Learning the Membership Function Contexts for Mining Fuzzy Association Rules by Using Genetic Algorithms

Jesús Alcalá-Fdez, Rafael Alcalá, María José Gacto, Francisco Herrera

Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain

Abstract

Different studies have proposed methods for mining fuzzy association rules from quantitative data, where the membership functions were assumed to be known in advance. However, it is not an easy task to know a priori the most appropriate fuzzy sets that cover the domains of quantitative attributes for mining fuzzy association rules.

This paper thus presents a new fuzzy data-mining algorithm for extracting both fuzzy association rules and membership functions by means of a genetic learning of the membership functions and a basic method for mining fuzzy association rules. It is based on the 2-tuples linguistic representation model allowing us to adjust the context associated to the linguistic term membership functions. Experimental results show the effectiveness of the framework.

Key words: Data mining, fuzzy association rules, genetic algorithms, genetic fuzzy systems, 2-tuples linguistic representation.

1 Introduction

Data Mining (DM) is the process for automatic discovery of high level knowledge by obtaining information from real data. Discovering association rules is one of the several data mining techniques described in the literature [15].

URLs: jalcala@decsai.ugr.es (Jesús Alcalá-Fdez), alcala@decsai.ugr.es (Rafael Alcalá), mjgacto@ugr.es (María José Gacto), herrera@decsai.ugr.es (Francisco Herrera).

Association rules are used to represent and identify dependencies between items in a database [36]. These are an expression of the type $X \rightarrow Y$, where X and Y are sets of items and $X \cap Y = \oslash$. It means that if all the items in X exist in a transaction then all the items in Y are also in the transaction with a high probability, and X and Y should not have a common item [1,2]. Many previous studies focused on databases with binary values, however the data in real-world applications usually consist of quantitative values. Designing DM algorithms, able to deal with various types of data, presents a challenge to workers in this research field.

Fuzzy set theory has been used more and more frequently in intelligent systems because of its simplicity and similarity to human reasoning [23]. The use of fuzzy sets to describe association between data extends the types of relationships that may be represented, facilitates the interpretation of rules in linguistic terms, and avoids unnatural boundaries in the partitioning of the attribute domains [9–11,22,34].

Different studies have proposed methods for mining fuzzy association rules from quantitative data [20,21,28,29,33], where the membership functions (MFs) were assumed to be known in advance. The given MFs may have a critical influence on the final mining results. For this reason, some approaches have also achieved a learning or tuning of the MFs [14,18,19,25–27,35].

Recently, a new linguistic rule representation model has been proposed to perform a genetic lateral tuning of MFs [3]. This new approach is based on the 2-tuples linguistic representation [17], that allows the symbolic translation of a linguistic term by considering only one parameter. In this way, two main objectives were achieved: to tune MFs by maintaining a high covering degree of the data, and to reduce the search space respect to the classic tuning [6] (usually considering three parameters in the case of triangular MFs), in order to easily obtain optimal models.

The automatic definition of fuzzy systems can be considered as an optimization or search process and nowadays Evolutionary Algorithms, particularly Genetic Algorithms (GAs), are considered as the better known and used global search technique. The genetic coding that GAs use allow them to include prior knowledge and to use it for leading the search up. For this reason, GAs have been successfully applied to learn and to tune fuzzy systems in the last years [5,6,16].

Based on the 2-tuples linguistic representation model, in this paper we present a new fuzzy data-mining algorithm for extracting both fuzzy association rules and MFs from quantitative transactions by means of a genetic learning of the MFs and the use of a basic method for mining the fuzzy association rules. In this way, the search space reduction provided by the 2-tuples linguistic representation helps the genetic search technique to obtain more suitable MFs.

Moreover, this way to work allows us to learn the most adequate context [7,8] for each fuzzy partition, which is necessary in different contextual situations with the aim of getting high quality fuzzy association rules.

The scheme considered for discovering both useful fuzzy association rules and suitable MFs from quantitative values is comprised of two stages (see Fig. 1):

(1) A genetic process to learn the MFs.
(2) A method to mine fuzzy association rules. The method presented in [20] will be considered for this task as a first approach.

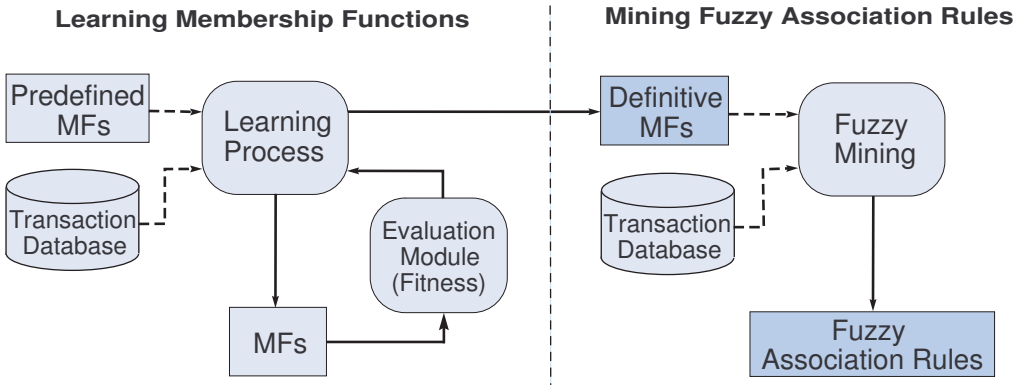

Fig. 1. Scheme for discovering both useful fuzzy association rules and suitable MFs

We will develop this approach in this paper. We will propose a genetic learning process for getting the MFs together with a mining process for getting the fuzzy association rules.

We will also present an experimental study for showing the behaviour of the proposed approach using a public data base, FAM95 [1] . We will develop a double study, first, we will show the results obtained by our proposal, comparing it with the classical one using the uniform partition and the well known approach presented by Hong et al. in [19], which also performs a genetic learning of the MFs. Second, we will revise the fuzzy association rules obtained with our approach via support and confidence and we will analyse the complexity and scalability of the proposed approach.

To do that, the paper is arranged as follows. The next section describes the linguistic rule representation model based on the linguistic 2-tuples. Section 3 details the genetic learning components proposed to obtain the MFs. Section 4 describes the proposed mining process. Section 5 shows the results of the proposed mining algorithm applied over a real-world database. Finally, Section 6 points out some concluding remarks.

[1] This data base was obtained from the UCLA Statistics Data Sets Archive website http://www.stat.ucla.edu/data/fpp.

2 Preliminaries: The 2-tuples linguistic representation

The 2-tuples linguistic representation scheme presented in [17], introduces a new model for rule representation based on the concept of symbolic translation (the lateral displacement of a linguistic term).

The symbolic translation of a linguistic term is a number within the interval [-0.5, 0.5) that expresses the domain of a linguistic term when it is moving between its two lateral linguistic term. Let us consider a set of linguistic terms S representing a fuzzy partition. Formally, we have the pair,

$$(s_i, \alpha_i), \;\; s_i \in S, \; \alpha_i \in [-0.5, 0.5).$$

Fig. 2 depicts the symbolic translation of a linguistic term represented by the pair $(S_2, \; -0.3)$, considering a set S with five linguistic terms represented by their ordinal values ($\{S_0, S_1, S_2, S_3, S_4\}$).

