MINING FOR OPTIMISED DATA USING CLUSTERING ALONG WITH FUZZY ASSOCIATION RULES AND GENETIC ALGORITHMS

G.V.S.N.R.V.Prasad[1],Y. Dhanalakshmi[2], V.Vijaya Kumar[3], I. Ramesh Babu[4]
1. Department of Computer Science Engineering,Gudlavalleru Engg., College,Gudlavalleru,A.P.India
guttaprasad@yahoo.co.in
2. Dept of Information Technology,V.R.Siddhartha Engg.,College,Vijayawada,A.P.,India.
abbinenidhanalakshmi@yahoo.co.in
3.Dept of Computer Science Engineering & Information Technology,G.I.E.T,Rajamundary,A.P,India.
4.Department of Computer Science Engineering,Acharya Nagarjuna University,Guntur,A.P.,India.
rinampudi@hotmail.com

Abstract:
Data mining also known as knowledge discovery in databases has been recognized as a promising new area for database research. The proposed work in this paper is about optimizing the data with clustering and fuzzy association rules using multi-objective genetic algorithms. This algorithm is implemented in two phases. In the first phase it optimizes the data to reduce the number of comparisons using clustering. In the second phase it is implemented with multi-objective genetic algorithms to find the optimum number of fuzzy association rules using threshold value and fitness function.

Keywords: *Fuzzy sets, Genetic algorithms, clustering and association rules.*

1. Introduction:

Data Mining - the process of finding patterns from very large volumes of data has received enormous attention by the research community, for its significance [2]. In data mining important evaluation criteria are efficiency and comprehensibility of knowledge.

The discovered knowledge should as well describe the characteristics of the data besides facilitating for better understanding while leading the way to use it effectively [2].

A few researchers have focused on the use of fuzzy sets in discovering association rules [5,7], but the results achieved are not realistic and it is extremely a hard bitten process of specifying the fuzzy sets. This experience has made the researchers turn in using algorithms along with fuzzy association rules for better results [8]. Uncompromising researchers have started using multi-objective genetic algorithms for optimum results [12,16, and 18].

This paper in our opinion is a true reflection of the efficient research – offering best results for optimized data, where data mining rules along with fuzzy logic and multi-objective genetic algorithms use threshold value and fitness function.

The paper comprises nine sections where it briefly introduces the data mining rules, fuzzy quantitative association rules and multi-objective optimization in sections 2,3 and 4 and extensively discusses the multi-objective genetic algorithm in section 5. It emphasizes the fitness evaluation in section 6. followed by the problem description-proposed algorithm and flow chart in section 7, while section 8 leads to the summary arriving at conclusion, section 9 gives the list of references.

2. **Data Mining Rules**:

Now a day's most of the data available all over the world are stored in databases. Data Mining also known as knowledge discovery in databases has been recognized as a promising new area for database research. This area can be defined as efficiently discovering interesting rules from large databases [4].

Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters. By clustering, one can identify dense and sparse regions and therefore, discover overall distribution patterns and interesting correlations among data attributes. In the hierarchical methods of clustering CURE (Clustering Using REpresentatives) that adopts a middle ground between centroid based and representative object based approach. The process of CURE can be summarized as follows. Starting with individual values as individual clusters, at each step the closet pair of clusters are merged to form a new cluster. This is repeated until only k clusters are left. As a result the values of each attribute in the database are distributed into k clusters. The centroids of the k clusters are the set of midpoints of the fuzzy sets for the corresponding attribute.

Data Mining is the step in knowledge discovery process that attempts to discover novel and meaningful patterns in data. One important topic in data mining research is concerned with discovery of interesting association rules. An association rule is an implication of the form $X \Rightarrow Y$, where both X and Y are sets of attributes or items; it is interpreted as: "for a specified fraction of the existing transactions, a particular value of X determines the value of Y as another particular value under a certain confidence" [7]. Support and confidence are the major factors in measuring the significance of an association rule. Simply support is the percentage of transactions that both contain X and Y while confidence is the ratio of the support of $\mathcal{X} \cup \mathcal{Y}$ to support of X. So the problem can be stated as finding interesting association rules that satisfy user-specified minimum support and confidence.

