *labels*), and the membership functions of these labels; these latter assign a membership value, μ label (*u*) to a given real value *u*(*R*, within some predefined range known as the universe of discourse). While the traditional definitions of Boolean logic operations do not hold, new ones can be defined. Three basic operations, and, or, and not, are defined in fuzzy logic as following equations:

$$\mu A\,(u) and\, \mu B\,(u) = min\,(\mu A\,(u), \mu B\,(u)) \quad (1)$$

$$\mu A\,(u)\ or\ \mu B\,(u) = max\,(\mu A\,(u), \mu B\,(u)) \quad (2)$$

$$\mu not\ A(u) = \neg \mu A(u) = 1 - \mu A(u) \quad (3)$$

Where *A* and *B* are fuzzy variables. Using such fuzzy operator's one can combine fuzzy variables to form fuzzy-logic expressions, in a Boolean logic. For example, in the domain of control, where fuzzy logic has been applied extensively, one can find expressions such as: **if** room temperature **is** Low, **then** increase ventilation fan speed [8, 9, 10, and 15]. A *fuzzy inference system* is a rule-based system that uses fuzzy logic, rather than Boolean logic, to reason about data. Its basic structure includes four main components, as depicted in Figure ("1"): (1) a fuzzifier which translates crisp (real-valued) inputs into fuzzy values; (2) an inference engine that applies a fuzzy reasoning mechanism to obtain a fuzzy output; (3) a defuzzifier which translates this latter output into a crisp value; and (4) a knowledge base which contains both the rule base, and the initial database. The decision-making process is performed by the inference engine using the rules contained in the rule base. These fuzzy rules define the connection between input and output fuzzy variables [7].

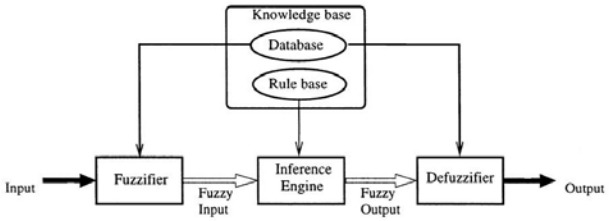

Fig.1 Basic structure of a fuzzy inference system

2.2 Genetic algorithm

GA is a combinatorial optimization technique based mechanics of the natural selection process (biological evolution) of a randomly chosen population of individuals can be thought of as a search through the space of possible chromosome values or search for an optimal solution to a given problem. The basic concept is that the strong tend to adapt and survive while the weak tend to die out. The evolutionary search process is influenced by the following main components of a GA [8]: an *encoding* of solutions to the problem as a chromosome or genome; a *function* to evaluate the *fitness*; *Initialization* of the initial population; *Selection* operators; and *Reproduction* operators. During each temporal increment, the structures in the current population are rated for their effectiveness as domain solutions (called a generation). GA has been successfully used in a wide variety of problem domains (Goldberg, 1989) [8, 11, 12, and 13].

2.3 Data mining

Data mining, "discovering hidden value in your data warehouse", is frequently described as *the process of extracting valid and actionable information from large, complex databases.* In other words, data mining derives patterns and trends that exist in data. These patterns and trends can be collected together and defined as a mining model. Mining models can be applied to specific business such as forecasting sales, determining specific customers and likely products to be sold [14, 16]. The evolved system trains the data base to classify customers in two groups *normal* and *risk*. According to some variables in specific period can define customer as normal or default. This classification helps bank to

remote the system for better customer relationship management.

3. Fuzzy genetic classifier model:

In the customer classification, there are two problems that can influence its accuracy. First, in credit scoring areas bank usually cannot label one customer as absolutely good who is sure to repay in time, or absolutely bad who will default certainly. Second, database is large, often complex and incomprehensible with many irrelevant variables. These redundant irrelevant variables spoil the classification, and increase many unwanted calculations and decrease the accuracy of customer classification. Moreover, it would be highly desirable to be in possession of a so-called *degree of confidence*: the system not only provides a binary classification (normal or risk), but also outputs a numeric value that represents the degree to which the system is confident about its response. Second, it would be highly beneficial for such a classification system to be human-friendly, exhibiting so-called *interpretability*. The proposed method combines two methodologies fuzzy systems and genetic algorithms, by using a simple GA with binary coding, to produce new generation of fuzzy rules based on a database of training cases (fitness) and formation of gene pool for extract good rules. The binary coding will represent all the membership functions associated to the linguistic labels belonging to each one of the linguistic term sets into a single chromosome, where are shifted along the x-axis freely. All the individuals in the population will represented by chromosomes with fixed same length. The individual (rule) will represented by the following form as Eq. (4):

$$If\ V_1\ is\ A_{i1}\ and\ V_2\ is\ A_{i2}\ \ and\ \dots\ \ and\ V_i\ is\ A_{ij}\ \ then\ Y\ is\ B_K \quad (4)$$

