Pattern Anal Applic (2008) 11:159–168
DOI 10.1007/s10044-007-0090-x

THEORETICAL ADVANCES

Extraction and optimization of fuzzy association rules using multi-objective genetic algorithm

P. Santhi Thilagam · V. S. Ananthanarayana

Received: 30 January 2006 / Accepted: 7 August 2007 / Published online: 9 October 2007

Abstract Association Rule Mining is one of the important data mining activities and has received substantial attention in the literature. Association rule mining is a computationally and *I/O* intensive task. In this paper, we propose a solution approach for mining optimized fuzzy association rules of different orders. We also propose an approach to define membership functions for all the continuous attributes in a database by using clustering techniques. Although single objective genetic algorithms are used extensively, they degenerate the solution. In our approach, extraction and optimization of fuzzy association rules are done together using multi-objective genetic algorithm by considering the objectives such as fuzzy support, fuzzy confidence and rule length. The effectiveness of the proposed approach is tested using computer activity dataset to analyze the performance of a multi processor system and network audit data to detect anomaly based intrusions. Experiments show that the proposed method is efficient in many scenarios.

Keywords Fuzzy association rules · Multi-objective genetic algorithms · Fuzzy *k*-means clustering

P. Santhi Thilagam (✉)
Department of Computer Engineering,
National Institute of Technology Karnataka,
Srinivasanagar, Surathkal 575025, India
e-mail: santhi_soci@yahoo.co.in

V. S. Ananthanarayana
Department of Information Technology,
National Institute of Technology Karnataka,
Srinivasanagar, Surathkal 575025, India
e-mail: anvs@nitk.ac.in

1 Introduction

The discovery of association rules in large databases has been the focus of a number of research papers since its introduction [1]. The structure of an association rule is in the form of a relationship between valuesets of attributes ($A \Rightarrow B$) in a database in the early proposals. Each rule has two measurements, support and confidence. Confidence is a measure of the rule's strength, while support corresponds to statistical significance.

However, an active research in this area over the recent years has lead to a variety of structures to characterize association rules. Tables in most business and scientific domains have richer attribute type. Attribute can be quantitative (e.g. age, income) or category-based (e.g. zip code, make of car). The problem is to identify association rules between quantitative and category-based data in relational tables. A quantitative association rule is of the form (Age: 30:39)^(Married: yes) $\Rightarrow$ (No. of cars: 2). Based on variations in support value in the formation of the association rules, we have Crisp association rules and fuzzy association rules [2]. Fuzzy association rules are the fine-tuned versions of quantitative association rules where fuzzy sets are used.

Existing algorithms (e.g. [1, 3, 4]) involve discretizing the domains of quantitative attributes into intervals so as to discover quantitative association rules. But, these intervals may not be concise and meaningful enough for human experts to easily obtain nontrivial knowledge [5]. On the other hand, we can use fuzzy association rules of the form: "if X is A then Y is B: (X is $A \Rightarrow Y$ is B)" which provides a smooth boundary, where each attribute x_k in X will have a fuzzy set f_{xk} in A such that f_{xk} belongs to F_{xk}, a set of all fuzzy sets and is similar for attribute Y [6, 7]. Linguistic

representation [8] makes the discovered rules to be much natural for human experts to understand.

Evolutionary Algorithms (EA) or Genetic Algorithms (GA) [9, 10] are randomized search procedures often used as optimization algorithms in most data mining applications. In this paper, the process of extracting optimized fuzzy association rules is modeled as a multi-objective optimization problem. Extraction and optimization of fuzzy association rules of different orders are done together using Multi-Objective Genetic Algorithm (MOGA). We applied this technique to computer activity dataset which describes about the performance of a multi processor system. Intrusion detection has become an important area of research since it is not technically feasible to build a system with no vulnerabilities [11]. Rule based expert systems, machine learning methods [12], time-based inductive machine [13], neural network-based systems [14] exist in the literature for solving this problem. More recently, techniques from the data mining area have been used to mine normal patterns from network audit data [15]. Intrusions in the network can be found by calculating the deviation of current network traffic data with the extracted patterns. In this paper, the above problem is solved by integrating fuzzy logic with data mining methods for intrusion detection [16]. Here, we defined two measures Intrusion Chance factor (ICF) and Average Confidence and Support Product (ACSP) for calculating the deviation. Our main contributions include:

- Identification of fuzzy attributes and defining the membership functions based on clustering techniques.
- Generation and Optimization of fuzzy association rules of different orders using multi-objective genetic algorithm.
- Experimental study on computer activity dataset and network audit data.

2 Related work

Evolution of various techniques for Fuzzy Association Rule Mining (FARM) is depicted in Fig. 1. Chan and Au introduced a novel technique, called F-APACS, for mining fuzzy association rules [7]. Existing algorithms involve discretizing the domains of quantitative attributes into intervals so as to discover quantitative association rules. Instead of using intervals, F-APACS employs linguistic terms to represent the revealed regularities and exceptions [6].