3. **Fuzzy quantitative association rules:**

The theory of fuzzy sets has been recognized as a suitable tool to model several kinds of patterns that can hold in data. A general model was developed to discover association rules among items in a crisp set of fuzzy transactions [10]. Some work has recently been done on the use of fuzzy in discovering association rules for quantitative attributes [3,5,7].

Given a database of transactions T, its set of attributes I it is possible to define some fuzzy sets for attribute i_k with a membership function per fuzzy set such that each value

of attribute i_k qualifies to be in one or more of the fuzzy sets specified for i_k. The degree of membership of each value of i_k in any of the fuzzy sets specified for i_k is directly based on the evaluation of the membership function of the particular fuzzy set with the specified value of i_k as input. The following form of fuzzy association rules were used [16].

Definition:

A fuzzy association rule is expressed as:

If $Q = \{u_1, u_2, u_p\}$ is $F_1 = \{f_1, f_2, f_p\}$
then
$R = \{v_1, v_2, \ldots, v_q\}$ is $F_2 = \{g_1, g_2, \ldots, g_q\}$,

where Q and R are disjoint sets of attributes called item sets, i.e. $Q \subset I$, $R \subset I$ and $Q \cap R = \phi$; F_1 and F_2 contain the fuzzy sets associated with corresponding attributes in Q and R, respectively, i.e., f_i is the class of fuzzy sets related to attribute u_i and g_j is the class of fuzzy sets related to attribute v_j .Finally, "Q is F_1" is called the antecedent of the rule while "R is F_2" is called the consequent of the rule. For a rule to be interesting, it should have enough support and high confidence value, larger than user specified thresholds [16]. For generating the fuzzy association rules the following formula is used to calculate the fuzzy support of item set Z and its corresponding set of fuzzy sets F which is denoted by and $|T|$ denotes the number of transactions in database T

$$S_{<Z,F>} = \frac{\sum_{t_i} \in T \prod_{Z_j \in Z} \mu_{z_j(} (f_j \in Ft_i[z_j]}{|T|}$$

4. Multi-objective optimization:

Contrary to single objective optimization problem, multi-objective optimization problem deals with simultaneous optimization of several incommensurable and often competing objectives such as performance and cost. For example, when the design of a complex hardware is considered, it is required that the cost of such systems be minimized while the performance is maximized. If there is more than one objective criterion as in the example mentioned above, some of them can be considered as constraints in the problem. For example, while trying to optimize a system for large performance at low cost, the size of the system must not exceed the given dimensions as a separate optimization criterion. By this way, a multi-objective optimization problem can be formalized as follows.

A multi-objective optimization problem includes, in general, a set of a parameters (called decision variables), a set of b objective functions, and a set of c constraints; objective functions and constraints are functions of the decision variables.

The optimization problem is modeled as:

min / max $y = f(x) = (f_1(x), f_2(x), ..., f_b(x))$

constraints $e(x) = (e_1(x), e_2(x), ..., e_c(x)) \geq 0$

with $x = (x_1, x_2, ..., x_a). \in X$
$y = (y_1, y_2, ..., y_b) \in Y$
where x is the decision vector, y is the objective vector, X denotes the decision space, and Y is called the objective space; the constraints $e(x) \geq 0$
determine the set of feasible solutions [16].

