Journal of Computer Science and Cybernetics, V.32, N.4 (2016), 335–350
DOI 10.15625/1813-9663/32/4/9145

HEDGES ALGEBRAS AND FUZZY PARTITION PROBLEM FOR QUALITATIVE ATTRIBUTES

TRAN THAI SON[1], NGUYEN TUAN ANH[2]

[1] *Institute of Information Technology, Vietnam Academy of Science and Technology*
[2] *University of Information and Communication Technology, Thai Nguyen University*
[1] *ttson1955@gmail.com;* [2] *anhnt@ictu.edu.vn*

Abstract. There have been many approaches proposed to derive membership functions for mining fuzzy association rules using genetic algorithms (GAs) and show their advantages. However, these approaches show that the number of linguistic terms needs to be predefined. In this paper, the author proposed a new method to construct the membership functions (MFs) based on database. The theory of hedge algrebra was used to build the membership functions and GA is applied to optimize them. The experimental results demonstrate the benefits of this method.

Keywords. Fuzzy Association Rules; Data mining; Hedge algebras; Genetic algorithms; Membership functions

1. INTRODUCTION

Recently, mining fuzzy association rules, such as "If students have high academic results and are passionate about researching, they will find a good job.", has been the topic that is considered and developed. To be able to get fuzzy rules, in the first step, we partition each attribute domain of a database into fuzzy sets, characterized by membership functions (MFs). Then, each value (number) in the database will be converted to the corresponding set of the degree of membership. Numerical Database is converted into a fuzzy database ready for fuzzy mining in next steps. Traditionally, the fuzzy database is often assumed to be available, i.e, the MF sets are defined for each DB attribute. Such MF sets are usually determined experimentally and independently from database. In fact, it was shown that the approach may inversely affect the quality of the mining rules. Therefore, researchers pay attention to the construction of optimal sets MFs according to some predefined criteria by different algorithms of DB. However, there are very few studies of contructing MF sets for extracting association rules. The majority of them are related to the problems of automatic classification and regression [8, 9, 10, 11, 24]. In addition, despite its flexibility in approach, the use of L.Zadeh fuzzy set theory still has certain limitations. In this paper we propose the approach using Hedge Algebra [18, 19] to build MFs for extracting association rules. The advantages of this method will be demonstrated in analysis of experimental results on standard data of the Population of US in 1995.

© 2016 Vietnam Academy of Science & Technology

2. RELATED WORK

2.1. Association rule

Let $I = I_1, I_2, ..., I_m$ be a set of items. Let D, the task-relevant data, be a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called TID [21].

Definition 1. [22] An association rule has the form of $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$.

Two important measures of association rule are support(s) and confidence(c) defined in [22].

Definition 2. [22] The support of association rule $X \Rightarrow Y$ is the probability that $X \cup Y$ exists in a transaction in the database D.

$$\text{support}(X \Rightarrow Y) = P(X \cup Y) = \frac{n\,(X \cup Y)}{N}. \tag{1}$$

Definition 3. [22] The confidence of the association rule $X \Rightarrow Y$ is the probability that $X \cup Y$ exists given that a transaction contains X, i.e.

$$\text{confidence}\,(X \Rightarrow Y) \;=\; P\left(\frac{X}{Y}\right) \;=\; \frac{n\,(X \cup Y)}{n\,(Y)}, \tag{2}$$

where $n(X)$ - transaction count (including X), N - total of transaction database.

Mining the association rules is used to find all rules having the degree of support and confidence that is greater than that of min support and min confidence determined by the user.

In fuzzy association rules, the degree of support of a fuzzy range s_k belonging to x_i is defined as following [3, 4]:

$$FS\left(A_{s_1}^{x_1},\, A_{s_2}^{x_2}, .., A_{s_k}^{x_k}\right) = \frac{1}{N}\sum_{j=1}^{N}\prod_{i=1}^{k}\mu_{s_i}^{x_i}\left(d_j^{x_i}\right). \tag{3}$$

And the reliability of a fuzzy range $s_1, s_2, .., s_k$ of items $x_1, x_2, .., x_k$ respectively is

$$FS\left(A_{s_1}^{x_1},\, A_{s_2}^{x_2},\, \ldots,\, A_{k}^{x_k}\right) \;=\; \frac{1}{N}\sum_{j\,=\,1}^{N} min\left(\mu_{s_1}^{x_1}\left(d_j^{x_1}\right),\, \mu_{s_2}^{x_2}\left(d_j^{x_2}\right),\, \ldots,\, \mu_{s_k}^{x_k}\left(d_j^{x_k}\right)\right), \tag{4}$$

where x_i is i^{th} item, s_j is fuzzy range belonging to item i^{th}, N is the total of transactions in the database, $\mu_{\mathrm{s_k}}^{\mathrm{x_i}}(\mathrm{d_j^{x_i}})$ is the membership degree of the value at the i^{th} column, row j into the fuzzy set s_k.