F-APACS employs adjusted difference analysis which has the advantage that it does not require any user-supplied thresholds which are often hard to determine. F-APACS effectively discovers rules from a real-life transactional

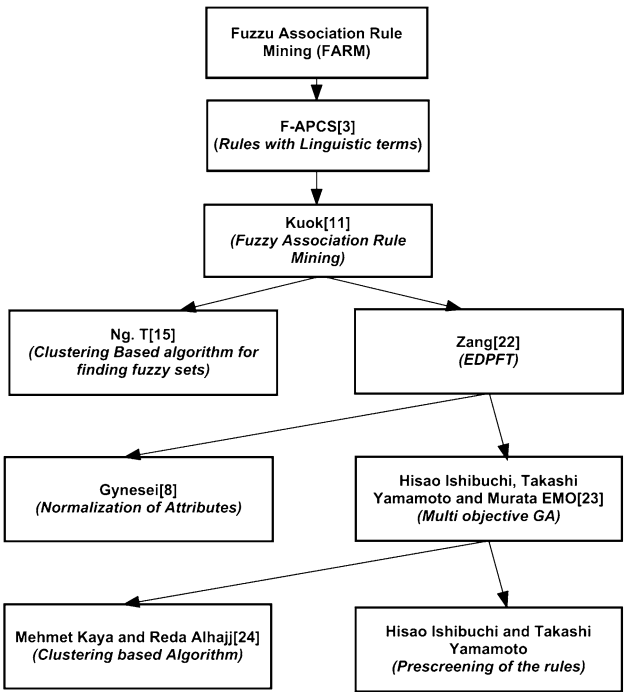

Fig. 1 Evolution of the different approaches

database of a PBX system provided by a telecommunication corporation.

Kuok' s algorithm [17] expects the user or an expert to provide the required fuzzy sets of the quantitative attributes and their corresponding membership functions. Fu argues that experts may not give the right fuzzy sets and their corresponding membership functions. Hence, he proposed a method to find the fuzzy sets based on clustering techniques [3].

Zhang' s EDPFT algorithm [18] discovers association rules that have raw domain values, intervals and fuzzy terms in both the antecedent and the consequent. Gyenesei [5] assigns each attribute with several fuzzy sets that characterize the quantitative attribute. He defines two different interest measures: fuzzy confidence and fuzzy correlation and two different methods of mining fuzzy quantitative association rules: with and without normalization.

Gyenesei [2] introduces the problem of mining weighted quantitative association rules based on fuzzy approach. He assigns weights to the fuzzy sets to reflect their importance to the user and proposes two different definitions of weighted support: with and without normalization similar to his previous method.

Ishibuchi et al. [19] demonstrated the effect of a three-objective formulation of fuzzy rule selection on the generalization ability of obtained rule sets. They have shown that many non-dominated solutions can be simultaneously obtained by the single run of Evolutionary Multi-Objective

optimization (EMO) algorithms which is the advantage over classical approaches and hence they have utilized this in genetic rule selection for the design of fuzzy rule-based classification systems [20].

Ishibuchi et al. extended the genetic algorithm-based rule selection method in Ref. [19] to the case where various fuzzy partitions with different granularities are used for each input. This extension increases the number of candidate rules. Hence, they proposed a prescreening procedure which is based on two rule evaluation criteria of association rules, for decreasing the number of candidate rules.

Kaya et al. [21] proposed an automated clustering method based on multi-objective genetic algorithms. This method automatically clusters the values of a given quantitative attribute in order to obtain large number of large itemsets in low duration. They compared their proposed approach with CURE-based approach. In addition to the autonomous specification of fuzzy sets, experimental results exhibit good performance over CURE-based approach in terms of runtime as well as the number of large itemsets and interesting association rules.

Bridges et al. [16] developed a prototype Intelligent Intrusion Detection System (IIDS) to demonstrate the effectiveness of data mining techniques that utilize fuzzy logic and genetic algorithms.

Au et al. developed a fuzzy technique, called Fuzzy Association Rule Mining II (FARM II) for mining very large Bank database [8]. They discovered some interesting characteristics about the customers who had once used the bank' s loan services but then decided later to cease using them. The bank translated what they discovered into actionable items by offering some incentives to retain their existing customers.

3 Statement of the research problem

Input: A dataset consisting of finite number records with quantitative and categorical attributes.

Steps: Consider a database consisting of a set of attributes $A = \{a_1, a_2, a_3, \ldots, a_n\}$, where a_i can be quantitative or categorical. For any quantitative attribute a_i that belongs to A can be represented using a linguistic term set $L = \{l_1, l_2, \ldots, l_k\}$. Each linguistic term is characterized by a fuzzy set F_{ij} (jth fuzzy set for ith attribute) defined on the domain of a_i and whose membership function can be given by

$$f_{ij} : \text{Domain}(a_i) \rightarrow [0, 1]$$

Categorical attributes are treated normally. The problem is to find the set of association rules in terms of the linguistic terms which are nothing but the defined fuzzy sets on the attributes. Genetic algorithm is used to generate and optimize the rules. Optimization of the association rules is treated here as the multi-objective optimization problem. The objective functions considered for the fuzzy association rule mining problem are maximizing support and confidence. Multi-objective optimization problem considers a set of 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 goal is expressed as:

$$\begin{aligned} \text{Min/Max } y = f(x) &= \{f_1(x), f_2(x), f_3(x), \ldots, f_b(x)\} \\ \text{where } x &= (x_1, x_2, x_3, \ldots, x_a) \text{belongs to } X \\ y &= (y_1, y_2, y_3, \ldots, y_b) \text{belongs to } Y \end{aligned}$$

where x is decision vector, y is objective vector, X denotes decision space and Y is called objective space.