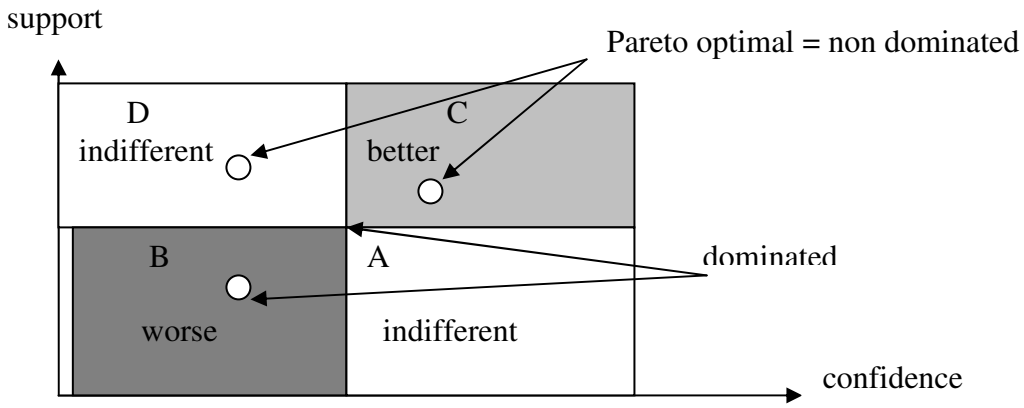


Fig.1.1 The concept of pareto optimality

Consider the above definition and assume that the two objectives performance (f_1) and cheapness (f_2), the inverse of cost, are to be maximized under size constraints (e_1). Then, an optimal design might be an architecture, which achieves maximum performance at minimal cost and does not violate the size of limitations. If such a solution exists, then actually we have to solve only a single-objective optimization problem. The optimal solution for either objective is also the optimal for the other objective. However, what makes multi-objective optimization problems difficult is the common situation when individual optima corresponding to the distinct objective functions are sufficiently different. In Fig.1.1, consider the values of support and confidence of fuzzy rules as objective functions. In this regard, a solution defined by corresponding decision vector can be better than, worse, or equal to, but also indifferent from another solution with respect to the objective values as shown in Fig.1.1. Better means a solution is not worse in any objective and at least better in one objective than another. For example, while the solution represented by point B is worse than the solution represented by point A, the solution with C is better than that of A. However, it cannot be said that C is better than D or vice versa. This is because one objective value of each point is higher than the other one. Using this concept, an optimal solution can be defined as a solution which is not dominated by any other solution in the search space. Such a solution is called Pareto optimal, and the entire set of optimal trade-offs is called the Pareto-optimal set, which is represented in Fig.1.1. In such an optimization problem, the objectives are conflicting and cannot be optimized simultaneously. Instead a satisfactory trade-off has to be found. Therefore, it is necessary to have a decision making process in which preference

information is used in selecting an appropriate trade-off. In the literature, strong association rules have been defined as the rules having the values of support and confidence above certain threshold values. If the attributes in a rule include quantitative values, another important measure for strong association rules becomes to find the number of appropriate fuzzy sets and their membership functions. In this case, if the interval of a membership function is shrunk, the value of support of the rule concerning that membership function decreases. However, it can be said nothing about the value of confidence of that rule.

5. Multi-objective genetic algorithms:

Genetic Algorithms are heuristic optimization methods whose mechanisms are analogous to biological evolution. A good general introduction to genetic algorithms is given in [1]. In Genetic algorithms the solutions are called individuals or chromosomes. After initial population is generated randomly, selection and variation functions are executed in a loop until some termination criteria is reached. Each run of the loop is called generation. The selection operator is intended to improve the average quality of the population by giving individuals of higher quality a higher probability to be copied into next generation. The quality is measured by fitness function.

Chromosome encoding has two different encoding schemes. The first tries to find the appropriate fuzzy sets in a certain rule such that the desired criterion in the previous section, whether a rule is interesting or not can be judged either subjectively or objectively. Ultimately, only the user can judge whether a given rule is interesting or not. Furthermore, this judgment may differ from one user to another. However, Objective Interestingness Criterion can be used as one step towards the goal of pruning uninteresting rules from presentation to the user.

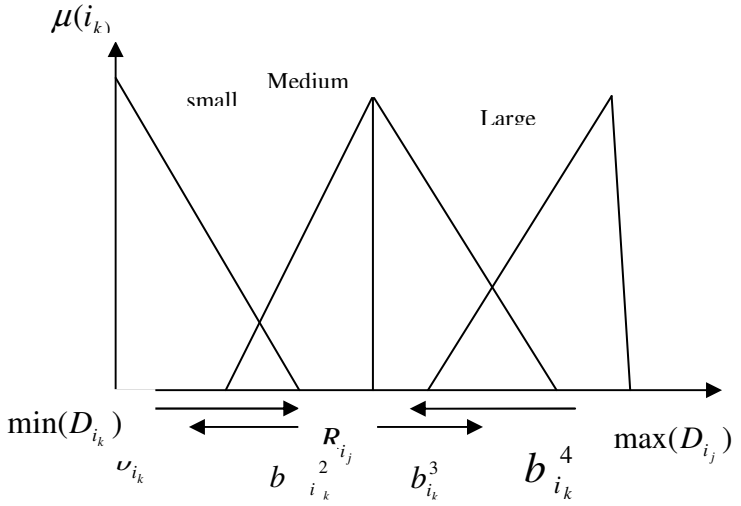


Fig.1.2 Membership functions and base variables of attribute i_k

$b^1_{i_k}$: [min(D_{i_k}), max(D_{i_k})]