2.2. Hedges algebras (HA)

Suppose that there is a set of linguistic values of any linguistic variable that includes:...< Very Negative < Negative < Little Negative < Zero < Little Positive < Positive < Very Positive <... These linguistic values appear in the language rules (LRB - Linguistic Rule Base) of approximately reasoning problems based on knowledge. Therefore, it is necessary to have a strict computing architecture which preserves order relation of the inherent linguistic values. Since then the semantic relationship of the linguistic value in the rules can be calculated.

HA [18, 19] is a mathematical structure which has the order of collection of linguistic items. The order relationship is defined by the semantics of the linguistic items from this collection. The quantitative semantics value of linguistic items through Semantically Quantifying Mappings SQMs allows performing a full description and showing rule set model and approximate inference process of inference in a logical and coherent way.

Definition 4. [19] Hedge algebras of the linguistic variable X can be represented as an algebraic structure which is the set of 5 components $AX = (X, G, C, H, \leq)$, where X is a set of items in X; $\leq$ denotes the natural semantic order relationships of the items in X; G $= \{c^-, c^+\}$, $c^- \leq c^+$ called the generating elements; $C = 0, W, 1$ is the set of constants, with the range $0 \leq c^- \leq W \leq c^+ \leq 1$, to indicate the elements that has the smallest, largest and neutral elements. The set of hedges $H = H^- \cup H^+$, with $H^- = \{h_j : 1 \leq j \leq q\}$ is the set of negative hedges, $H^+ = \{h_j : 1 \leq j \leq p\}$ is the positive hedges.

Apparently, the set X consists of linguistic items which have the order of the linguistic variable X and are generated by the impact of hedges $h \in H$ on the generating elements $c \in G$. The components in AX has some following properties:

- $\forall h \in H, x \in X : hx \leq$ or $hx \geq x$
- $\forall h \in H$, $hx = x$ then x is the permanent element. We have $H(x) = \{x | x \in C\}$
- $\forall h \in H^-$ then $hc^+ \leq c^+$, $hc^- \geq c^-$; $\forall h \in H^+$ then $hc^+ \geq c^+$ $hc^- \leq c^-$
- $h,\ k \in H^+$, $h \geq k$ if $hc^+ \geq kc^+$ (or $hc^- \leq kc^-$)
- $h,\ k \in H^-,\ h \geq k$ if $hc^+ \leq kc^+$ (or $hc^- \geq kc^-$)
- $h, k \in H$, $x \in X$, h is positive for k if $hkx < kx < x$ (or $x < kx < hkx$), h is negative for k if $x < kx < hkx$ (or $x < hkx < kx$)

From the above characteristics, we can define a function sgn as follows:

Definition 5. [25, 26] sgn : $X \rightarrow \{-1, 0, 1\}$. With $k, h \in H, c \in G, x \in X$

1) $\text{sgn}(c^+) = +1$ and $\text{sgn}(c^-) = -1$
2) $\{h \in H^+ | \text{sgn}(h) = +1\}$ and $\{h \in H^- | \text{sgn}(h) = -1\}$
3) $\text{sgn}(hc^+) = +\text{sgn}(c^+)$ if $hc^+ \geq c^+$ or $\text{sgn}(hc^-) = +\text{sgn}(c^-)$; if $hc^- c^-$ and $\text{sgn}(hc^+) = -\text{sgn}(c^+)$; if $hc^+ \leq c^+$ or $\text{sgn}(hc^-) = -\text{sgn}(c^-)$; if $hc^- \geq c^-$ or $\text{sgn}(hc) = \text{sgn}(h)\,\text{sgn}(c)$

4) $\operatorname{sgn}(khx) = +\operatorname{sgn}(hx)$ if k is positive for h ($\operatorname{sgn}(k, h) = +1$) and $\operatorname{sgn}(khx) = -\operatorname{sgn}(hx)$ if k is negative for h ($\operatorname{sgn}(k, h) = -1$)

5) $\operatorname{sgn}(khx) = 0$ if $khx = hx$.

Fuzzy measurement of a concept $x \in X$ is equal to the radius of the set $H(x)$, denoted as $fm(x)$ and can be calculated recursively from the fuzzy measurement of the generating elements $fm(c^-)$, $fm(c^+)$ and the fuzzy measurement of the hedges $\mu(h)$, $h \in H$, called the fuzzy parameters of X. $fm(x)$ is defined recursively as follows:

Definition 6. [20] $fm : X \to [0,\ 1]$, $x \in X$ where

1) $fm(C^-) + fm(C^+) = 1$ and $\sum_{h\in H} fm(hu) = fm(x)$, for all $x \in X$

2) $fm(x) = 0$, with $\forall x$, $H(x) = \{x\}$. Especially, $fm(0) = fm(W) = fm(1) = 0$

3) $\forall x, y \in X$, $\forall h \in H$, $\dfrac{fm(hx)}{fm(x)} = \dfrac{fm(hy)}{fm(y)}$ does not depend on specific elements and is called the fuzziness measure of h, denoted by $\mu(h)$.