Where the linguistic labels A_i and B_K associated to the linguistic variables V_i and Y respectively, where $i = 1, \dots, n, k = 1, \dots, m, j = 0, \dots, 3$ where n is the number of the variables. As shown in figure ("2"), fuzzy variable u with two possible fuzzy values labeled *Low* and *High*, so the associated membership function are two points *(p, d)*, each point is represented by a binary number with a fixed number of bits determined by u_n which is the number of possible values for the variable V_i. P defines the start point to measure the degree of membership versus input values, and d defines the length of membership function edges which separates between two labels (*Low*, *High*).

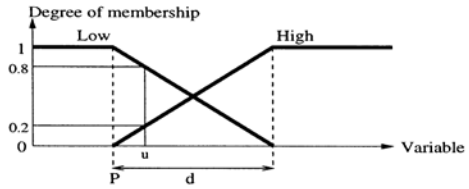

Fig. 2 example of a fuzzy variable

Each rule will be encoded in pieces of the chromosome $C_{ri}, i = 1, \dots, n$ where n is the number of the variables in the following way:

$$C_{ri} = (p_{i1}, d_{i1}, A_{i1}, \dots\dots, p_{in}, d_{in}, A_{in}) \quad (5)$$

The proposed model for classification problem consists of a fuzzy system and threshold system as shown in figure ("3"). The fuzzy system computes a continuous appraisal value of the risky customers, based on the input values. The threshold unit then outputs a normal or risk classifier according to the fuzzy system's output.

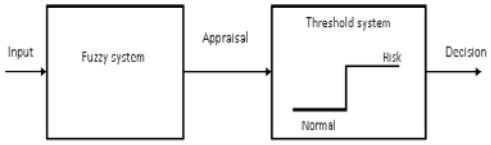

Fig.3 The proposed classifier system, note that the fuzzy subsystem displayed in Fig. 1

The proposed algorithm used to search for three parameters the relevant variables, the input membership function values, and the antecedents of rules. They are constructed as follows:

• Membership function parameters. There are *i* variables (V_i–V_n), each with two parameters *P* and *d*.

• Antecedents. The *i-th* rule has the form:

if *(V_1* ***is*** *A_{j1})* ***and****…****and*** *(V_8* ***is*** *A_{j8})* ***then*** *(output* ***is*** *normal)*

So A_{ij} represents the membership function applicable to variable V_i. A_{ij} takes the values: 1(*Low*), 2 (*High*), and 0 or 3 (*Other*).

• Relevant variables are searched for reduction by letting the algorithm chooses number of existent membership functions as valid antecedents; in such a case, the respective variables is considered irrelevant. For example, the rule

if (V_1 **is** High) **and** (V_2 **is** Other) **and** (V_3 **is** Other) **and** (V_4 **is** Low) **and** (V_5 **is** Other) **and** (V_6 **is** Other) **and** (V_7 **is** Other) **and** (V_8 **is** Low) **then** (output **is** normal),

is interpreted as:

if (V_1 is High) and (V_4 **is** Low) **and** (V_8 **is** Low) **then** (output **is** normal).

The parameters encoding are shown in figure ("4"), which form a single individual's genome. For each V_i have two points (p, d) with the same number of bits for each one and relevant variable (A_{ij}) where $i = 1, \dots, n$ and $j = 0, \dots, 3$. Each chromosome has a fixed length for all individuals can be calculated according to following equation:

$$\ell_{\mathbf{c}} = (\ |p|\ +\ |d|\ +\ |A|\ ) * n \qquad (6)$$

n is the number of variables which represented in the chromosome in bits.

P_1		d_1		A_1		.....		P_4		d_4		A_4		.....		n		.....	
1	0	1	0	0	1	1	1	0	0	1	1	1	0	0	0	1	0	1	1

Fig. 4 encoding of the chromosome

The structure of the chromosome generated randomly with initial database that used to tune the database parameters (*p*,*d*) and rule base constituted by *m* control rules.