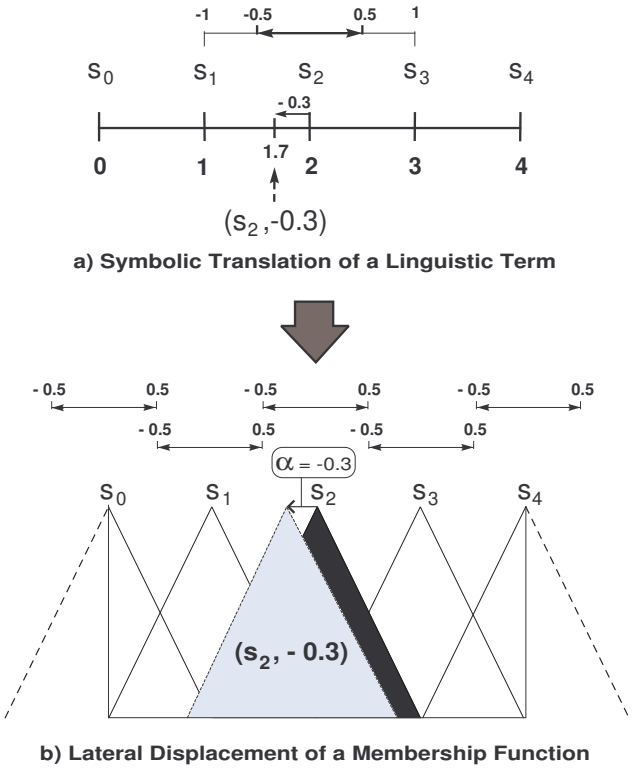

Fig. 2. Symbolic translation of a linguistic term and lateral displacement of the involved MF.

In [17], both the 2-tuples linguistic representation model and the needed elements for linguistic information comparison and aggregation are presented and applied to the Decision Making framework. In [3], a new rule representation model has been presented based on these concepts to perform a tuning of

complex linguistic fuzzy models. In this work, we extend its use for fuzzy association rule representation. Below we present this approach considering a simple mining problem.

Let us consider a simple problem with two items (age and weight) and three linguistic terms with their associated MFs (see Fig. 3). Based on this definition, an example of classic fuzzy association rule and 2-tuples fuzzy linguistic representation based rule is:

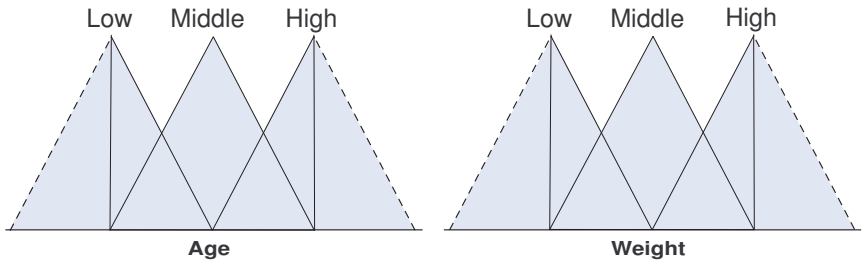

Fig. 3. Items and linguistic terms in a simple problem

Classic Fuzzy Association Rule,
If **Age** is Middle then **Weight** is High.

Rule with 2-Tuples Fuzzy Linguistic Representation,
If **Age** is (Middle,0.3) then **Weight** is (High,-0.1).

This proposal decreases the size of the tuning search space, since the three parameters usually considered per linguistic term [6] are reduced to only one symbolic translation parameter. Moreover, from the point of view of interpretability:

- the original shapes of the MFs are maintained (in our case triangular and symmetrical), by laterally changing the location of their supports,
- the lateral variation of the involved MFs is restricted to a short interval, *ensuring overlapping between two adjacent MFs* to some degree but preventing their vertex points from crossing, and
- the 2-tuples representation based linguistic terms can be interpreted with respect to the initial ones.

Analysed from the rule interpretability point of view, we could interpret the previous 2-tuples linguistic representation based rule in the following way:

If **Age** is (higher than Middle)
then **Weight** is (a bit smaller than High).

3 Genetic Learning Process Components to Obtain the MFs

In this paper, we will consider the use of GAs to design the proposed learning method of the MFs. A good genetic model is the CHC genetic model [12]. The CHC algorithm is a GA that presents a good trade-off between exploration and exploitation, being a good choice in problems with complex search spaces.

In the following, the components needed to design this GA are explained. They are:

- CHC genetic model.
- MFs codification and initial gene pool.
- Chromosome evaluation.
- Crossover operator.
- Restart Approach.

3.1 CHC Genetic Model

We will consider a population-based selection approach, by using the CHC genetic model [12] in order to perform an adequate global search. The genetic model of CHC makes use of a "Population-based Selection" approach. N parents and their corresponding offspring compete to select the best N individuals to take part of the next population. The CHC approach makes use of an incest prevention mechanism and a restarting process to provoke diversity in the population, instead of the well known mutation operator.

This incest prevention mechanism will be considered in order to apply the crossover operator, i.e., two parents are crossed if their hamming distance divided by 2 is over a predetermined threshold, L. Since we will consider a real coding scheme, we have to transform each gene considering a Gray Code with a fixed number of bits per gene ($BITSGENE$) determined by the expert. In this way, the threshold value is initialised as:

$$L = (\#Genes * BITSGENE)/4.0$$

where $\#Genes$ is the number of genes in the chromosome (for more information, see [13]). Following the original CHC scheme, L is decremented by one when there is no new individuals in the population in one generation. In order to make this procedure independent of $\#Genes$ and $BITSGENE$, in our case, L will be decremented by a $\varphi\%$ of its initial value (being φ determined by the user, usually 10%). The algorithm restarts when L is below zero.

A scheme of this algorithm is shown in Fig. 4.

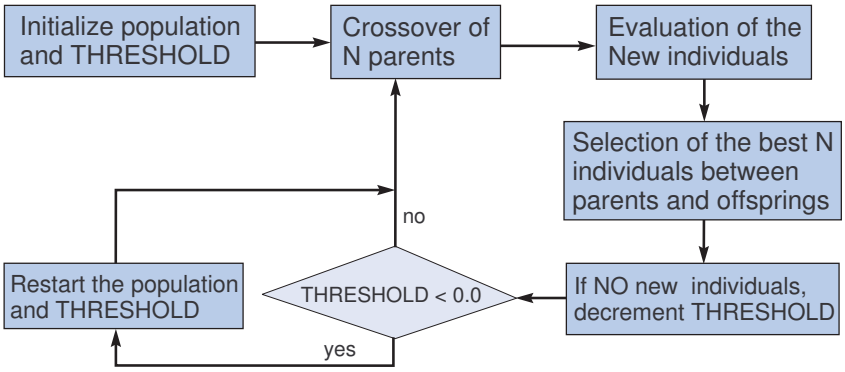

Fig. 4. Scheme of CHC

3.2 MFs Codification and Initial Gene Pool

A real coding scheme is considered, i.e., the real parameters are the GA representation units (genes). Each chromosome is a vector of real numbers with size $n * m$ (n items with m linguistic terms per item) in which the displacements of the different linguistic terms are coded for each item. Then, a chromosome has the following form (where each gene is the displacement value of the corresponding linguistic term),

$$(c_{11}, \ldots, c_{1m}, c_{21}, \ldots, c_{2m}, \ldots, c_{n1}, \ldots, c_{nm})$$

Fig. 5 graphically depicts an example of correspondence between a chromosome and its associated MFs. Notice that, the three parameters usually considered per linguistic term (in the case of triangular MFs) are reduced to only one parameter.

To make use of the available information, the initial MFs obtained from expert knowledge are included in the population as an initial solution. To do so, the initial pool is obtained with the first individual having all genes with value '0.0', and the remaining individuals generated at random in [-0.5, 0.5).