R_{i_k} : [min(D_{i_k}), max(D_{i_k})]

$b^2_{i_k}$:[min(D_{i_k}),R_{i_k}]

$b^3_{i_k}$:[R_{i_k} ,max(D_{i_k})]

$b^4_{i_k}$:[min(D_{i_k}),max(D_{i_k})]

Chromosome encoding:

To illustrate the encoding scheme used, membership functions for a quantitative attribute i_k having three fuzzy sets and their base variables are shown in Fig.1.2.Each base variable takes finite values. For instance, the search space of the base value b^1_{ik} lies between the minimum and maximum values of attribute i_k, denoted min((D_{i_k})) and max(D_{i_k}), respectively. So, based on the assumption of having three fuzzy sets per attribute, as is the case with attribute i_k, a chromosome that consists of the base lengths and the intersection point is represented in the following form:

$$b^1_{i_1} b^2_{i_1} R_{i_1} b^3_{i_1} b^4_{i_1} b^1_{i_2} b^2_{i_2} R_{i_2} b^3_{i_2} b^4_{i_2} .....b^1_{i_m} b^2_{i_m} R_{i_m} b^3_{i_m} b^4_{i_m}$$

To illustrate the process, consider 5 quantitative attributes. It is assumed that each attribute can have at the most 5 fuzzy sets. So, a chromosome that consists of the base lengths and the intersecting points is represented in the following form.

$$w_{i_1} b^1_{i_1} b^{12}_{i_1} R^1_{i_1} b^2_{i_2} b^3_{i_1} R^2_{i_1} b^4_{i_1} b^5_{i_1} R^3_{i_1} b^6_{i_1} b^7_{i_1} R^4_{i_1} b^8_{i_1} b^9_{i_1} R^5_{i_1} b^{10}_{i_1} b^{11}_{i_1} .....w_{i_5} b^1_{i_5} b^{12}_{i_5} .....R^5_{i_5} b^{10}_{i_5} b^{11}_{i_5}$$

where, the gene w_{i_j} denotes the number of fuzzy sets for attributes i_j . If the number of fuzzy set equals 2, then while decoding the individual, the first two base variables are considered and the others are omitted. However, if w_{i_j} indicates to 3, then the next three variables are taken into account more. So, as long as the number of fuzzy set increases, the number of variables to be taken into account is also enhanced. The chromosomes are represented as a floating point number and their genes are real parameters when using a real-valued coding. The value of gene is reflected under its own search interval by the following formula

$$b^k_{i_j} = \min(b^k_{i_j}) + \frac{g}{g_{\max}} (\max(b^k_{i_j}) - \min(b^k_{i_j}))$$

Where g is the value of gene in search

$g_{\max}$ is the maximum value of the gene g may take

$\min b^k_{i_j}$ Are the minimum value of the reflected area?

$\max b^k_{i_j}$ are the maximum values of the reflected area

. In multi-objective problems both fitness assignment and selection must allow for several objectives. One of the methods used for fitness assignments is to make direct use of the concepts of Pareto dominance [17].

6. Fitness evaluation:

In a given population the fitness function is measured as the goodness of an individual and also the success of a genetic algorithm is to optimize the fitness function. This fitness function should be carefully set, by taking into considerations all the factors that play an important role in optimizing the problem under investigation. The new population is generated in the process is evaluated with respect to the fitness function. The evaluation process is the main source for providing the mechanism for evaluating the status of each chromosome, and is also the main criteria for linking the genetic algorithms and the system. The decoded chromosome which accepts the fitness function produces an objective value as a measure of performance of the input chromosome. The aim of the genetic algorithms employed in this study is to maximize the large item sets and minimum support values in a given interval. The fitness function of genetic algorithms is calculated as follows

$$\text{Fitness} = \sum_{i=2}^{n} (\min\_\sup[i] - \min\_\sup[i-1] \times l \arg e - itemsets[i-1]$$

Where n is number of iterations in a given interval of minimum support values

7.. Problem Description:

They are many algorithms proposed in [18]. However a trial is taken to improve the algorithm using Genetic Algorithms. Normal Distribution is taken as the situation arises out a more natural phenomenon. A fuzzy threshold value ϵ [0, 1] is used. If the fuzzy value of the items to be inserted at any stage is greater than equal to ($\geq$) threshold then only it is accepted otherwise it is discarded.