3.1 Cluster explanation:

The proposed method starts with a set of solutions to the problem under examination, the solutions set (represented by chromosomes in GA) is called the population which generated randomly. Every evolutionary step, known as a generation, the individuals in the current population are decoded and evaluated according to some predefined quality criterion, referred to as the fitness, or fitness function.

Each individual generates a fuzzy rule. This rule is trained by the database values which presented as input values to the fuzzy system. The membership value of each variable is then computed as $\mu Low(u)$ and$\mu\, high(u)$, as shown in figure ("2"). Therefore, the inference engine goes on to compute the truth value of each rule by applying the fuzzy logic operator (i.e. complement, intersection, union) to combine the antecedent clauses (the membership values) in a fuzzy manner (1), this results in the output truth value, namely, a continuous value which represents the rule's degree of activation [15]. The defuzzifier producing the final continuous value of the fuzzy inference system; this latter value is the appraisal value that is passed on to the threshold unit, Figure ("3") which calculated as follow:

$$Appraisal = \frac{W(a) * \mu A(u) \vee \mu B(u) * B_{min} + w(d) * B_{max}}{W(a) * \mu A(u) \vee \mu B(u) + W(d)} \qquad (7)$$

where $W(a)$ and $W(d)$ are weights of the active rule and default rule, respectively. The continuous appraisal value would be in the range of (B_{min} ,B_{max}) discrete values of output membership function. This value is then passed along to the second subsystem *threshold subsystem* which produces the final binary output (B_{min} or B_{max}). The threshold subsystem simply the outputs B_{max} if the appraisal value is above a fixed threshold value and outputs B_{min} otherwise as illustrated in figure ("3"). The threshold value can be rewritten as follows:

$$\theta = (B\_min + B\_max)/2 \qquad (8)$$

Then the system computes the fitness function δ_{a^k}; means the percentage of cases correctly classified by the following equation:

$$\delta_{a^k} = \sum_{E \in cases}^{0} ((d_k(E) - a_k(E))^2) \qquad (9)$$

where the $k = 1, .., m$, m is the output decision for the rule and the $d_k(E)$ is the desired output for the case and $a_k(E)$ is the actual output of case estimated by the system. The proposed model uses genetic algorithm with a fixed population size of μ individuals to evolve the fuzzy inference system, and fitness-proportionate selection (higher fitness more likely) and genetic operators. To form a new population (the next generation), individuals are selected according to their fitness. Thus, high-fitness ('good') individuals stand a better chance of 'reproducing', while low-fitness ones are more likely to disappear. The evolved system puts these individuals in mating pool to generate good children by using genetic operators. Crossover operation is used to obtain a new individual by combining different chromosomes to generate new better child using crossover operator *pc*, the new solution carried out by flipping bits at random; with usually small probability *pm* [12]. The evolved system made it changeable to get optimized solution and small to ensure that the good solutions are not distorted too much. In each iteration, the evolved system runs the generating method for choosing the best chromosome. The algorithm terminates when the maximum number of generations is reached.

3.2 The fuzzy genetic classifier algorithm*:*

Input: training set, control parameters
Output: the optimized rule set
Begin:
1) Initialize control parameters;
2) t=0 , generation counter
3) Generate initial population randomly, $C(0)$, of μ individuals;
for each individual, X_i (t) ∈ C(t) **do**
Evaluate the fitness, f (X_i (t));
End
4) For each generation

a) Choose 2 parents at random;
b) Create offspring X'_i (t) through application of crossover operator on parent genome
c) Mutate offspring according to mutation operator;
d) Evaluate the fitness of offspring, f(X'_i(t))
e) If f(X'_i (t)) > f(X_i (t)) then
add X'_i (t) to c (t+1)
Else add X_i (t) to c (t+1),

5) Select the individuals to form rule set, which satisfy the higher fitness;
End

4. Results

According to the database which was collected from Barclays bank in the period from 30/12/2008 to 30/5/2009 as a six month used for training data the new model exhibits classification of the bank's customers in two groups normal and risk. The database consists of eight measured variables (V_1-V_8) as follow :(Average Revenue as V_1, Internal Transfer as V_2, Foreign Transfer as V_3 , Loan Count as V_4 , Loan Over Due as V_5 , Guarantees as V_6, Insurance as V_7, CC _ Over Due as V_8) as shown in table("1").