3.3 Chromosome Evaluation

To evaluate a determined chromosome we will use the fitness functions defined in [18]. The fitness value of a chromosome C_q is defined as:

$$fitness(C_q) = \frac{\sum_{x \in L_1} fuzzy\_support(x)}{suitability(C_q)}$$

where L_1 is the set of large 1-itemsets obtained by using the set of MFs in

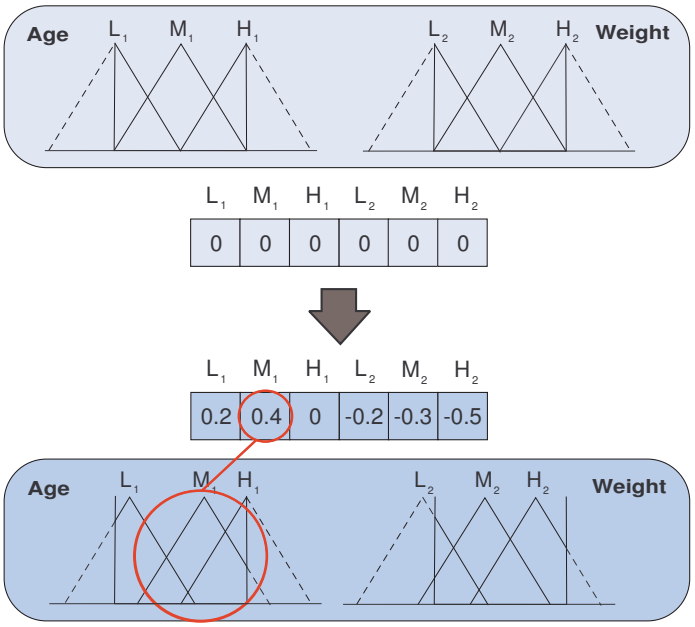

Fig. 5. Example of coding scheme

C_q, $fuzzy\_support(x)$ is the fuzzy support of the 1-itemset x from the given transaction database [24], and $suitability(C_q)$ represents the shape suitability of the MFs from C_q. The suitability of the set of MFs in a chromosome C_q is defined as:

$$suitability(C_q) = \sum_{k=1}^{n} [overlap\_factor(C_{qk}) + coverage\_factor(C_{qk})]$$

where n is number of items, $overlap\_factor(C_{qk})$ is the overlap factor of the MFs for an item I_k in the chromosome C_q, and $coverage\_factor(C_{qk})$ is the coverage factor of the MFs for an item I_k in the chromosome C_q.

The overlap factor represents the overlap ratio of the MFs for an item I_k in the chromosome C_q. The overlap ratio of two MFs R_i and R_j $(i < j)$ is defined as the overlap length divided by the minimum of the right span of R_i (right extreme minus vertex) and the left span of R_j (vertex minus left extreme). If the overlap length is larger than the minimum of the above two spans, then these two MFs are thought of as a little redundant. Appropriate punishment must then be considered in this case. Thus, the overlap factor of the MFs for an item I_k in the chromosome C_q is defined as:

$$overlap\_factor(C_{qk}) = \sum_{i=1}^{m} \sum_{j=i+1}^{m} [max(\frac{overlap(R_i, R_j)}{min(spanR_{R_i}, spanL_{R_j})}, 1) - 1]$$

where $overlap(R_i, R_j)$ is the overlap length of R_i and R_j, $spanR_{R_i}$ is the right span of R_i, $spanL_{R_j}$ is the left span of R_j and m is the number of MFs for

I_k. Notice that, in our case $spanR_{R_i}$ and $spanR_{R_j}$ are the same size because the displacements of the MFs are performed on the uniform partition and the original shapes of the MFs are maintained (triangular and symmetrical).

The coverage factor represents the coverage ratio of the MFs for an item I_k in the chromosome C_q. The coverage ratio of MFs for an item I_k is defined as the coverage range of the functions divided by the maximum quantity of that item in the transactions. The more the coverage ratio is, the better the derived MFs are. Thus, the coverage factor of the MFs for an item I_k in the chromosome C_q is defined as:

$$coverage\_factor(C_{qk}) = \frac{1}{\frac{range(R_1,...,R_m)}{max(I_k)}}$$

where $range(R_1, R_2, ..., R_m)$ is the coverage range of the MFs and $max(I_k)$ is the maximum quantity of I_k in the transactions. Notice that the coverage factor is always 1 because in our case *the 2-tuples linguistic representation ensures the coverage in all the domain*, reducing the computation time. Thus, the suitability of the set of MFs in a chromosome C_q is therefore defined as:

$$suitability(C_q) = \sum_{k=1}^{n} [overlap\_factor(C_{qk}) + 1]$$

The suitability factor can reduce the occurrence of the two bad kinds of MFs shown in Fig. 6, where the first one is too redundant, and the second one is too separate. The overlap factor in $suitable(C_q)$ is used for avoiding the first bad case, and the 2-tuples linguistic representation prevents the second one.

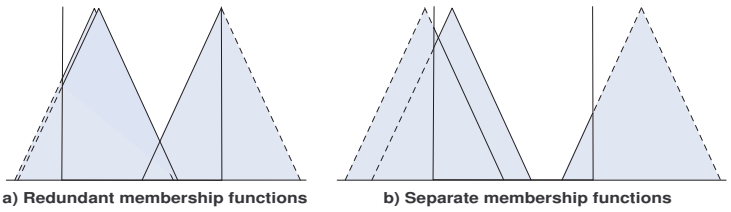

Fig. 6. Two bad kinds of membership functions

3.4 Crossover Operator

The crossover operator is based on the concept of neighbourhood. These kinds of operators present a good cooperation when they are introduced within genetic models forcing the convergence by pressure on the offspring (as the case of CHC). Particularly, we consider the Parent Centric BLX (PCBLX) operator [31], which is based on the BLX-α. Fig. 7 shows the performance

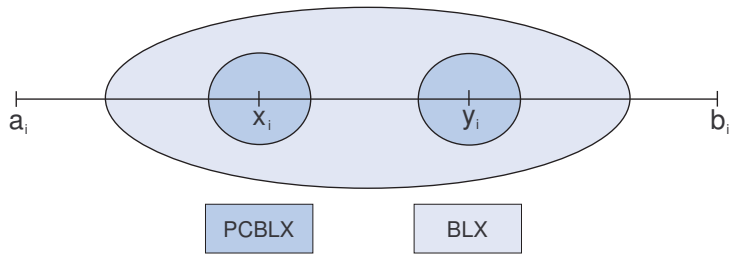

Fig. 7. Diagram of the performance of the crossover operators based on environments

of these kinds of operators, which allow the offspring genes to be around the genes of one parent or around a wide zone determined by both parent genes.

The PCBLX operator is described as follows. Let us assume that $X = (x_1 \cdots x_n)$ and $Y = (y_1 \cdots y_n)$, $(x_i, y_i \in [a_i, b_i] \subset \Re, i = 1 \cdots n)$, are two real-coded chromosomes that are going to be crossed. We generate the two following offspring:

- $O_1 = (o_{11} \cdots o_{1n})$, where o_{1i} is a randomly (uniformly) chosen number from the interval $[l_i^1, u_i^1]$, with $l_i^1 = max\{a_i, x_i - I_i \cdot \alpha\}$, $u_i^1 = min\{b_i, x_i + I_i \cdot \alpha\}$, and $I_i =\mid x_i - y_i \mid$.
- $O_2 = (o_{21} \cdots o_{2n})$, where o_{2i} is a randomly (uniformly) chosen number from the interval $[l_i^2, u_i^2]$, with $l_i^2 = max\{a_i, y_i - I_i \cdot \alpha\}$ and $u_i^2 = min\{b_i, y_i + I_i \cdot \alpha\}$.

3.5 Restart Approach

To get away from local optima, this algorithm uses a restart approach [12]. In this case, the best chromosome is maintained and the remaining are generated at random within the corresponding variation intervals [-0.5, 0.5). It follows the principles of CHC [12], performing the restart procedure when a threshold value is reached or all the individuals coexisting in the population are very similar.

4 Genetic based Mining Process

According to the above description, the proposed algorithm for mining both MFs and fuzzy association rules is described below.

INPUT: T quantitative transaction data, a set of n items, each with m predefined linguistic terms, a support threshold α, a confidence threshold λ and a population size N.

OUTPUT: A set of fuzzy association rules with its associated set of MFs.

Stage 1. *Genetic learning of the MFs.*

Step 1: Generate the initial population with N chromosomes.