Output: Set of optimized fuzzy association rules of different orders, which are in terms of linguistic terms, both in antecedent and consequent parts.

4 Mining fuzzy association rules

4.1 Design

The proposed solution approach for finding the association rules in terms of fuzzy terms is shown in Fig. 2. This approach consists of five phases.

4.1.1 Database selection and cleaning

The first phase is selection of an appropriate dataset for the rule extraction. Cleaning of the dataset involves converting the attributes to the required form. For example, the format of time attribute (HH:MM:SS) is not suitable for defining fuzzy sets. So it has to be converted into seconds or any other suitable format.

4.1.2 Preprocessing

Preprocessing step includes selecting the attributes, finding the continuous attributes, defining the attribute hierarchies for the selected attributes. Continuous attributes can be partitioned into groups of certain range [18]. But this leads to the problem of sharp boundaries. Though this problem can be solved by using fuzzy set theory, it is difficult to define the membership functions for each and every attribute based on intuition [17]. Hence, a clustering based approach [3] has been used for finding membership for each attribute value.

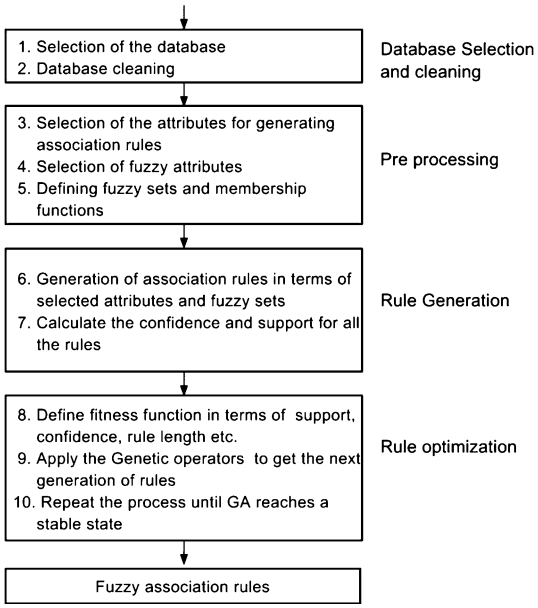

Fig. 2 Steps for extracting fuzzy association rules

4.1.3 *Rule generation and Rule optimization*

This phase deals with the generation and optimization of the rules which possess high support and high confidence using multi-objective genetic algorithm [22].

4.1.4 *Fuzzy association rules*

After the optimization phase, all the extracted rules are represented using linguistic terms and the attribute names. Output of this procedure is a set of fuzzy association rules of different orders.

The data flow diagram for mining fuzzy association rules is shown in Fig. 3. Initially all the quantitative attributes of the dataset are given as input for the fuzzy *k*-means clustering module for calculating the membership functions. Calculated function values are given to the rule generation module for obtaining initial rules. Then, the generated initial rules and membership function values are given to the optimization module for generating optimized rules. These rules are added to the main set of rules. Output of all these phases is set of fuzzy association rules.

4.2 Solution

4.2.1 *Linguistic term*

Consider a dataset R which consists of attributes $A = \{A_1, A_2, A_3,\ldots, A_n\}$, where each A_i can be continuous or

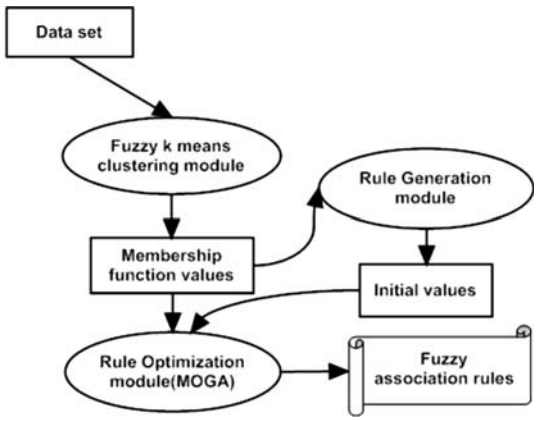

Fig. 3 Data flow diagram for fuzzy association rule mining

categorical. Let $L = \{l_1, l_2, l_3,\ldots, l_m\}$ be a set of linguistic terms defined over a continuous attribute a_i. Any continuous attribute a_i is represented by a linguistic variable L_i, whose values is a linguistic term in $T(L_i) = \{l_{ij} / i = 0, 1, 2, \ldots, s_i\}$ where l_{ij} is linguistic term characterized by a fuzzy set F_{ij}, that is defined on domain of A_i.

$$U_{F_{ij}} = \text{Domain}(A_i) \rightarrow [0, 1]$$

The fuzzy sets F_{ij}, $i = 1, 2, \ldots, s_i$ are represented as

$$\begin{aligned} F_{ij} &= \sum_{\text{Domain}(A_i)} \frac{U_{F_{ij}}(a_i)}{a_i} \text{ if } A_i \text{ is discrete} \\ F_{ij} &= \int_{\text{Domain}(A_i)} \frac{U_{F_{ij}}(a_i)}{a_i} \text{ if } A_i \text{ is continuous} \end{aligned} \quad (1)$$

where $a_i \in \text{Domain}(A_i)$. The degree of compatibility of $a_i \in \text{Domain}(A_i)$ with linguistic term l_{ij} is given by $\bigcup_{F_{ij}}(a_i)$ [8].