With $x \in X$, $x = h_n, h_{n-1}, ..., h_1c, h_j \in H$, $c \in G$, we have the characteristics of $fm(x)$ and $\mu(h)$ as follows:

$$fm(hx) = \mu(h)\, fm(x), \forall x \in X, \tag{5}$$

$$\sum_{i=-q, i\neq 0}^{p} fm(h_i c) = fm(c), with\ C \in \left\{C^-, C^+\right\}, \tag{6}$$

$$\sum_{i=-q, i\neq 0}^{p} fm(h_i x) = fm(x), \tag{7}$$

$$fm(x) = fm(h_n h_{n-1} \dots\ h_1 c) = \mu(h_n)\, \mu(h_{n-1}) \dots \mu(h_1)\, fm(c), \tag{8}$$

$$\sum_{i=-1}^{-q} \mu(h_i) = \alpha,\ \text{ and } \sum_{i=1}^{p} \mu(h_i) = \beta, \text{ with } \alpha, \beta > 0 \,\text{and}\, \alpha + \beta = 1. \tag{9}$$

With advance $fm(c^-), fm(c^+)$ and $\mu(h), h \in H$ specifically, semantically quantifying value is recursively determined by the semantically quantifying mapping function v as follows.

Definition 7. [19] $v : X \to [0, 1]$

$$v(W) = fm(c^-),\ v(c^-) = \theta - \alpha fm(c^-) = \beta fm(c^-), \tag{10a}$$

$$v\left(c^+\right) = \theta + \alpha fm\left(c^+\right) = 1 - \beta fm\left(c^+\right), \tag{10b}$$

$$v\left(h_j x\right) = v\left(x\right) + sign\left(h_j x\right)\left\{\sum_{i=sign(j)}^{j} fm\left(h_j\right) - \omega\left(h_j x\right) fm\left(h_j x\right)\right\}, \tag{11a}$$

$$v\left(h_j x\right) = v\left(x\right) + sign\left(h_j x\right)\left\{\sum_{i=sign(j)}^{j} fm\left(h_j\right) - \omega\left(h_j x\right) fm\left(h_j x\right)\right\}, \tag{11b}$$

where $\omega\left(h_j x\right) = \frac{1}{2}\left[1 + \mathrm{sgn}\left(h_p, h_j\right)\left(\beta - \alpha\right)\right]$, $i \in [-q \wedge q] = [-q,\ p] \setminus \{0\}$.

The definition and properties of the HA can be found in [1].

Similarly it is also set $I_{(k)} = \cup_{l=1,\ldots,k} I_l$.

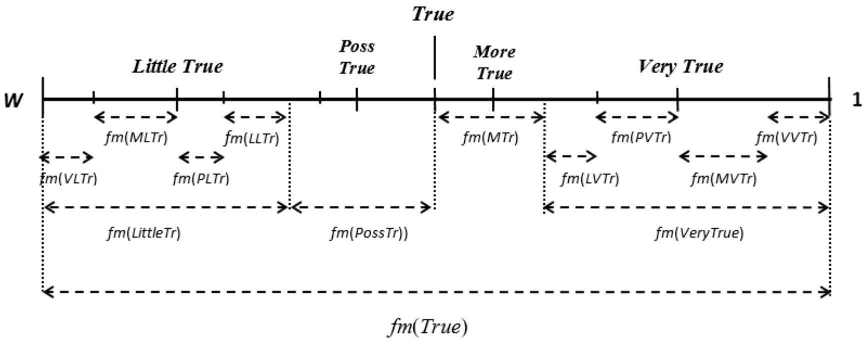

Figure 1. Fuzziness measure of TRUTH

In HA, the set (of elements) of the same capacity (length) will create a partition with domain-specific attributes. Each fuzzy interval corresponds to a fuzzy measure of an element.

These fuzzy intervals are arranged in the order that reflexes the "strength" of the semantics values (aka, quantitative semantics value). These results give us a simple method to establish a set of domain-specific partition MFs [26].

3. FUZZY PARTITIONS IN QUANTITATIVE ATTRIBUTES

3.1. Using fuzzy logic method for Partitions

There are several methods for partitioning domain values using fuzzy logic as follows:

- *Random partitioning:* In this method, a fixed number of sub-domains (partitions) is chosen (conventionally, 3) so that each sub-domain has the same length. This method is simple and effective when there is not enough information about the database. However, it does not take into account the distribution of data.