Step 2: Evaluate the population. For each chromosome:

- For each transaction datum D_i, i=l to T, and for each item I_j, j=l to n, transfer the quantitative value $v_j^{(i)}$ ($D_i = (v_1^{(i)},...,v_n^{(i)})$) into a fuzzy set $f_j^{(i)}$ represented as:

$$f_j^{(i)} = \{\frac{f_{j1}^{(i)}}{R_{j1}} + ... + \frac{f_{jm}^{(i)}}{R_{jm}}\}$$

  using the corresponding MFs represented by the chromosome, where R_{jk}, is the $k-th$ linguistic term of item I_j, $f_{jk}^{(i)}$ is $v_j^{(i)}$'s fuzzy membership value in region R_{jk}, and m is the number of linguistic terms for I_j.
- For each linguistic term R_{jk}, calculate its count on the transactions as follows:

$$count_{jk} = \sum_{i=1}^{T} f_{jk}^{(i)}$$

- For each R_{jk}, $1 < j < n$ and $1 < k < m$, check whether its $count_{jk}$ larger than or equal to the minimum support threshold α. If R_{jk} satisfies the above condition, put it in the set of large 1-itemsets (L_1). That is:

$$L_1 = \{R_{jk} \mid count_{jk} \geq \alpha, 1 \leq j \leq n \; and \; 1 \leq k \leq m\}$$

- Set the fitness value of the chromosome as the sum of the fuzzy support (the count / T) of the linguistic terms in L_1 divided by $suitability(C_q)$. That is:

$$fitness(C_q) = \frac{\sum_{x \in L_1} fuzzy\_support(x)}{suitability(C_q)}$$

Step 3: Initialise the threshold value L.

Step 4: Generate the next population:

- Shuffle the population.
- Select the parents two by two. Each pair is crossed if the hamming distance between the parent Gray codings divided by 2 is over L.
- Evaluate the new individuals.
- Join the parents with their offspring and select the best N individuals to take part of the next population.

Step 5: If the best chromosome does not change or there are no new individuals in the population, $L = L - (L_{initial} * 0.1)$.

Step 6: If $L < 0$, restart the population.

Step 7: If the maximum number of evaluations is not reached, go to Step 4.

Stage 2. *Basic method for mining fuzzy association rules.*

Step 8: The set of the best MFs is then used to mine fuzzy association rules from the given quantitative database. The fuzzy mining algorithm proposed in [20] is then adopted to achieve this purpose.

5 Experimental Results

To evaluate the usefulness of the proposed approach several experiments have been carried on a real-world database with 63,756 transactions, FAM95. In these experiments, we compare the proposed approach with one uniform fuzzy partition and with Hong et al.'s approach proposed in [19], which also performs a genetic learning of the MFs.

In the following subsections, first we describe the real-world database, then we show the results obtained from the comparison with other approaches, later on we revise the fuzzy association rules via supports and confidences, and finally we analyse the complexity and scalability of the proposed approach.

5.1 Problem Description and Experiments

The real-world database FAM95 contains data for the 63,756 families that were interviewed in the March 1995 Current Population Survey, conducted by the Bureau of the Census for the Bureau of Labor Statistics. C. Yarbrough (Santa Rosa) and D Freedman (Berkeley) transcribed the data from a public-use microdata tape supplied by the Bureau of the Census and they are responsible for any errors of transcription or interpretation.

This database consists of 63,756 family records with 23 attributes each one [2]. To develop the different experiments, we extracted the 10 quantitative at-

[2] This data set was obtained from the UCLA Statistics Data Sets Archive website http://www.stat.ucla.edu/data/fpp.

tributes from them: age of head of the family, number of persons in the family, number of children, hours head worked last week, head's personal income, family income, taxable income for head, federal tax for head, final sampling weight and March supplement weight for income and tax.

The initial linguistic partitions are comprised by 3 and 5 linguistic terms with uniformly distributed triangular MFs giving meaning to them. The following values have been considered for the parameters of each approach [3]:

- Genetic process: 50 individuals, 10,000 evaluations, 30 bits per gene for the Gray codification, 0.6 as crossover probability (0.01 as mutation probability and 0.35 for the factor d in the max-min-arithmetical crossover for Hong et al.'s approach).
- Method for mining fuzzy association rules: 0.8 for the confidence threshold.

5.2 Results and Analysis

The results obtained in the genetic process by the analysed approaches are presented in Table 1, where Sup stands for the minimum support, Fit for the fitness value, F_{sup} for the sum of the fuzzy support of the large 1-itemsets, $Suit$ for the suitability and $\#1I$ for the number of large 1-itemsets.

Table 1
Results obtained in the genetic process.

	Proposed Approach				Hong et al's Approach				Uniform Fuzzy Partition			
Sup	Fit	F_{sup}	Suit	#1I	Fit	F_{sup}	Suit	#1I	Fit	F_{sup}	Suit	#1I
With 3 Linguistic Terms												
0.2	0.99	11.68	11.85	20	0.68	10.83	15.83	19	0.92	9.24	10.00	16
0.5	0.94	11.68	12.39	17	0.53	10.28	19.45	15	0.76	7.55	10.00	10
0.7	0.66	6.98	10.63	9	0.37	6.55	17.94	8	0.57	5.71	10.00	7
0.9	0.28	2.80	10.00	3	0.00	0.00	14.75	0	0.00	0.00	10.00	0
With 5 Linguistic Terms												
0.2	0.95	10.46	10.99	22	0.53	10.22	19.27	22	0.94	9.43	10.00	21
0.5	0.77	9.92	12.92	15	0.38	7.95	20.63	12	0.46	4.57	10.00	7
0.7	0.61	7.69	12.57	10	0.20	3.96	19.54	5	0.24	2.36	10.00	3
0.9	0.10	0.92	10.00	1	0.06	0.90	15.01	1	0.00	0.00	10.00	0

[3] With these values we have tried to ease the comparisons selecting standard common parameters that work well in most cases instead of searching very specific values for each approach.

Analysing the results presented in Table 1, we can highlight the following conclusions:

- The best results are obtained by the proposed approach, presenting a good relationship between the size of the search space and the results obtained, and getting a good trade-off between fuzzy support and suitability. Fig. 8 shows the average fitness values of the chromosomes along with different numbers of evaluations of the proposed approach and Hong et al.' approach with 3 linguistic terms and 0.2 as minimum support.

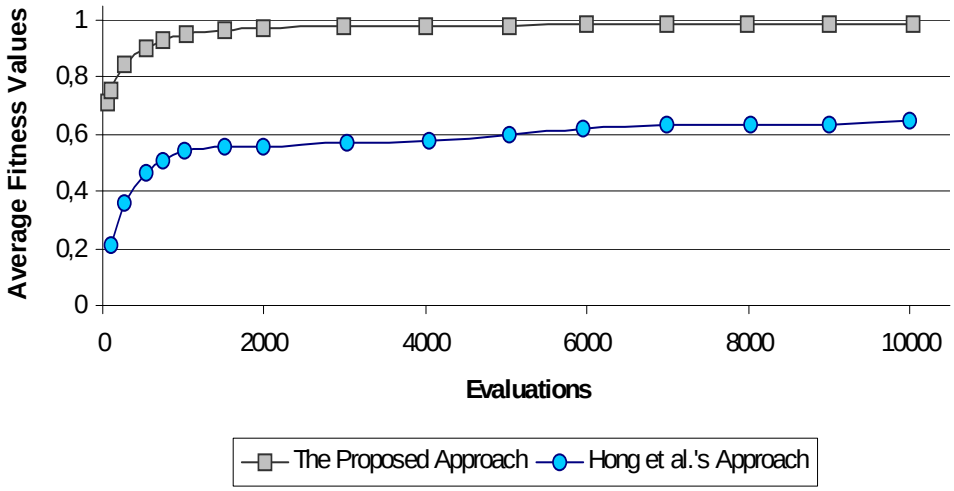

Fig. 8. The average fitness values along with different numbers of evaluations

- The proposed approach achieves larger or equal number of large 1-itemsets than the remaining approaches, which make easy to obtain larger number of rules. Fig. 9 shows the relationship between the number of large 1-itemsets