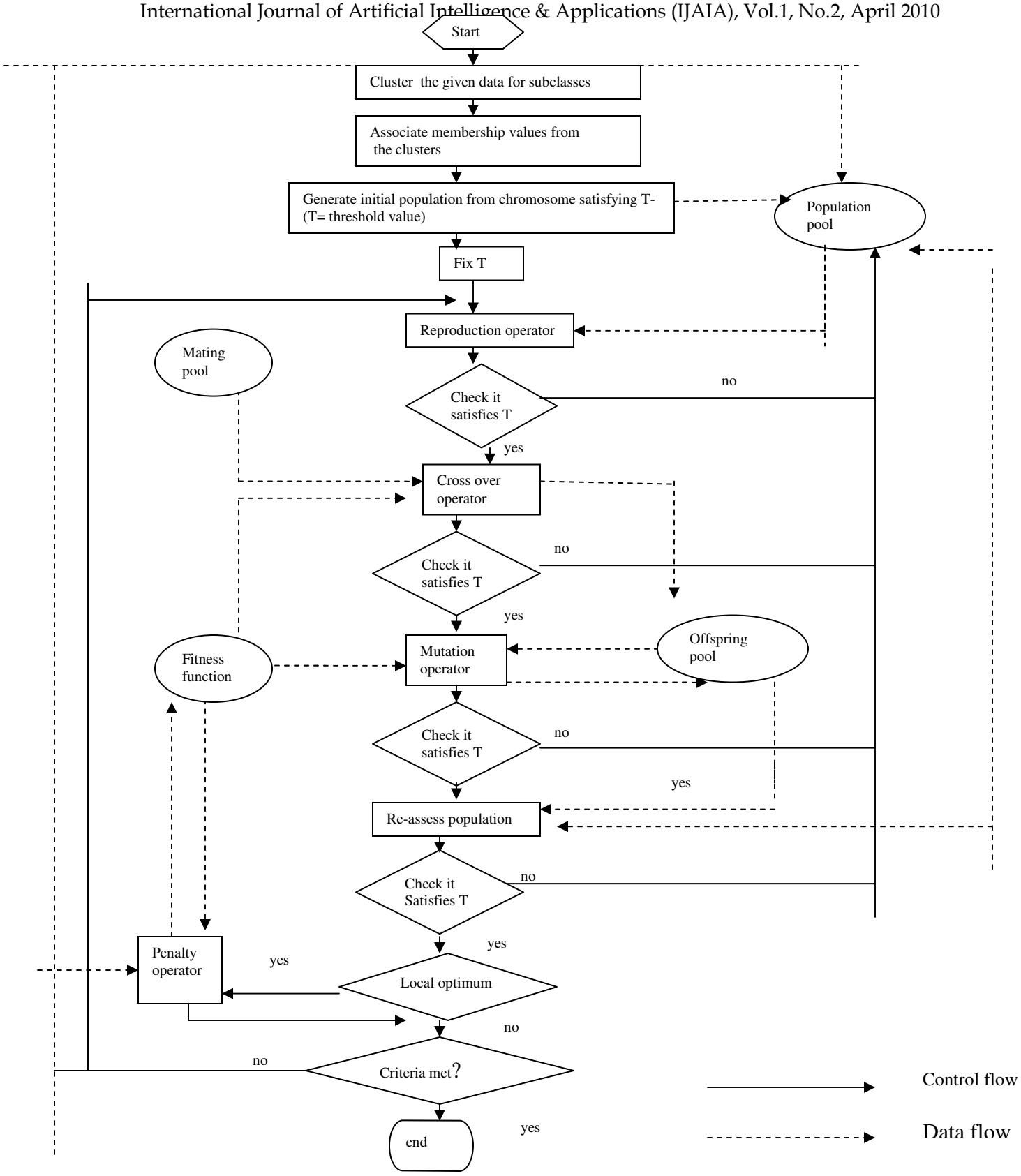


Fig.1.3 Flow chart showing proposed algorithm

The proposed algorithm:

Input: Population size: N
Maximum number of generations: G
Crossover probability: p_c
Mutation rate : p_m
Threshold value: Th
Output: Nondominated set: S
1 Set $P_0 = \Phi$ and t = 0, For h = 1,to N do
a) Choose i∈ I , where i is an individual and I is the individual space, according to normal distribution,
b) Set $P_0 = P_0 + \{i\}$
2 a)Randomly for each individual i and P_t ,
b) Determine the encoded decision vector and objective vector and calculate the scalar fitness value F(i) with respect to the approach mentioned above.
3 Set P' =Φ: For h= 1 to N do
a) Select one individual i∈ P_t with respect to its fitness value F(i)
b) Set P' = P' + {i} if the fuzzy threshold value of i ≥ Th.
4 Set P" =Φ For h = 1 to N/2 do
a) Choose two individuals i, j∈ . P' and remove them from P'
b) Recombine i and j using crossover The resulting offsprings are k, l ∈ I
c) Insert k, l into P" with probability p_c, otherwise insert i, j into P" if the threshold value of i, j ≥Th
5 Set P"' =Φ, For each individual i∈ P" do
a) Mutate i with mutation rate p_m. The resulting individual is j∈ I
b) Set P"' = P"' + {j } if the threshold of j ≥ Th.
6. Set P_{t+1} = P" and t = t +1.
If t = G or another termination criterion is satisfied then return S = p(P_t), where the p(P_t) gives the set of non-dominated decision vectors in P_t . In other words, the set p(P_t) is the non-dominated set regarding P_t .
Otherwise go to step 2, i.e, execute steps 2 to 6.

Fig 1.4 The proposed algorithm :

7. Summary and Conclusions:

The data mining process from numerical attributes is more difficult for analysts than itemized data in general. In this paper we propose multi-objective genetic algorithms for optimizing data mining rules based on the fitness function and threshold value. These algorithms with fuzzy sets are used to propose a new algorithm for optimizing the data through clustering and association rules. The main purpose of the second stage is to

reduce the size of data objects to a moderate one so as to be suitable for genetic algorithms in second stage. So it is possible to obtain more appropriate solutions, with a number of large item sets and interesting association rules. At every stage of iterations of genetic algorithms, number of items is reduced. As a result of all these advantages it shows that the proposed approach is more appropriate and can be used more effectively in achieving an optimal solution than the classical methods.

9. References:

[1] G.E. Goldberg, "Genetic Algorithms in Search optimization and Machine learning, " Addison Wesley, New York, 1989.

[2] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in Proc. 20th VLDB Conf. Sept 1994, pp.478-499

[3] R.R. Yager "Fuzzy Summaries in Database Mining" Proc. Of Artificial Intelligence for Applications pp. 265-269, 1995.

[4] R. Srikant and R. Agrawal " Mining quantitative Association rules in large relational tables " in Proc. ACM SIGMOD Int. Conf. Management Data 1996, 1-12.

[5] K.C.C.Chan and W.H. Au " Mining Fuzzy Association Rules" Proc. Of ACM CIKM pp.209-215,1997.

[6] Hisao Ishibuchi, Tadahiko Murata "A multi objective Genetic Local Search Algorithm and its application to Flow shop Scheduling IEEE Trans. System Man and Cybernetics, Part C, Vol.28. No.5, Aug 1998, pp.392-403.

[7] W. Zhang "Mining Fuzzy quantitative Association rules", Proc. Of IEEE, ICTAI Pp.99-102, 1999.

[8] Been-Chien,ZinLongLin,Tzung-PeiHong "An efficient clustering Algorithm for mining fuzzy quantitative association rules", IEEE,2001,pp.1306-1311,

[9] Plamen Angelov, Richard Buswell "Identification of evolving fuzzy rule-based Models" IEEE Transactions on Fuzzy Systems Vol.10, No.5, Oct 2002, pp.667-677.

[10] Miguel Delgado, Nicolas Marin, Daneil Sanchez and Maria-Amparo-Vila" Fuzzy Association Rules: General model and Applications", IEEE Transactions on Fuzzy systems, Vol. 11. No.2, April 2003.