- *Partition clustering:* In this method, data is partitioned into clusters based on the proximity of a certain criteria between them. Popular algorithm is k-means clustering.

Normally, the number of clusters are chosen in advance and fixed (usually, 3). Unlike the previous one, this method takes into account the distribution of data. However, it is also very time-consuming.

- *Dynamically constrained partitioning:* In this method, data is partitioned into fuzzy domain under constraints on the membership functions to ensure a given criteria. According to [11], these criteria might be:

  1) The number of partions must be reasonably small.

  2) The membership functions are distinguished, for example, two MFs are not specific to the same linguistic label.

  3) Each MF is normalized ie if it reaches the value 1 at least at one point of the range.

  4) The range of values is covered entirely by the corresponding fuzzy domain and at least one MF received value $x > 0$ at any point in the value domain.

3.2. Using HA method for Partitions

In section 2.2., we presented some basic concepts of HA. These values refer to different fuzzy sets of the partition for each numeric dataset attribute. It is easy to calculate the value of membership degree based on fuzzy membership functions as in [17]. Having defined the domain of the item (the normalization on the interval [0,1]), any value between two quantification of semantics value of consecutive fuzzy intervals or coincides with a quantification of semantics value will create domain-specific partition fuzzy intervals.

Thus, the distance from x_{ij} to two quantitative semantics value can be taken as the membership degree x_{ij} in fuzzy sets (if x_{ij} is equal to a quantitative semantics value, only one membership degree): The smaller the distance is, the higher the membership degree will be (maximum value 1 when being a full member). In [14], the authors built MFs from quantification of semantics values. Membership functions are triangular with vertices having coordinates $(v(x_i), 1)$. The coordinates of the three vertices of a triangle are: $(v(x_i), 1), (v(x_{i-1}), 0), (v(x_{i+1}), 0)$ with $v(x_{i-1}), v(x_i), v(x_{i+1})$ being consecutive quantification of semantics values (in Figure 2).

It can be seen that in Figure 2, the two methods are used. Assuming E is an arbitrary point in the determining area of property I_i. With the first method, the distance between $Ev(x_2)$ and $Ev(x_3)$ will be used to determine the membership degree of E to fuzzy sets represented by triangle membership functions $v(x_1)Bv(x_3)$ and $v(x_2)Cv(x_4)$ (based on the standardization so that membership degree is always in the range [0,1]. Using the second method, EG and EF are the membership degree of E to the two fuzzy sets. EG is parallel to $v(x_2)B$, so $\dfrac{EG}{v(x_2)B} = \dfrac{Ev(x_3)}{v(x_2)v(x_3)}$. Similarly, $\dfrac{EF}{v(x_3)C} = \dfrac{v(x_2)E}{v(x_2)v(x_3)}$. Furthermore, $v(x_2)B = v(x_3)C = 1$, so $\dfrac{EF}{EG} = \dfrac{Ev(x_2)}{Ev(x_3)}$. Therefore, it is equivalent to use the membership degree in these two methods. Especially, determining the membership degree based on HA is also reasonable and close to the natural experience.

Using the HA to build membership functions has some advantages:

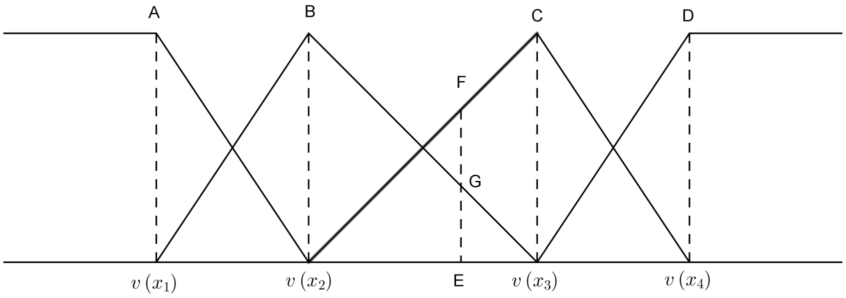

Figure 2. Build MFs using HA

a) Since constructing HA is based on the semantics that human beings feel, it is sensible that the membership functions bring good reflection on the semantics of the fuzzy set it represents.

b) Coverage of MFs as well (always identified domain covered). If need optimal suitability of MFs, we just need to optimize performance and usability overlap. Optimization problem according to the parameters of the overlapping HA and usefulness can be solved by a genetic algorithm (GA).

c) The number of optimal parameters is limited. When changing the initial parameters of the HA, it is easy to rebuild the MFs. Therefore, this method is simple and reasonable. To prove the effectiveness of this approach, we tested on the standard sample FAM95 data as presented in the following section (FAM95 database conducted by the Bureau of the Census for the Bureau of Labor Statisticsin 1995).