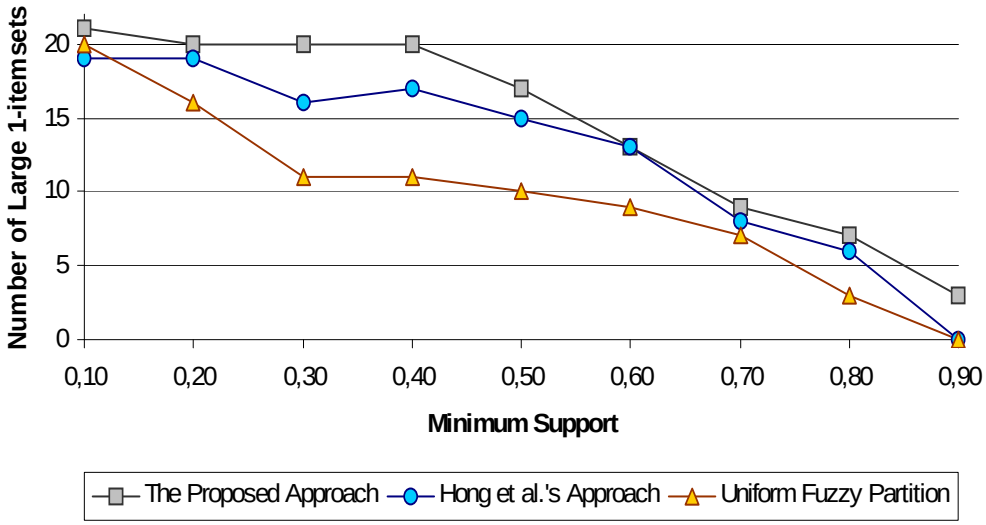

Fig. 9. Relationship between large 1-itemsets and minimum support

and the values for the minimum support with 3 linguistic terms.

- Obviously, the uniform fuzzy partition always obtains the best results for the suitability. However, the proposed approach obtains values of suitability very near to the uniform partition and better Hong et al.'s approach for the different values of minimum support, presenting the MFs obtained a good shape suitability. Furthermore, the MFs obtained are interpretables in a high level since the original shapes of the initial MFs are maintained and the new ones are directly related to the initial ones by means of the 2-tuples representation.

Table 2 presents the results obtained in the genetic process by Hong et al.'s approach with the 2-tuples linguistic representation. Comparing the results obtained with the results presented in Table 1 we can highlight that the 2-tuples linguistic representation allows us to highly improve the fitness values obtained by Hong et al.'s approach, achieving suitability values similar to the proposed approach.

Table 2
Results obtained in the genetic process.

Hong et al's Approach with the 2-tuples				
Sup	Fit	F_{sup}	Suit	#1I
With 3 Linguistic Terms				
0.2	0.97	10.90	11.18	20
0.5	0.89	11.36	12.64	18
0.7	0.59	6.20	10.33	7
0.9	0.26	2.79	10.52	3
With 5 Linguistic Terms				
0.2	0.93	10.18	10.93	22
0.5	0.64	7.39	11.80	11
0.7	0.41	4.76	11.60	6
0.9	0.08	0.91	10.92	1

Fig. 10 and Fig. 11 depict the final MFs obtained with 3 linguistic terms and 0.2 as minimum support by the proposed approach and Hong et al.'s approach, respectively. Fig. 10 shows how small displacements in the MFs lead to important improvements in the number of obtained large 1-itemsets. Furthermore, the MFs are more or less well distributed, which makes easy to find their corresponding meanings for an expert. Fig. 11 shows how the MFs obtained by Hong et al.' approach also are more or less well distributed but they present a larger overlap.

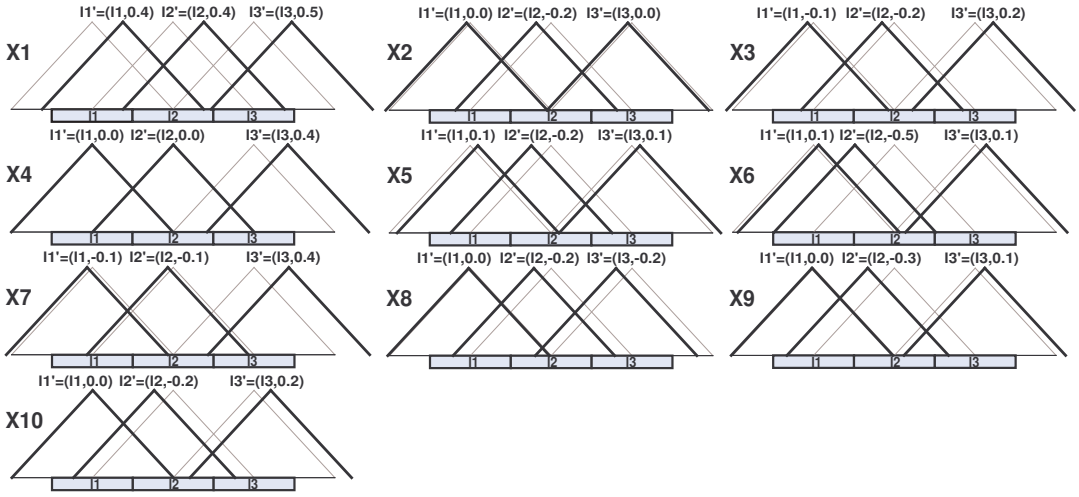

Fig. 10. MFs with/without lateral displacements (black/gray) and displacements of the MFs obtained by the proposed approach with 3 linguistic terms

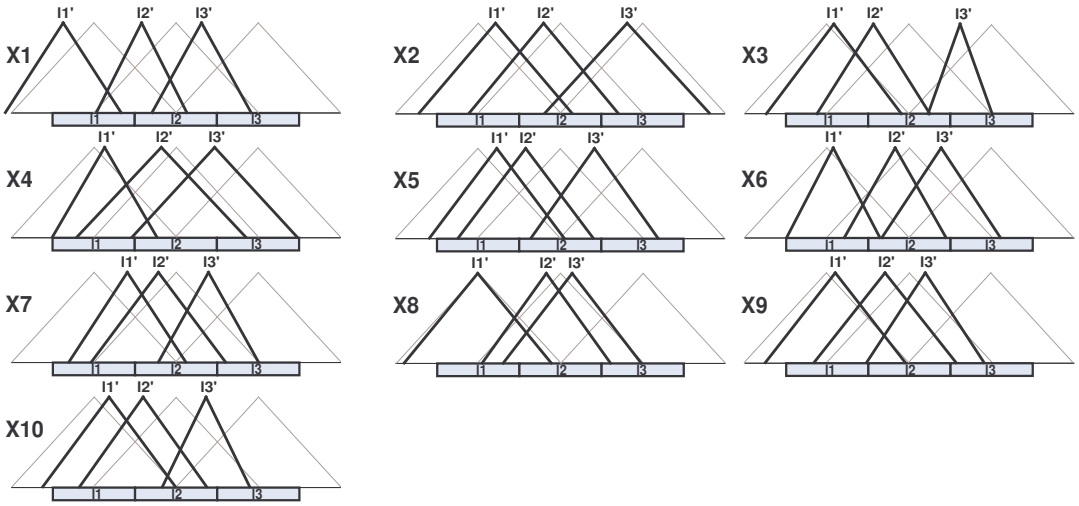

Fig. 11. MFs with/without displacements (black/gray) obtained by Hong et al.' approach with 3 linguistic terms

The number of fuzzy association rules obtained with 3 linguistic terms by the different approches is presented in Fig. 12 and Fig. 13. Fig. 12 depicts the relationship between the number of fuzzy association rules and the minimum support with 0.8 for the confidence threshold. In this figure we can highlight that the proposed approach extracts the best number of fuzzy association rules in 8 of the 9 values for the minimum support. On the other hand, Fig. 13 depicts the relationship between the number of fuzzy association rules and the confidence threshold with 0.2 for the minimum support. Analysing this figure we can highlight that, although the derived number of fuzzy association rules decreased along with the increase of the minimum confidence value, the proposed approach extracts more than twice as fuzzy association rules as remaining approaches with all the values of the confidence threshold.