Table 1: customer classification

Customer id	V1	V2	…….	V8	Decision
100	1300	8	……	5	Normal
101	3000	6	……	10	Risk
……	……	…..	……	…..	……
970	450	6	……	3	Normal

The fuzzy system setup

Logical parameters
• Reasoning mechanism: singleton-type fuzzy system.
• Fuzzy operators: min.
• Input membership function type: orthogonal.
Structural parameters
• Number of input membership functions: two membership functions denoted *Low* and *High*.
• Number of output membership functions: two singletons are used, corresponding to the *normal* and *risk* classifier.
• Number of rules: The rule itself is to be found by the genetic algorithm.
Connective parameters
• Antecedents of rules: to be found by the algorithm.
• Consequent of rules: the algorithm finds rules for the *normal* class; the *risk* class is an else condition.
• Rule weights: active rules have a weight value 1 and the else condition has a weight of 0.25.
Operational parameters
• Input membership function values: to be found by the evolutionary algorithm.
• Output membership function values: we used a value of 3 for *normal* and 5 for *risk*.
• Threshold value: 4.

The evolutionary algorithm setup

• $\ell_{\mathbb{C}}$: length of the chromosome = 124 bit.
• *Selection method*: roulette wheel selection.
• *Population size*: 200.
• *Crossover operator*: 0.60.
• *Mutation operator*: 0.032.

The evolutionary performed into eight categories according to partitioning data into two distinct sets training set and test set as shown in table ("2"); The table lists the average performance over all 100 evolutionary runs, where the averaging is done over the best individual of each run. The performance value denotes the percentage of cases correctly classified. Three such performance values are shown, (1) performance over the training set; (2) performance over the test set; and (3) overall performance, considering the entire database. The choice of the training set is done randomly. The number of rules per system was fixed between one and six determined by the final structure of the genome.

Table 2: results summary of 100 evolutionary runs

Training/test cases	Performance training set	Test set	Overall
870/0	-	-	97.2
750/120	97.2	96.4	96.8
650/220	97.4	96	96.7
550/320	97.7	94.7	96.2
450/420	97.9	94.1	96
350/520	98	93.8	95.9
250/620	98.3	93.3	95.8
150/720	98.6	92.8	95.7

Figure ("6") shows the accuracy of the proposed model according to number of used cases in the training, the classification rate accuracy increases with the number of cases used after 100 evolutionary runs.

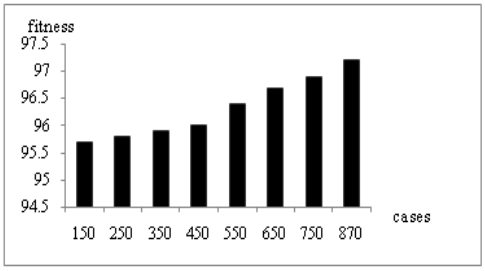

Fig. 6 classification rate

Finally, table ("3"): delineates the best one-rule system found through proposed evolutionary approach with its initial database and its rule base. It obtains 97.2% correct classification rate overall the customer cases.

Table 3: the best evolved fuzzy classification system with one rule

Data base								
	V1	V2	V3	V4	V5	V6	V7	V8
P	4773	9	2	2	1	39	7736	1
d	7257	18	6	2	5	95	6593	7
A	3	2	1	0	1	2	0	3
Rule base								
Rule	If (v2 is high) and (v3 is low) and (v5 is low) and (v6 is high)then (output is normal)							
default	Else (output is risk)							

[18] Uses fuzzy decision trees for classification of customers' problems by automatically creating fuzzy regions around tree nodes. [19] Proposes a fuzzy 'if-then' rule based classifier to predict bankruptcy in banks. After comparison with the proposed method on the data base of Barclays bank the following results achieved as shown in the figure ("7"). The proposed method (F.G) achieved higher classification accuracy on the data base than the other previous methods.

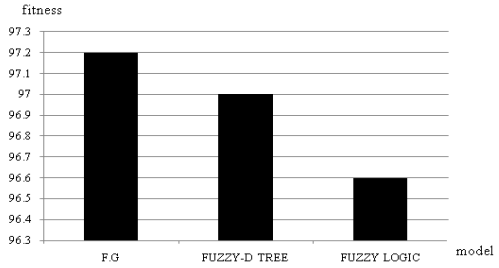

Fig. 7 comparison between classification rates

5. Conclusion

The paper explains the proposed algorithm (hybrid fuzzy- genetic algorithm) to solve the bank customers' classification problem. Proposed evolved fuzzy system presents both characteristics, first: attain high classification performance with possibility of measure to the output, second the results provide simple rules with its interpretable. Proposed model powerful in financial treatments because it could determine if the customer good or otherwise, with few determined measured variables. Fuzzy-natural computing hybrid techniques have been successfully applied to several fields of soft computing for intelligent data mining. For future work will add artificial immune systems as a natural computing technique with fuzzy logic for better classification.