4. GENETIC MINING PROCESS

In [7], the authors used methods to evaluate the membership functions. MFs' prominence was assessed by three factors: The overlap factor represents the overlap ratio of the MFs, The coverage factor represents the coverage ratio of the MFs, and usage factor. In this paper, we use GA algorithm as in [5] to optimize the MFs according to the above criteria.

The components required to design this GA include:

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

4.1. CHC genetic model

In this section, the population-based selection approach is considered using the CHC genetic model [18] to perform the adequate global search. In the genetic CHC model, N

parents and their offsprings compete to select the best N individuals for the next population. The CHC uses the "incest prevention" mechanism and a process of restarting to create the diversity in the population (instead of mutation).

The incest prevention mechanism is used when provoking the crossover operator. While considering a real coding scheme, each gene is transformed using a Gray Code with a fixed number of bits per gene. This number is determined by experts. The crossover threshold L is calculated as following:

$$L = (\#Genes\_BITSGENE)/4.0$$

where $\#Genes$ is the number of genes, BITSGENE is the number of bits per gene. In the original CHC model, if there is no new individuals in the popuplation in a generation, L is decreased by one. To make L indepedent of $\#Genes$ and BITSGENE, we decrease L by $\varphi\%$ (φ is set by users, normally 10%). The algorithm restarts when $L < 0$.

4.2. MFs codification and initial gene pool

In this paper, we use structured HA as follows:
$AT = (X, G,\ H, \leq)$,
$G = \{C^- = \{low\} \cup C^+ = \{high\}\}$,
$H = \{H^- = \{Little\} \cup H^+ = \{Very\}\}$,
$\alpha = \mu\,(Little) = 1 - \mu\,(Very)$, $\beta = 1 - \alpha$,
$w = fm\,(low) = 1 - fm\,(high)$.

We performed a chromosome, a real number array size $n \times 2$ (where n is the number of items, 2 corresponds to the parameters α and w in each HA):

$$(\alpha_1, \ldots, \alpha_n, w_1, \ldots,\ w_n)\,. \tag{12}$$

For each pair (α_i, w_i) are parameters of a HA.

Initialize population consisting of N chromosomes: based on the experience of the value α and w will receive a random value in the interval [0.2 to 0.8].

4.3. Chromosomal evaluation

To evaluate the chromosome, we use fitness function to define in [7]. The fitness function of a chromosome C_q is defined as follows:

$$\text{fitness}\,(Cq) = \frac{\sum_{x \in L_1} \text{fuzzy\_support}\,(x)}{\text{suitability}\,(C_q)} \tag{13}$$

with

- L_1 is frequent 1-itemset using MFs in C_q,
- fuzzy_support(x): fuzzy support of 1-itemset x is calculated from the transaction database,
- suitability (C_q): appropriate coverage of MF in C_q.

The suitability of the set of chromosomes of MFs in C_q is defined as follows:

$$\text{suitability}(C_q) = \sum_{k=1}^{n} [\text{overlap\_factor}(C_{qk}) + \text{coverage\_factor}(C_{qk})] \tag{14}$$

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 presents the ratio of overlying between the MFs (denoted R_i and R_j, $i < j$) for an item I_k in the chromosome C_q. This ratio is defined as the length of the overlap region from the right span of R_i and the left span of R_j. If this length is larger than the minimum of the spans, it is said to be "little redundant". Thus, the overlap factor is defined as

$$\text{Overlap\_factor}\left(C_{qk}\right) = \sum_{k=1}^{m} \sum_{j=i+1}^{m} \left[\max\left(\frac{\text{overlap}\left(R_i, R_j\right)}{\min\left(\text{span}R_{R_i}, \text{span}L_{R_j},\right)}, 1 \right) - 1 \right] \tag{15}$$

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.

In this case, spanR_{R_i} and spanR_{R_j} have the same size as the MFs are shifted in the uniform partition (and thus maintain the original MFs' shapes).

The coverage factor is the ratio of coverage of the MFs for an item I_k in the chromosome and is defined as the coverage range divided by the maximum number of items in the transaction. The larger the coverage ratio is, the better the derived MFs will be.

The coverage factor of the MFs for an item I_k in the chromosome C_q is defined as:

$$\text{Coverage\_factor}\left(C_{qk}\right) = \frac{1}{\dfrac{\text{rang}\left(R_1, \ldots, R_m\right)}{\max\left(I_k\right)}}, \tag{16}$$

where range$(R_1, R_2, ..., R_m)$ - coverage range of the MFs and max(I_k) - maximum quantity of I_k in the transactions.