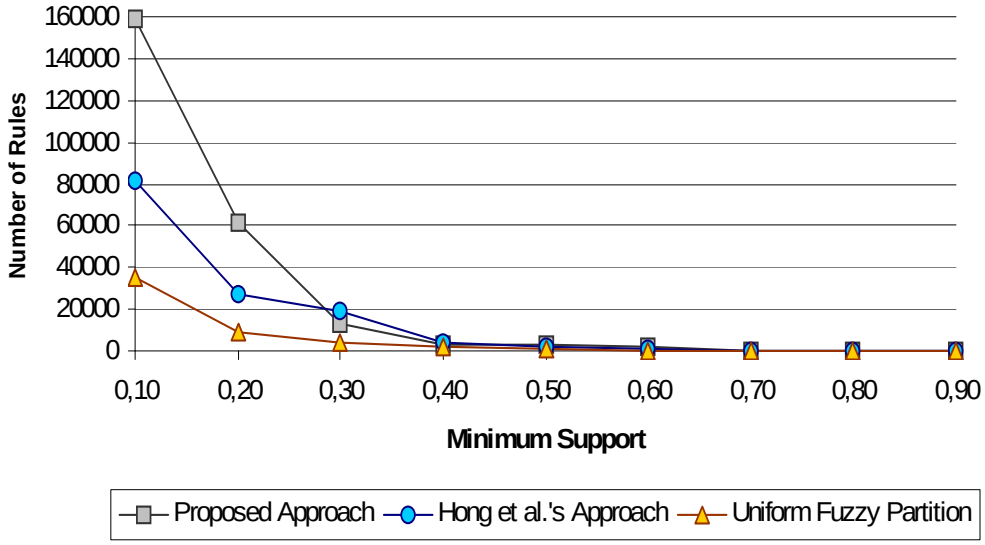

Fig. 12. Relationship between the number of rules and the minimum support

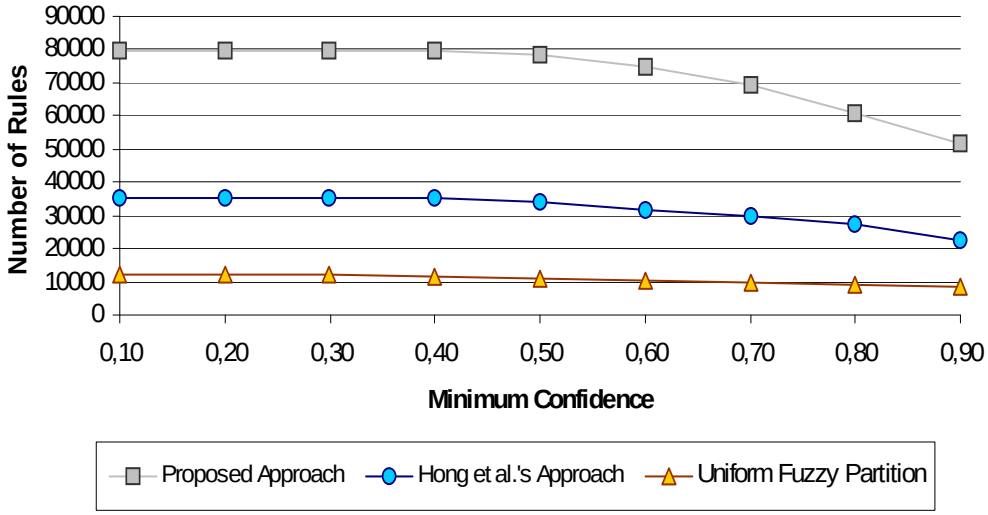

Fig. 13. Relationship between the number of rules and the confidence threshold

A crucial problem in association rule mining concerns the often huge number of frequent itemsets and interesting rules that can be found in a database. In this paper, we have considered the method presented in [14] as a first approach for mining fuzzy association rules. However, we could consider other approach which allows us to reduce the number of rules presented to the user. For example, we could use a method for mining multi-level fuzzy association rules [30], weighted association rules [33], etc.

In this section several experiments have been carried to analyse the fuzzy association rules obtained by the proposed approach. Fig. 14 shows the relationship between the number of fuzzy association rules derived by the final MFs and the minimum supports along with different minimum confidences. We can see that the number of rules decreases along with the increase of the minimum support values. Besides, the curves have similar shapes and the differences among them are small (mainly with minimum support values larger than 0.2). It means that the proposed method allows us to obtain interesting fuzzy association rules since most of the fuzzy association rules can satisfy the confidence threshold even with large values of minimum confidence.

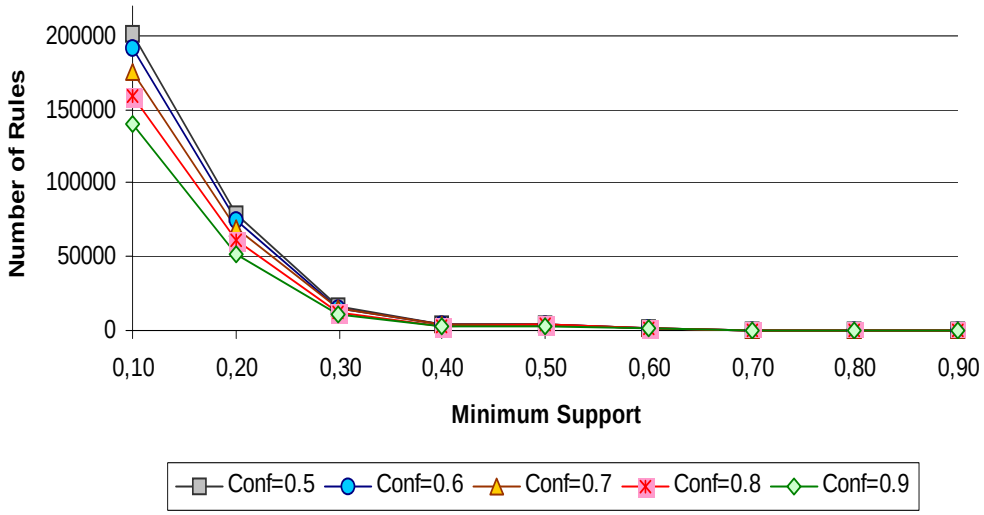

Fig. 14. Relationship between the number of fuzzy association rules and the minimum support along with different confidence thresholds

Fig. 15 shows the relationship between the number of association rules derived by the final MFs and the confidence threshold along with different minimum supports. We can see that the number of rules decreases slowly with the increase of the confidence threshold values. Notice that this figure shows clearer how most of the fuzzy association rules satisfy the confidence threshold when the confidence threshold value is increased. Besides, the curve with a large minimum support value are smoother than those with a small value, meaning that the confidence threshold value has a larger effect on the number of fuzzy association rules when smaller minimum support values are used.