Reference

[1] A. E. Elalfi, R. Haque, and M. E. Elalami, "Extracting rules from trained neural network using GA for managing E-business", Applied Soft Computing, Vol. 4, NO. 1, February 2004, pp. 65-77.

[2] F. Hoffmann, B. Baesens, and J. Martense, "Comparing a genetic fuzzy and a neurofuzzy classifier for credit scoring", international journal of intelligent systems, vol. 17, NO. 11, 2002, pp. 1067-1083.

[3] Z. Zhang, L. Zhang, and Sh. Niu, "A Parallel Classification Algorithm Based on Hybrid Genetic Algorithm", IEEE international conference on Intelligent control and automation, 2006, vol. 6, pp.3237-3240.

[4] X. Chuansheng, X. Xin, and H. Wentian, "Power Customer Credit Rating Based on FCM and the Differential Marketing Strategy Research", information science and engineering (ICISE), 2010, 2nd international conference,17 January 2011,vol. 2, pp.416-418.

[5] H. Wang, and Y. Xiang, "Study on Customer Classification Based on BP Neural Networks", IEEE international conference on Wireless Communications, Networking and Mobile computing, 2008, WiCOM'08, 4th international conference, 18 November 2008, vol. 4, pp 1.

[6] C. Chiu, "a case based customer classification approach for direct marketing", expert systems with applications, February 2002, vol.22, pp.163-168.

[7] G. Yang, and X. Yuan, "bank customer classification model based on Elman neural network optimized by PSO", IEEE international conference on Wireless Communications, Networking and Mobile computing, 2007, pp. 5672-5675.

[8] A. P. Engelbrecht, "Computational Intelligence: An Introduction", 2 edition, Wiley, England, 2007.

[9] F. Herrera, M. Lozano, and J. Verdegay, "Generation fuzzy rules from examples using genetic algorithms", fifth international conference of information processing and management of uncertainty in knowledge-based system, Paris, 1994, Vol. 5, pp 675-680.

[10] D. Fasel, "a fuzzy data warehouse approach for the customer performance measurement for a hearing instrument manufacturing company", IEEE international conference on Fuzzy Systems and Knowledge Discovery, 2009, Vol. 6, pp. 285-289.

[11] R. R. Yager, and D. P. Filev., "Essentials of Fuzzy Modeling and Control", SIGART Bulletin, Vol.6, NO. 4, 1994.

[12] O. Ahmed, M. Nordine, S. Sulaiman, and W. Fatimah, "Study of Genetic Algorithm to Fully-automate the Design and Training of Artificial Neural Network" , IJCSNS International Journal of computer Science and Network security, Vol.9 No.1, January 2009,pp. 217-226

[13] P. Makvandi, J. Jassbi, and S.Khan, "Application of Genetic Algorithm and Neural Network in Forecasting with Good Data", WSEAS International Conference on neural network, Lisbon, Portugal, 2005, Vol. 6, pp.56-61.

[14] J. Han, and M. Kamber, "Data mining: concepts and techniques", 2nd Edition, Morgan Kaufmann, 2006.

[15] C.A. Pena Reyes, and M.A. Sipper, "fuzzy-genetic approach to breast cancer diagnosis", Artificial Intelligence in Medicine, Vol. 17, No. 2, 1999, pp. 131-155.

[16] A. Berson, and Kurt Thearling, "Building Data Mining Application for CRM", USA, 1999.

[17] Z. Michalewicz, "Genetic Algorithms + Data Structures=Evolution Programs", 3rd edition. Berlin Heidelberg, Springer-Verlag, inc., 1996.

[18] K. Crockett, and Z. Bandar, "Fuzzification of Discrete Attributes From Financial Data in Fuzzy Classification Trees", IEEE International Conference on Fuzzy Systems, Korea, 2009, Vol. 18, pp 1320-1325.

[19] P.R. Kumar, and V. Ravi, "Bankruptcy Prediction in Banks by Fuzzy Rule Based Classifier", IEEE International Conference on Digital Information Management , 2006, Vol. 1,pp 222-227.