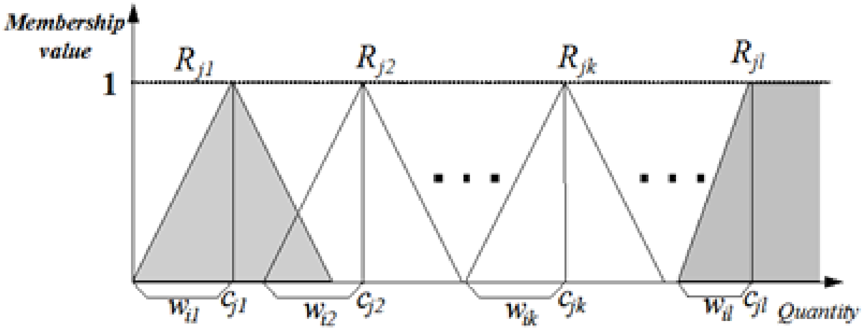

Figure 3. MFs of Item I_j

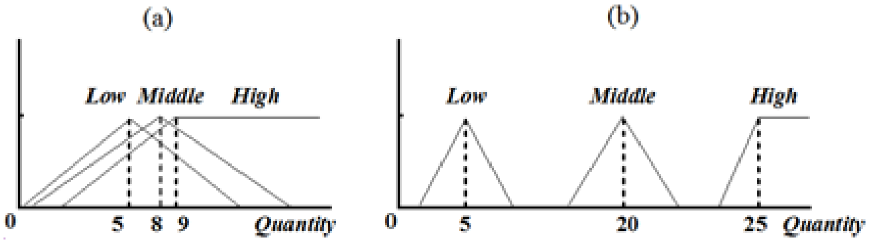

Figure 4. Two bad kinds of membership functions

These suitable factors in the fitness functions may help to prevent two unfavorable types of MFs (redundant or separated): the overlap factor - for the redundant MFs, the coverage factor - for the separated MFs.

Recently, we also use the concept of strong fuzzy partition to build collective MF [12]. This concept is defined as follows: MFs make a strong fuzzy partition domain if they cover property values and at any point on each identified domain, the total's membership degree on this point to all of the partitions MFs valued at 1. Strong fuzzy partition MFs also generate relatively good distribution.

With this measure, it is possible to use genetic algorithms to get optimal MFs.

4.4. Crossover operator

The crossover operators create a cooperation within genetic model generating the convergence by pressuring on the offsprings. Particularly, the Parent Centric BLX (PCBLX) operator based on the $BLX-\alpha$ is considered here.

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

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

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

5. PROPOSED MINING ALGORITHM

In this section, a proposed algorithm for mining MFs and association rules is described in detail.

Input: Transaction data with T quantities, n-item set (each item has m predefined linguistic terms), support threshold min_Support, confidence threshold min_confidence, population size N.

Output: Set of association rules with associated set of MFs.

Phase 1: Learning the MFs.

Step 1: Initial population generation with N chromosomes.

Chromosome representation of the form $(\alpha_1, \ldots, \alpha_n, w_1, \ldots, w_n)$, with each pair of (α_i, w_i) is a HA.

Step 2: Population evaluation. For each chromosome:

The MFs is a string encryption as described in Section 4.2. Based on the HA has been in step 1, to build the MFs for a property in the database as described in Section 3.2.

Step 3: Calculate the fitness function for each chromosome in the population as follows:

- For each transaction D_i, $i = 1$ to T, and for each item I_j, $j = 1$ to n, convert the quantitative value $v_j^{(i)}$ $(D_i = (v_1^{(i)}), ..., v_n^{(i)}))$ into a fuzzy set $f_j^{(i)}$ represented as

$$f_j^{(i)} = \left(\frac{f_{j1}^{(i)}}{R_{j1}} + \frac{f_{j2}^{(i)}}{R_{j2}} + \ldots + \frac{f_{jl}^{(i)}}{R_{jl}} \right)$$

  using the corresponding MFs, where R_{jk} is the k-th linguistic term of item I_j, $f_{jk}^{(i)}$ is $v_j^{(i)}$'s membership value in R_{jk} region, and m is the number of linguistic terms for I_j.

- For each linguistic term R_{jk}, assess its count on the transactions:

$$\text{count}_{jk} = \sum_{i=1}^{n} f_j^{(i)} \tag{17}$$

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

$$L_1 = \left\{ R_{jk} | \text{count}_{jk} \geq \alpha, 1 \leq j \leq m, 1 \leq k \leq |I_j| \right\}. \tag{18}$$

- 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

$$\text{fitness}\left(C_q\right) = \frac{\sum_{x \in L_1} \text{fuzzy\_support}\left(x\right)}{\text{suitability}\left(C_q\right)}. \tag{19}$$

Step 4: Threshold value (L) initialization.

Step 5: Next population generation:

- Shuffle the population.

- Select parents two by two.

- Evaluate new individuals.

- Join the parents with their offspring and select the best N individuals to put into the next population.