4.2.2 *Fuzzy* k*-means clustering*

We have used fuzzy *k*-means clustering to find the membership function values for all continuous attributes [23]. This method divides the values of each attribute into *k*-clusters. The steps given below are used for clustering the attribute values:

1. Place *k* points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the *k*-centroids.
4. Repeat steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

5. And this algorithm aims at minimizing the objective function (squared error function [23] in our case).

$$J = \sum_{i=1}^{N}\sum_{j=1}^{k}(u_{ij})^m(x_i - c_j)^2 \tag{2}$$

where $(x_i - c_j)^2$ is a chosen distance measure between attribute value $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the N (number of records) attribute values from their respective cluster centers. The resultant clusters have to be associated with k linguistic terms. These linguistic terms are associated based on cluster centers and attribute nature. Then the membership value is calculated for each value using Cauchy function. It is given as [23]

$$u_{ij} = f(d) = \frac{1}{d^a + b} \tag{3}$$

where d is the distance from the cluster center is given by

$$d = \sum_{m=1}^{k}\left(\frac{|x - c_j|}{|x - c_m|}\right)^2 \tag{4}$$

where x is the required attribute value, c_i is the mean of ith cluster, k is the total number of clusters for that attribute and a,b are constants. The (unnormalized) degree of membership is computed as the inverse squared distance from the cluster center.

4.2.3 Adjusted difference

In order to decide whether the association between a linguistic term, L_{pq} and another linguistic term, $L_{\varphi k}$ is interesting, adjusted difference method [24] has been used. This is defined as

$$d_{L_{pq}L_{\varphi k}} = \frac{Z_{L_{pq}L_{\varphi k}}}{\sqrt{\gamma_{L_{pq}L_{\varphi k}}}} \tag{5}$$

where $Z_{L_{pq}L_{\varphi k}}$ the standardized difference, is given by

$$Z_{L_{pq}L_{\varphi k}} = \frac{\deg_{L_{pq}L_{\varphi k}} - e_{L_{pq}L_{\varphi k}}}{\sqrt{e_{L_{pq}L_{\varphi k}}}} \tag{6}$$

$e_{L_{pq}L_{\varphi k}}$ is the sum of degrees to which records are expected to be characterized by L_{pq} and $L_{\varphi k}$ and is calculated by

$$e_{L_{pq}L_{\varphi k}} = \frac{\sum_{i=1}^{m}\deg_{L_{pq}L_{\varphi i}}\sum_{j=1}^{n}\deg_{L_{pj}L_{\varphi k}}}{\sum_{j=1}^{n}\sum_{i=1}^{m}\deg_{L_{pj}L_{ki}}} \tag{7}$$

where m and n are number of linguistic terms defined on attributes φ and p and $\gamma_{L_{pq}L_{\varphi k}}$ is the maximum likelihood estimate of the variance of $Z_{L_{pq}L_{\varphi k}}$ and is given by

$$\gamma_{L_{pq}L_{\varphi k}} = \left(1 - \frac{\sum_{i=1}^{m}\deg_{L_{pq}L_{\varphi i}}}{\sum_{i=1}^{m}\sum_{j=1}^{n}\deg_{L_{pj}L_{\varphi i}}}\right) \times \left(1 - \frac{\sum_{i=1}^{n}\deg_{L_{pi}L_{\varphi k}}}{\sum_{i=1}^{m}\sum_{j=1}^{n}\deg_{L_{pj}L_{\varphi i}}}\right) \tag{8}$$

If $d_{lik ljp} > 1.96$ (the 95‰ of the normal distribution), we can conclude that the association between L_{pq} and $L_{\varphi k}$ is interesting. The first order rules which satisfy this condition are taken as the initial population.

4.2.4 Algorithm in detail

The first-order fuzzy association rules can be defined as rules involving one linguistic term in their antecedent. The second order fuzzy association rules can be defined as rules involving two linguistic terms in their antecedent. Similarly, higher order fuzzy association rules can be defined using many linguistic terms in their antecedent. The algorithm for extracting fuzzy association rules is given in Fig. 4.

Initially all the first order rules (R_1) are generated based on the adjusted difference. Those rules which are having $d_{lik ljp} > 1.96$ are selected as the first order rules and added to R which will store all order rules. The rules in R_1 are then used to generate second-order rules, which are, in turn, stored in R_2. The rules in R_2 are then used to generate third-order rules, which are stored in R_3 and so on for fourth and higher orders. This procedure is repeated to generate association rules upto a certain higher-order.

Population initialize ($\boldsymbol{R}_{k\text{-}1}$)
Begin
  for (i = 0; i < *populationsize*; i + +)
  begin
    for (j = 0; j < *nrules*; j + +)
    $chromosome_i$ $rule_j$ = *rand* ($\boldsymbol{R}_{k\text{-}1}$)
  end
P_{k1} = ∪ $chromosome_i$
End

Once the initial population is generated, fitness of the chromosomes is calculated in terms of fuzzy support and confidence [8]. Chromosome for generating rules of any order consists of "n" number of rules (alleles). The structure of a chromosome for generating second order rules is shown in Fig. 5. The cells from $L_{a_1t_1}$ to $L_{a_3t_3}$ represent a second order rule and the arrow shows the nth rule.