Finally, an example of classic fuzzy association rule mined out with one uniform fuzzy partition and with the proposed approach is:

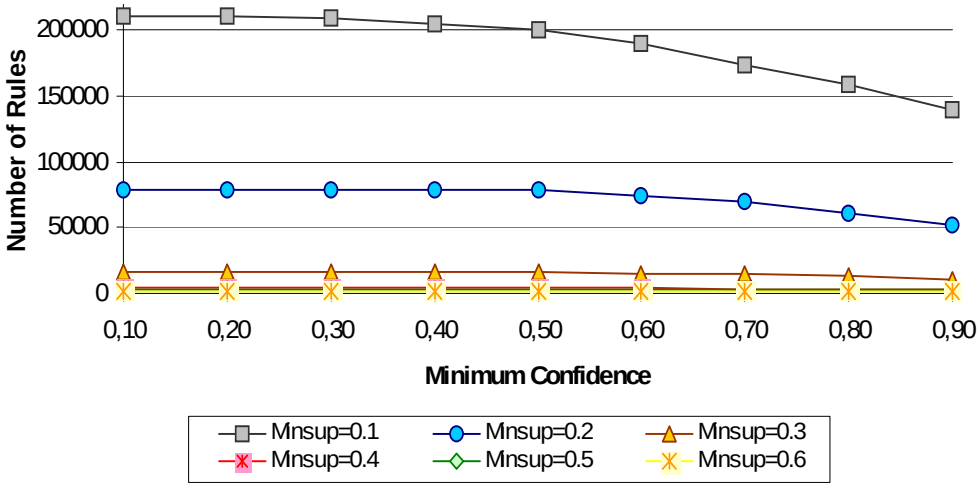

Fig. 15. Relationship between the number of fuzzy association rules and the confidence threshold along with different minimum supports

Classic Fuzzy Association Rule,
If **number of children** is Low and
hours head worked last week is Low
then **head's personal income** is Low.
Factor of confidence 0.87

Rule with 2-Tuples Fuzzy Linguistic Representation,
If **number of children** is (Low, -0.16) and
hours head worked last week is (Low, -0.06)
then **head's personal income** is (Low, 0.1)
Factor of confidence 0.99

This example shows how the proposed approach improves the confidence of the fuzzy association rules obtained with one uniform fuzzy partition. Furthermore, the interpretability of the rules is maintained in a high level since the original shapes of the initial MFs are not changed and the new ones are directly related to the initial ones by means of the 2-tuples linguistic representation.

5.4 Analysis of Complexity and Scalability

Several experiments have been carried to analyse the complexity and scalability of the proposed approach. All of the experiments were performed using a Pentium Centrino, 2.4 GHz CPU with 2Gb of memory and running Windows

XP. Fig. 16, 17 and 18 show the relationship between the runtime and the number of transactions, attributes and linguistic terms, respectively. It can be easily seen from these figures that the reduction of the search space provided by the 2-tuples linguistic representation allows the proposed approach to decrease its runtime regarding Hong et al's approach as we increase the size of problem. Moreover, the results plotted in these figures show that the proposed approach scales quite linearly for the database used in the experiments.

On the other hand, we can see how the proposed approach expend a reasonable time for the database used. However, an interesting further work could be the use of a parallel distributed implementation [32] or of a data reduction [4] to improve the scalability of the proposed approach.

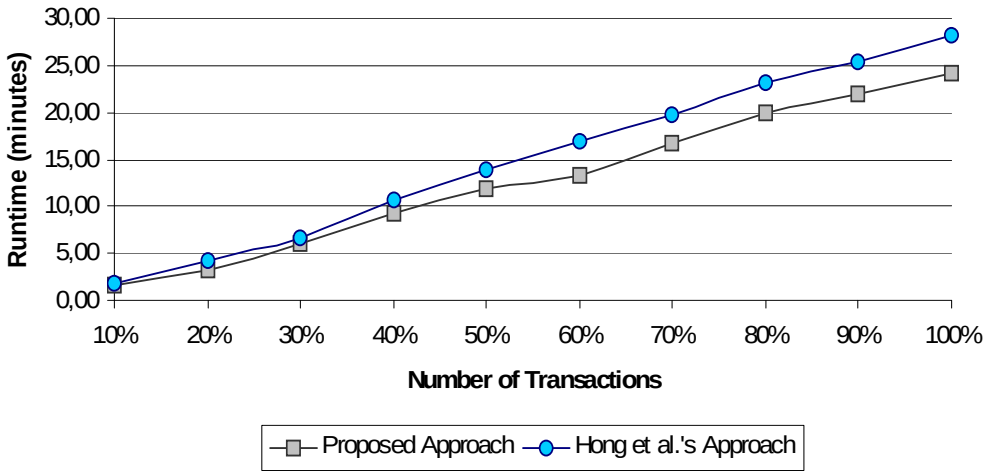

Fig. 16. Relationship between the runtime and the number of transactions with 10 attributes and 3 linguistic terms.

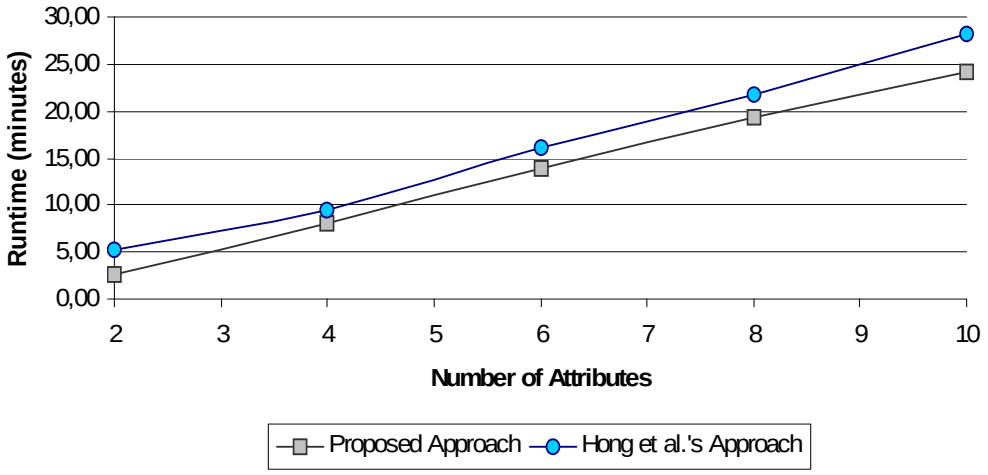

Fig. 17. Relationship between the runtime and the number of attributes with the 100% of transactions and 3 linguistic terms.

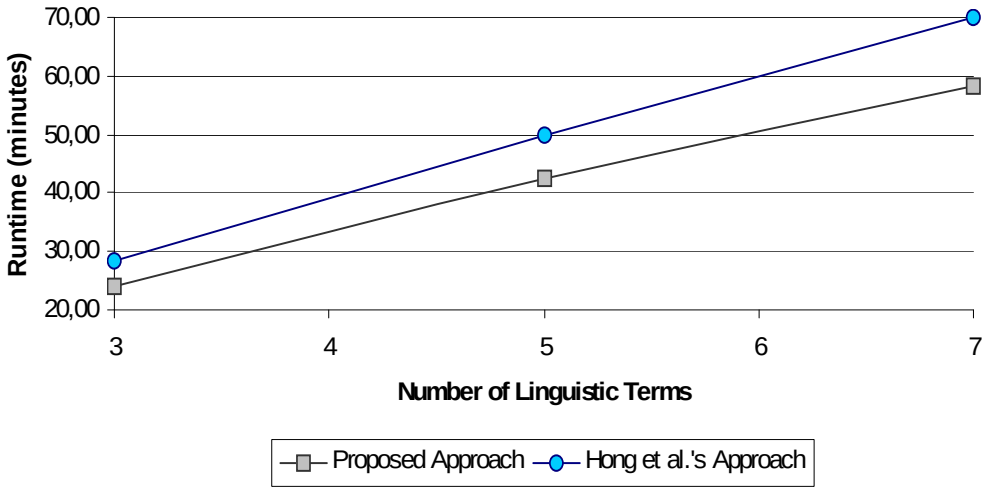

Fig. 18. Relationship between the runtime and the number of linguistic terms with the 100% of transactions and 10 attributes.

6 Conclusions

In this paper, a new rule representation scheme by using the 2-tuples linguistic representation model has been considered to extracting both MFs and fuzzy association rules from quantitative transactions. To do that, we have proposed a genetic learning process for getting the MFs together with a basic method to mine fuzzy association rules. Here, we present our conclusions and further considerations:

- The 2-tuples linguistic representation model allows an important reduction of the search space from the optimization point of view.
- The coverage ranges of the final MFs contain all the items possible quantities in the transactions since the 2-tuples linguistic representation maintains the original shapes of the MFs and restricts the lateral variation to a short interval, ensuring overlapping between two adjacent MFs.
- The learning scheme together with the 2-tuples linguistic representation model and the used fitness function offers a good mechanism to obtain MFs with a a good trade-off between fuzzy supports and suitability, allowing us to mine out a larger number of interesting fuzzy association rules.