Step 6: In case the best chromosome does not change or there is no new individual in the population, $L = L - (L_{\text{initial}} * 0.1)$.
Step 7: If $L < 0$, restart the population.
Step 8: If the maximum number of evaluations is not reached, go to Step 5.
Phase 2: Basic method for mining association rules.
Step 9: The set of the best MFs is then applied in mining fuzzy association rules from the given quantitative database using the algorithm proposed in [23].

6. EXPERIMENTAL RESULTS

The source of the data is taken from FAM95 database, conducted by the Bureau of the Census for the Bureau of Labor Statistics in 1995. We selected 10 attribute numbers that include: age of the 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 for weight and March supplement income and tax.

Table 1. Results obtained in the genetic process

	Proposed Approach				Hong et als Approach				Uniform Fuzzy Partition			
Sup	Fit	Fsup	Suit	#1I	Fit	Fsup	Suit	#1I	Fit	Fsup	Suit	#1I
0.2	0.98	9.83	10	22	0.53	10.22	19.27	22	0.94	9.43	10	21
0.5	0.79	7.87	10	10	0.38	7.95	20.63	12	0.46	4.57	10	7
0.7	0.66	6.62	10	8	0.2	3.96	19.54	5	0.24	2.36	10	3
0.9	0.09	0.94	10	1	0.06	0.9	15.01	1	0	0	10	0

In these experiments, the proposed approach with one uniform fuzzy partition is compared to the Hong et al.s work [7]. For our method, each item is divided into 5 fuzzy domains with the corresponding labels in the HA. They are $\{0, C^-, W, C^+, 1\}$.

We performed a genetic algorithm with the following parameters: 50 individuals, 10,000 evaluations, 0 bits per gene for the Gray codification, 0.6 as crossover probability.

The results are shown in Table 1, with the total level of support, with 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.

The results from the table 1 show that:

- With the minsupp value of 20, the large 1-itemsets calculated by the HA and by Hong et al. have the same values which are better than the others.
- The Hong method brings better result only with the minsupp value of 0.5. However, the HA approach wins the others.

Herrera et al. also conducted experiments on the same data using a method called 2-tuples. This method shows better large 1-itemsets (22, 15, 10, 1, accordingly). However, the received MF sets (after the GA calculating) are quite poor (fig. 6.: MFs for 10 attributes with minimum support value of 20%).

It is shown that, in these situations, the received MFs sets all have a pair of overlapping MFs, which does not satisfy the overlap criteria. These results prove that the fuzzy domain