The fuzzy support of a linguistic term $L_{a_1t_1}$, is represented by $f\ \sup(L_{a_1t_1})$ and it is defined as follows

$$f\ \sup(L_{a_1t_1}) = \frac{\sum_{t\in D}\lambda_{L_{a_1t_1}}(t)}{\sum_{t\in D}\sum_{j=1}^{k}\lambda_{L_{a_1t_j}}(t)} \tag{9}$$

```
R1 ← {lik ⇒ ljp / i != p and d_jpik > 1.96 }
R = ∪ R1
k ← 2
While (k < m)
Begin
Population pk1 = initialize (Rk-1)
for j=2,...,n interations do
    Begin
        for each chromosome i in population pkj do
        Begin
          Calculate fuzzy support and fuzzy confidence
        End
        for each chromosome i in population pkj do
        Begin
            Rank of i th chromosome ri = 1 + m
            (m is number of chromosomes that dominates i)
        End
        for each rank ri
        Begin
            f(ri) = f(ri) + 1
        End
        for each chromosome i do
        Begin
            fitness fi = N − Σ_{k=1}^{ri−1} f(k) − 0.5(f(ri) − 1)
        End
        For each solution i in rank r
        Begin
            Calculate niche count nci
            Calculate shared fitness sfi
            Scale the shared fitness sfi
        End
        Construct roulette wheel (sfi)
        for i = 1, 2, . . . population size/2 do
        Begin
            Chrom1 = select from wheel ()
            Chrom2 = select from wheel ()
            pkj (chromi , chrom2i) = Crossover(chrom1, chrom2)
        End
        for mr (popsize * pm ) chromosomes do mutation(pkj )
    End
        Rk = pkn
        R = ∪Rk
End
```

Fig. 4 Algorithm for extracting fuzzy rules

where $L_{a_1t_1}$ represents a set of continuous attributes and corresponding terms associated with that attribute. The fuzzy support of the association $L_{a_1t_1} \Rightarrow L_{a_2t_2}$, $f\sup(L_{a_1t_1} \Rightarrow L_{a_2t_2})$ us given by

$$f\sup(L_{a_1t_1} \Rightarrow L_{a_2t_2}) = \frac{\sum_{t\in D}\min(\lambda_{L_{a_1t_1}}(t), \lambda_{L_{a_2t_2}}(t))}{\sum_{t\in D}\sum_{j=1}^{m}\sum_{k=1}^{n}\min(\lambda_{L_{a_1t_j}}(t), \lambda_{L_{a_2t_k}}(t))} \tag{10}$$

The fuzzy confidence of the association $L_{a_1t_1} \Rightarrow L_{a_2t_2}$, $f\,\mathrm{conf}(L_{a_1t_1} \Rightarrow L_{a_2t_2})$ is given by,

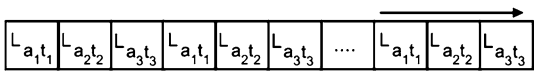

Fig. 5 Structure of a chromosome

$$f\,\mathrm{conf}(L_{a_1t_1} \Rightarrow L_{a_2t_2}) = \frac{f\sup(L_{a_1t_1} \Rightarrow L_{a_2t_2})}{f\sup(L_{a_1t_1})} \tag{11}$$

The fuzzy support and confidence are calculated for all the rules in a chromosome which are used to calculate the fitness of each chromosome. In order to calculate the fitness, rank of all the chromosomes has to be calculated. Rank of each chromosome is calculated as the number of chromosomes which are dominating the given chromosome both in fuzzy support and fuzzy confidence. Fitness of a chromosome i is given by

$$\text{Fitness} \quad f_i = N - \sum_{k=1}^{r_i-1} f(k) - 0.5(f(r_i) - 1) \tag{12}$$

where N is the population size, $f(k)$ is the number of chromosomes which are having rank k and r_i is the rank of the given chromosome. We have used niching method [9, 25] to maintain a diverse set of solutions. Niche count of each solution/chromosome nc_i is given by

$$nc_i = \sum_{j=1}^{f(r_i)} sh(d_{ij}) \tag{13}$$

where sh is sharing function, it is defined as

$$sh(d_{ij}) = 1 - \frac{d_{ij}}{\sigma} \quad \text{if } d_{ij} < \sigma \text{ otherwise } 0 \tag{14}$$

And d_{ij} is given by

$$d_{ij} = \sqrt{\sum_{k=1}^{M}\left(\frac{f_k^i - f_k^j}{f_k^{\max} - f_k^{\min}}\right)^2} \tag{15}$$

where M is the number of objectives i.e fuzzy support and confidence. $f_k^{\max}$ and $f_k^{\min}$ are the maximum and minimum objective function values of the kth objective. Then shared fitness sf_i is calculated as follows:

$$sf_i = \frac{f_i}{nc_i} \tag{16}$$

To preserve the average fitness, shared fitness is scaled using

$$sf_i = \frac{f_i * f(r_i)}{\sum_{k=1}^{f(r_i)} sf_k} * sf_i \tag{17}$$

Shared fitness of all the chromosomes having same rank is summed up in order to calculate shared fitness of an individual chromosome. Roulette wheel selection method is used for the selection of chromosomes for the cross over operation. Roulette wheel is constructed using the shared fitness of all the chromosomes. Two point cross over is used to produce next generation. Crossover operation is

clearly depicted in Fig. 6. Two points are selected randomly for each chromosome. After applying the cross over, two child chromosomes will be generated. These chromosomes are added to next generation (j) population p_{kj}.