7 Acknowledgment

This paper has been supported by the Spanish Ministry of Science and Technology under Project TIN2005-08386-C05-01 and the Andalusian Government under Project P05-TIC-00531.

References

[1] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in: SIGMOD, Washington D.C. (USA), 1993, pp. 207–216.

[2] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: International Conference on Very Large Data Bases, Santiago de Chile (Chile), 1994, pp. 487–499.

[3] R. Alcalá, J. Alcalá-Fdez, F. Herrera, A proposal for the genetic lateral tuning of linguistic fuzzy systems and its interaction with rule selection, IEEE Transactions on Fuzzy Systems 15 (4) (2007) 616–635.

[4] J. Cano, F. Herrera, M. Lozano, Evolutionary stratified training set selection for extracting classification rules with trade-off precision-interpretability, Data and Knowledge Engineering 60 (1) (2007) 90–108.

[5] O. Cordón, F. Gomide, F. Herrera, F. Hoffmann, L. Magdalena, Ten years of genetic fuzzy systems: Current framework and new trends, Fuzzy Sets and Systems 141 (1) (2004) 5–31.

[6] O. Cordón, F. Herrera, F. Hoffmann, L. Magdalena, GENETIC FUZZY SYSTEMS. Evolutionary tuning and learning of fuzzy knowledge bases. Advances in Fuzzy Systems - Applications and Theory, World Scientific, 2001.

[7] O. Cordón, F. Herrera, L. Magdalena, P. Villar, A genetic learning process for the scaling factors, granularity and contexts of the fuzzy rule-based system data base, Information Sciences 136 (2001) 85–107.

[8] O. Cordón, F. Herrera, P. Villar, Generating the knowledge base of a fuzzy rule-based system by the genetic learning of the data base, IEEE Trans. Fuzzy Syst. 9 (4) (2001) 667–674.

[9] M. Delgado, N. Marín, D. Sánchez, M. Vila, Fuzzy association rules: General model and applications, IEEE Transactions on Fuzzy Systems 11 (2) (2003) 214–225.

[10] D. Dubois, E. Hullermeier, H. Prade, A systematic approach to the assessment of fuzzy association rules, Data Mining and Knowledge Discovery 13 (2) (2006) 167–192.

[11] D. Dubois, H. Prade, T. Sudamp, On the representation, measurement, and discovery of fuzzy associations, IEEE Transactions on Fuzzy Systems 13 (2) (2005) 250–262.

[12] L. Eshelman, The chc adaptive search algorithm: How to have safe serach when engaging in nontraditional genetic recombination, in: G. Rawlin (ed.), Foundations of Genetic Algorithms, vol. 1, Morgan Kaufmann, 1991, pp. 265–283.

[13] L. Eshelman, J. Schaffer, Real-coded genetic algorithms and interval schemata, in: D. Whitley (ed.), Foundations of Genetic Algorithms, vol. 2, Morgan Kaufmann, 1993, pp. 187–202.

[14] A. Fu, M. Wong, S. Sze, W. Wong, W. Wong, W. Yu, Finding fuzzy sets for the mining of fuzzy association rules for numerical attributes, in: International symposium on intelligent data engineering and learning, Hong Kong, China, 1998, pp. 263–268.

[15] J. Han, M. Kamber, Data Mining: Concepts and Techniques. Second Edition, Morgan Kaufmann, San Fransisco, 2006.

[16] F. Herrera, Genetic fuzzy systems: Taxonomy, current research trends and prospects, Evolutionary Intelligence 1 (2008) 27–46.

[17] F. Herrera, L. Martínez, A 2-tuple fuzzy linguistic representation model for computing with words, IEEE Transactions on Fuzzy Systems 8 (6) (2000) 746–752.

[18] T. Hong, C. Chen, Y. Wu, Y. Lee, Using divide-and-conquer ga strategy in fuzzy data mining, in: IEEE International Symposium on Fuzzy Systems, Budapest, Hungary, 2004, pp. 116–121.

[19] T. Hong, C. Chen, Y. Wu, Y. Lee, A ga-based fuzzy mining approach to achieve a trade-off between number of rules and suitability of membership functions, Soft Computing 10 (11) (2006) 1091–1101.

[20] T. Hong, C. Kuo, S. Chi, Trade-off between time complexity and number of rules for fuzzy mining from quantitative data, International Journal Uncertain Fuzziness Knowledge-Based Systems 9 (5) (2001) 587–604.

[21] T. Hong, Y. Lee, An overview of mining fuzzy association rules, in: H. Bustince, F. Herrera, J. Montero (eds.), Studies in Fuzziness and Soft Computing, vol. 220, Springer Berlin/Heidelberg, 2008, pp. 397–410.

[22] E. Hullermeier, Y. Yi, In defense of fuzzy association analysis, IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics 37 (4) (2007) 1039–1043.

[23] H. Ishibuchi, T. Nakashima, M. Nii, Classification and Modeling with Linguistic Information Granules: Advanced Approaches to Linguistic Data Mining, Springer-Verlag, Berlin, 2004.

[24] H. Ishibuchi, T. Nakashima, T.Yamamoto, Fuzzy association rules for handling continuous attributes, in: IEEE International Symposium on Industrial Electronics Proceedings, Pusan, Korea, 2001, pp. 118–121.

[25] M. Kaya, Multi-objective genetic algorithm based approaches for mining optimized fuzzy association rules, Soft Computing 10 (7) (2006) 578–586.

[26] M. Kaya, R. Alhajj, Genetic algorithm based framework for mining fuzzy association rules, Fuzzy Sets and Systems 152 (3) (2005) 587–601.

[27] M. Kaya, R. Alhajj, Utilizing genetic algorithms to optimize membership functions for fuzzy weighted association rules mining, Applied Intelligence 24 (1) (2006) 7–15.

[28] C. Kuok, A. Fu, M. Wong, Mining fuzzy association rules in databases, SIGMOD Record 17 (1) (1998) 41–46.

[29] Y. Lee, T. Hong, W. Lin, Mining fuzzy association rules with multiple minimum supports using maximum constraints, in: KES, vol. 3214 of Lecture Notes in Computer Science, Springer-Verlag, 2004, pp. 1283–1290.

[30] Y. Lee, T. Hong, T. Wang, Multi-level fuzzy mining with multiple minimum supports, Expert Systems with Applications 34 (1) (2008) 459–468.

[31] M. Lozano, F. Herrera, N. Krasnogor, D. Molina, Real-coded memetic algorithms with crossover hill-climbing, Evolutionary Computation 12 (3) (2004) 273–302.

[32] Y. Nojima, I. Kuwajima, H. Ishibuchi, Data set subdivision for parallel distributed implementation of genetic fuzzy rule selection, in: International Conference on Fuzzy Systems, London, UK, 2007, pp. 23–26.

[33] J. Shu-Yue, E. Tsang, D. Yengg, S. Daming, Mining fuzzy association rules with weighted items, in: IEEE International Conference on Systems, Man and Cybernetics, Nashville, Tennessee, 2000, pp. 1906–1911.

[34] T. Sudkamp, Examples, counterexamples, and measuring fuzzy associations, Fuzzy Sets and Systems 149 (1) (2005) 57–71.

[35] W. Wang, S. Bridges, Genetic algorithm optimization of membership functions for mining fuzzy association rules, in: International Joint Conference on Information Systems, Fuzzy Theory and Technology Conference, Atlantic City, N.Y., 2000, pp. 1–4.

[36] C. Zhang, S. Zhang, Association Rule Mining: Models and Algorithms Series, Lecture Notes in Computer Science, LNAI 2307, Springer-Verlag, Berlin, 2002.