[11] Hisao Ishibuchi and Takashi Yamantoo "Fuzzy rule selection by multi-objective Genetic local search algorithms and rule evaluation measures in data mining" Fuzzy sets and systems, 141(1); 2004,pp.59-88.

[12] Mehmet Kaya, Reda Alhaji "A clustering Algorithm with Genetically Optimized membership functions for fuzzy association rules", Proc. The IEEE International conference on Fuzzy Systems, IEEE,2003.

[13] Mehmet Kaya, Reda Alhaji " Facilitating Fuzzy Association Rules Mining by using multi-objective Genetic algorithms for automated clustering",Proc. of The Third International Conference on Data Mining ICDM,IEEE,Melbourne USA, 2003.

[14] S.S.Weng, Y.J.Lin and F.Jen,"A study on searching for similar documents based on multiple concepts and distribution of concepts", Expert system with applications 25,2003,pp.355-368.

[15] Hisao Ishibuchi and Takashi Yamantoo " Fuzzy rule selection by multi-objective Genetic local search algorithms and rule evaluation measures in Data Mining",Fuzzy sets and Systems, 141(1): 2004, pp. 59-88.

[16] Mehmet Kaya, Reda Alhaji "Multi-Objective Genetic Algorithm Based Method for Mining optimized Fuzzy Association Rules", 5th International Conference on Intelligent Data Engineering and Automated Learning IDEA Exeter,UK, 2004, pp 758-764.

[17] Mehmet Kaya "Multi-objective genetic algorithm based approaches for mining optimized fuzzy association rules",Soft Computing, 2006, pp.578-586.

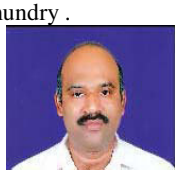
[18]S.M Khalessizadeh., R.Zaefarian, S.H.Nasseri, and E.Ardil "Genetic Mining:Using Genetic Algorithm for Topic based on Concept Distribution "Transactions on Engineering Computing and Technology,Vol.13, World Enformatika Society,May 2006,pp.144-147.

Authors:

Prof. G.V.S.N.R.V.Prasad did his MS Software Engineering in BITS Pilani and M.Tech in Computer Science and Technology in Andhra University .He has 15 years of teaching experience. Published 7 Research Papers in various National and International Conferences He is a member in various Professional Bodies . Presently working as Professor and Head in CSE at Gudlavalleru Engineering College , Gudlavalleru ,A.P. His area of interest is Data Mining, Network Security and Image Processing

Dr.Y.Dhanalakshmi did her MCA,M.Phil,and did her Ph.D in Computer Science & Engg., from Acharya Nagarjuna University. She is presently working in V.R.Siddharatha Engg.,College She has published 9 papers in international and national journals. Her area of interest is Data Mining, Network security.

Prof..Vijaya Kumar did his MS Engineering in Computer Science
[USSR –TASHKENT STATE UNIVERSITY] and Ph.D in Computer Science . Worked as Associate Professor in Department of CSE and School of Information Technology (SIT) at Jawaharlal Nehru Technological university (JNTU) Hyderabad . Having a total of 13 years of experience. He Published 60 Research Papers in various National and International Conferences /Journals. Guiding 10 Research scholars . He is a Member for various National and Inter National Professional Bodies .Presently working as Dean for CSE & IT at
ODAVARI INSTITUTE OF ENGINEERING AND TECHNOLOGY Rajamundry .



Prof. I.Ramesh Babu .did his M.Tech in Andhra University and Ph.D in Computer Science & Engg., from Nagarjuna University. He joined as an Assistant Professor in the Department of Computer Science and Engineering in Acharya Nagarjuna University in1988,and became a Professor in 2004.He held many positions in Acharya Nagarjuna University as Executive council member, Dean of Engineering, Chairman Board of Studies, Head, Director Computer Centre, Member of academic senate, member of the standing committee of academic senate. He is also a member of Board of Studies for other universities. He has published many research papers in International, national journals and presented papers in international conferences also. His research areas of interest include Image Processing, Computer Graphics, Cryptography, Network Security and Data Mining. He is member of IEEE, CSI, ISTE,IETE, IGISS, Amateur Ham Radio (VU2UZ)