Chromosomes for the next generation are formed by applying crossover operation for populationsize/two times. These chromosomes become the new population for the next iteration. Mutation operation is applied to the new population with a probability pm, i.e. pm*populationsize chromosomes are selected for mutation and they are reconstructed from $R_{k\text{-}1}$ rule set. Resultant population is considered as the initial population for the next iteration.

mutation (population p_{ki}*).*
Begin
  for($i = 0$; $i <$ *popsize*muatationrate*; $i++$)
  begin
    Chrom1 = rand (p_{ki})
    Initialize (chrom1, $R_{k\text{-}1}$)
  end
End

This procedure is repeated on the newly generated population for a certain number of iterations or same chromosome is selected more than 40% of the time for the crossover. Finally, rules are constructed from the final population P_{kn} in to kth order rule set R_k and these rules are added to R. Total time complexity of the algorithm for generating mth order rules is O ($n*k*m*N$), where n is the population size, k is the number of iterations and N is the number of records in the data set, n and k are constants for each rule generation.

5 Results

We applied the proposed method on computer activity dataset which describes the performance of a multi-processor system [26]. The computerActivity database

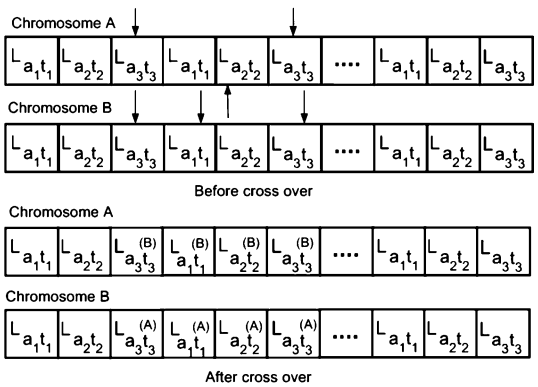

Fig. 6 Crossover operation on two chromosomes A and B

consists of various performance measures, such as the bytes read/written from sytem memory, from a Sun Sparctation 20/712 with 2 CPUs and 128 Mb of main memory. The scope of this application includes the analysis of the processor utilization and system utilization with respect to various parameters. This dataset consists of 8146 records and each is characterized by 25 attributes. Among all the 25 attributes, we identified system utilization and user utilization as the primary attributes. We used only these attributes in the consequent part of the rules and all other attributes in the antecedent part. Experimental set up consists of a Pentium IV 2.53 GHz CPU with 512 MB of memory and Linux operating system (Fedora core 2).

Three linguistic terms low, medium and high were defined for the attribute "runqsz" (process run queue size) and the attribute values were divided into these three clusters. Because of three clusters, all the attribute values belong to any one of the clusters with a membership value of greater than 0.5. The mean values are found to be {[5.59658], [2.91103] and [1.34045]} for these three clusters as calculated using fuzzy k-means clustering method based on the computed means, the linguistic terms were associated as high, medium and low, respectively. Membership values calculated are 0.00, 0.01 and 0.99 for the above attribute when its value is 1.2. Number of rules generated with different ranges of fuzzy support and fuzzy confidence and the following parameters are shown in Tables 1, 2 and 3.

Population size = 30
Number of iterations = 10
Mutation rate = 0.1
Number of alleles = 3
Rule order = 3

The obtained results clearly show that when the order of the rules increases support decreases. Based on the support and confidence values, the domain expert can find the rules which are required for his/her domain analysis.

Time taken (in seconds) to generate rules of different sizes and for different orders with 20 iterations and mutation rate 0.1, is given in Table 4 and the corresponding graph in Fig. 7. In Table 4, fourth order rules of size 120 and 150 have taken less time compared to other rule sizes. This is due to the selection of same chromosome for more than 40% of the times for crossover during rule generation, so the evolution process was terminated before the required number of iterations (20 iterations). Some of the rules which are discovered by this methodology are given below:

If fork = very high then sys = very high
If rchar = low then sys = high
If fork = low and rchar = low then sys = high

Table 1 Second-order rules

Fuzzy support range	Fuzzy confidence range			
	20–30	30–40	40–50	>50
0–10	0	1	3	0
10–20	16	7	18	3
20–30	12	13	5	2
>30	2	4	4	0

Table 2 Third-order rules

Fuzzy support range	Fuzzy confidence range			
	20–30	30–40	40–50	>50
0–10	6	13	13	0
10–20	14	12	14	1
20–30	4	2	5	1
>30	0	1	4	0

Table 3 Fourth-order rules

Fuzzy support range	Fuzzy confidence range			
	20–30	30–40	40–50	>50
0–5	7	1	12	1
5–10	15	15	19	1
>10	2	4	3	0

5.1 Intrusion Detection

We obtained a set of TCPDUMP data, available at [27]. The data set consists of three runs of tcpdump on generated network intrusions and one tcpdump run on normal network traffic (with no intrusions). Our experiments focused on building an anomaly detection model from the normal dataset using fuzzy association rule mining. As part of preprocessing, we developed a script to scan each tcpdump data file and extracted the connection level information about the network traffic. For each TCP connection, the script processes packets between the two ports of the participating hosts and calculate the statistics which include duration of the connection, bytes sent in each direction, control packets ratio and data packets ratio. Since UDP is connectionless (no connection state), we simply treat each packet as a connection. Attribute name, nature and the number of fuzzy sets defined on each attribute are given in Table 5.

Table 4 Time (in seconds) taken to generate rules

Number of rules	Order of the rules			
	First	Second	Third	Fourth
30	0.31	10.93	13.09	16.5
60	0.57	20.57	48.78	62.1
90	1.68	51.81	69.02	77.98
120	2.05	57.65	84.69	45.63
150	2.26	68.35	90.03	67.47

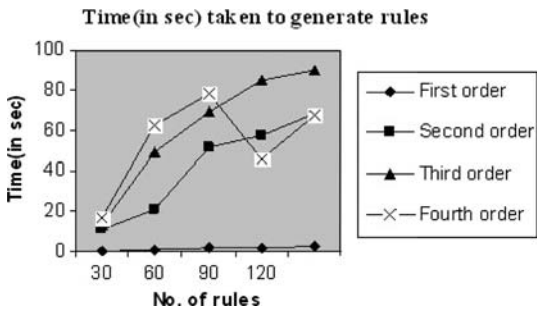

Fig. 7 No. of rules versus time (in seconds)

We used fuzzy *k*-means clustering for defining the membership functions for all the continuous attributes. Four linguistic terms medium, low, high and very low were associated for the attribute "duration" using fuzzy *k*-means clustering.

Initially, the proposed algorithm is executed to generate the rules in terms of the fuzzy terms from the audit data. To make these rules consistent, algorithm is executed many times under different settings. For each new run, we compute its rule set from the audit trail and update the (existing) aggregate rule sets using a count for each rule. Counts are incremented when there is a match between the newly extracted rule and the rule in the aggregate rule set. When the rule set stabilizes (there are no new rules added), we can stop the data gathering process since we have produced a near complete set of audit data for the normal runs. We then prune the rule set by eliminating the rules with low count.

Table 5 Attribute nature and fuzzy sets

Attribute name	Type	No. of fuzzy sets
Connection duration	Continuous	4
Protocol	Categorical	TCP/UDP
Bytes sent	Continuous	5
Bytes received	Continuous	5
Control packets ratio	Continuous	4
Data packets ratio	Continuous	5

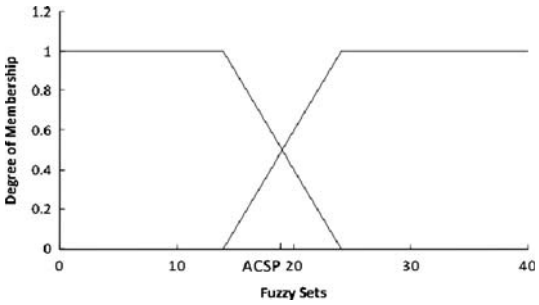

Fig. 8 Fuzzy sets for ICF

Table 6 Evaluation of datasets

Data set name	ICF	Result	Membership
Normal1	19.42	No intrusion	0.64
Normal2	25.60	No intrusion	1.00
Intrusion 1	15.77	Intrusion	0.72
Intrusion 2	10.63	Intrusion	1.00
Intrusion 3	12.23	Intrusion	1.00

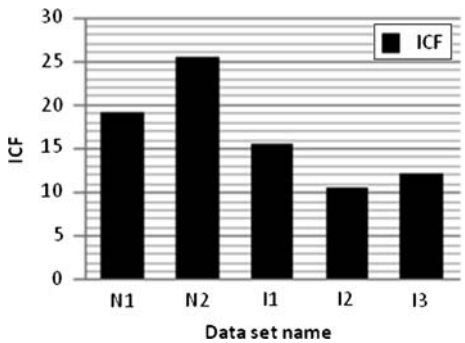

Fig. 9 ICF values of different datasets

In order to detect the intrusions in the network, we defined a new measure called "Intrusion Chance Factor" which is given by

$$\mathrm{ICF} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n}\sum_{k=1}^{m}\deg_{L_{jk}}}{(\text{matched\_records})} \tag{18}$$

where r is the number of records in the new dataset, n is the number of rules in Aggregate rule set, m is the order of the rule and matched_records is the number of matched records. A matched record is one which has at least one nonzero membership attribute in a given rule. We evaluated five new datasets, in which two were normal datasets and the other three were intruded datasets. All the five datasets are initially converted to the connection specific format using earlier mentioned approach. We defined another measure called Average Confidence and Support Product (ACSP), which is given as the product of average support and average confidence of all the rules in the aggregate rule set. We defined two fuzzy sets and two normal sets due to the fuzziness of ICF. Two normal sets define the regions where there is no chance of intrusion and another with full chance of intrusion. Whereas the fuzzy sets define the Intrusion or no intrusion with certain membership value. Based on the dominance, decision has to be made. Fuzzy sets on ICF are clearly shown in Fig. 8.

Fuzzy sets are defined based on the value of ACSP. In this case, we defined two fuzzy sets within a region of distance α to the either sides of ACSP. In this case, we selected α as 5. ACSP value for our aggregate data set is approximately 18. The calculated ICF, ACSP and result for all the five datasets are shown in Table 6 and the corresponding graph is shown in Fig. 9.

6 Conclusion

In this paper, we proposed a solution approach for extracting fuzzy association rules in terms of linguistic terms. Unlike other methods which define membership functions based on intuition, our approach uses the fuzzy k-means clustering for calculating fuzzy membership functions. We have used adjusted difference for finding the interestingness among the attributes in order to generate the initial rules. The important feature of this approach is that it uses MOGA for extracting the rules. MOGA finds the set of trade off optimal solutions by considering fuzzy support and fuzzy confidence as objectives.

References

1. Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of Items in large databases. In: Proceedings of the ACM SIGMOD international conference, Washington, pp 207–216
2. Gyenesei A (2000b) A fuzzy approach for mining quantitative association rules. TUCS Technical Report No. 336, pp 1–18
3. Ng RT, Han J (1994) Efficient and effective clustering methods for spatial data mining. In: Proceedings of the 20th VLDB conference, pp 144–155
4. Srikant R, Agrawal R (1996) Mining quantitative association rules in large relational tables. In: Proceedings of the ACM SIGMOD internationl conference, Canada, pp 1–12
5. Gyenesei A (2000a) Mining weighted association rules for fuzzy quantitative items TUCS Technical Report No. 346, pp 1–12
6. Yager RR (1996) On linguistic summaries of data, knowledge discovery in database, pp 347–363
7. Chan KCC, Au W (1997) Mining fuzzy association rules. In: Proceedings of the sixth international conference on information and knowledge management, pp 209–215
8. Au W-H, Chan KCC (2003) Mining fuzzy association rules in a bank-account database. IEEE Trans Fuzzy Syst, 11(2)
9. Fonseca CM, Fleming PJ (1993) Genetic algorithms for multiobjective optimization: formulataion, discussion and

generalization. In: Proceedings of fifth international conference on genetic algorithms, pp 416–423
10. Goldberg DE (1989) Genetic algorithms for search, optimization and machine learning. Addison-Wesley, Reading
11. Lunt T (1993) Detecting intruders in computer systems. In: Proceedings of 1993 conference on auditing and computer technology
12. Ilgun K, A Kemmerer (1995) State transition analysis: a rule-based intrusion detection approach. IEEE Trans Softw Eng 21(3):181–99
13. Teng H, Chen K, Lu S (1990) Adaptive real-time anomaly detection using inductively generated sequential patterns. In: Proceedings of IEEE computer society symposium on research in security and privacy, California, pp 278–84
14. Debar H, Becker M, Siboni D (1992) A neural network component for an intrusion detection system. In: Proceedings of IEEE computer society symposium on research in security and privacy, California, pp 240–50
15. Lee W, Stolfo S, Mok K (1998) Mining audit data to build intrusion detection models. In: Proceedings of the fourth international conference on knowledge discovery and data mining, New York, pp 66–72
16. Bridges SM, Vaughn RB (2000) Fuzzy data mining and genetic algorithms applied to intrusion detection. In: The national information systems security conference, Baltimore
17. Kuok CM, Fu A, Wong MH (1998) Mining fuzzy association rules in databases. In: Proceedings of the ACM SIGMOD conference on management of data, pp 41–46
18. Zhang W (1999) Mining fuzzy quantitative association rules. In: Proceedings of 11th IEEE international conference on tools with artificial intelligence, pp 99–102
19. Ishibuchi H, Nakashima T, Murata T (2001) Three-objective genetics-based machine learning for linguistic rule extraction. Inf Sci 136:109–133
20. Barandela R, Valdovinos RM, Sanchez JS (2003) New applications of ensembles of classifiers. Pattern Anal Appl 6:245–256
21. Kaya M, Alhajj R (2003) Facilitating fuzzy association rules mining by using multi-objective genetic algorithms for automated clustering. In: Proceedings of the third IEEE international conference on data mining (ICDM'03), pp 561–564
22. Deb K, Pratap A, Agarwal S, Meyarivan TA (2002) Fast and elitist multi-objective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 6(2):181–197
23. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum, New York
24. Au W-H, Chan KCC (1999) FARM: a data mining system for discovering fuzzy association rules. In: Proceedings of 8th IEEE international conference on fuzzy systems, Seoul, Korea, pp 1217-1222
25. Laumanns M, Thiele L, Deb K, Zitzler E (2002) Combining convergence and diversity in evolutionary multiobjective optimization. Evol Comput J 10(3):263–282
26. Computer Activity Dataset: http://delve/data/compactiv/compActivDetail.html
27. TCPdump Dataset: http://iris.cs.uml.edu:8080/network.html

Author Biographies

P. Santhi Thilagam received her Bachelor degree in Computer Science and Engineering in 1991 from Thiagarajar College of Engineering, Madurai, India. She received her Master degree in Computer Science and Engineering in 2000 from Anna University, Chennai, India. Since 2003 she has been pursuing the Doctoral degree and working as a Sr.Lecturer in the Department of Computer Engineering, National Institute of Technology Karnataka, Surathkal, India. Her research interests are mainly in the areas of database systems, data mining.

V. S. Ananthanarayana received his Bachelor degree in Computer Science and Engineering in 1990 from Karnatak University, India, his Master degree in Computer Science and Engineering in 1995 and his Doctoral degree in 2001 from Indian Institute of Science, Bangalore, India, and Post Doctoral Fellow in 2005 from Memorial University of Newfoundland, Canada. He is currently a professor of Department of Information Technology, National Institute of Technology Karnataka, Surathkal, India. His research interests are in the areas of database systems, data mining, distributed computing and web services.