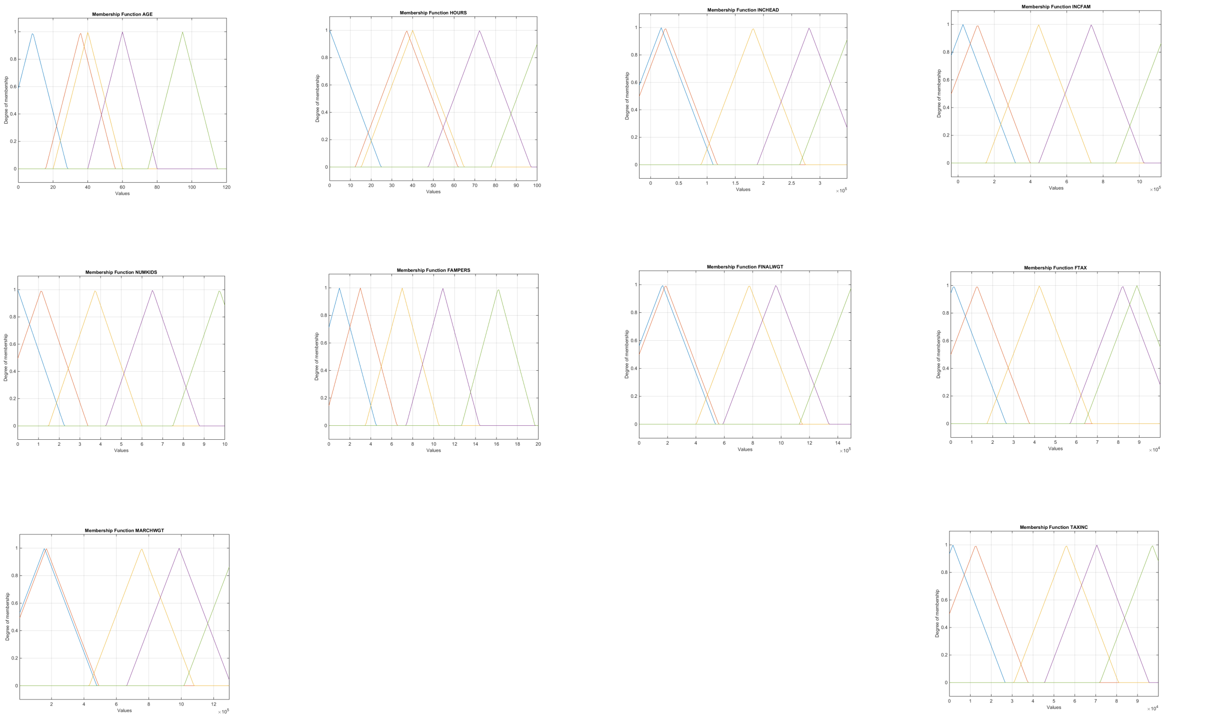

Figure 5. MFs obtained by Herrera et al.' approach with 5 linguistic terms

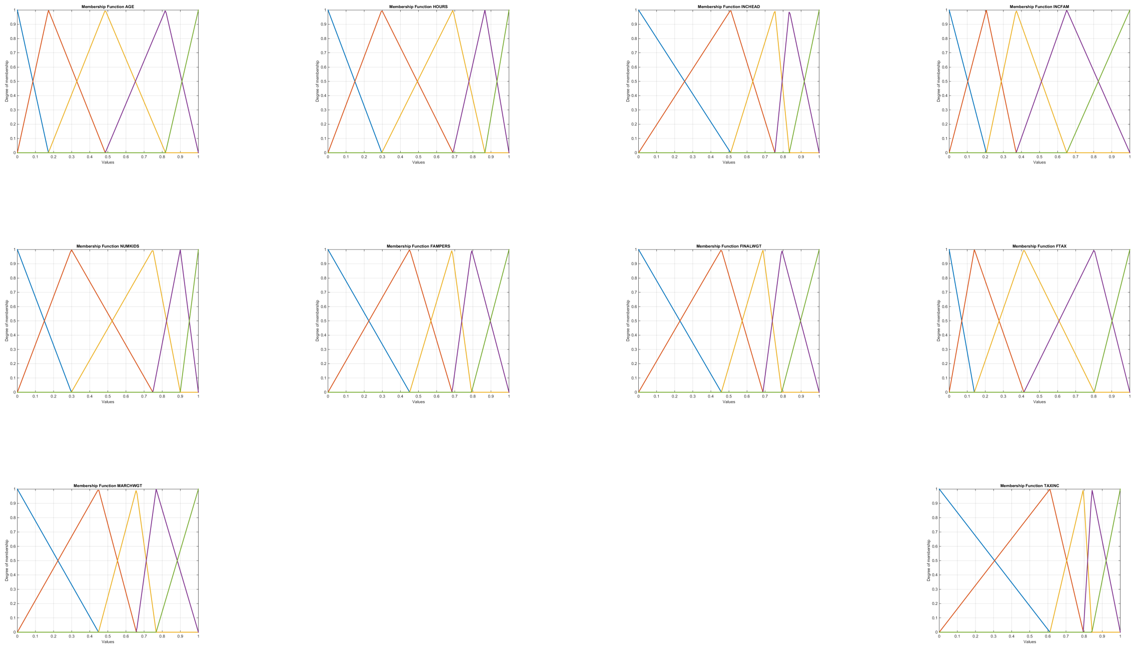

Figure 6. MFs displacements of the MFs with 5 linguistic terms

partitioned by this method is not good enough (ex. In this situation, its more reasonable to get 4 partitions, which reduce the amount language labels to 4 (instead of 5)). Selecting the MFs with fixed amount of items as well as determining the number of items are important issue since the final results are dependent on the amount of the MF for each item. We will return to this problem in the next paper with algorithms for optimizing the quantity and parameter of the attribute MFs in order to obtain the best result in data mining through the concept of multi domain particle fuzzy partitioning. Figure 6. introduces the illustration of MFs calculated by HA approach. The triangles representing the MFs make up a strong partitioning.

7. CONCLUSIONS

Hedge algebra, a new and quite effective tool, can be used instead of the fuzzy theory in many cases due to the order structure on the set of natural language elements. This paper demonstrates this trend through the method of building the membership function to split fuzzy item sets in the fuzzy data mining problem. This is an important step that has not still been much invested and researched. In order to meet the requirements of optimization problem for both the quantity and the MF parameters mentioned above, the expansion of hedge algebra (not only for 5 terms) will not only solve data minings problem better but also promote Hedge Algebras strength. Moreover, the number of classes can be easily increased and still ensure a strong partitions to divide the identified domain of items based on this method.

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under Grant No. 102.01-2017.06

REFERENCES

[1] N.C. Ho, Wechler W., Hedge algebra: An algebraic approach to structures of sets of linguistic truth values, Fuzzy Sets and Systems, vol. 35, pp. 281–293, 1990.

[2] L. Eshelman, The CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination, in: G. Rawlin (ed.), Foundations of Genetic Algorithms, 265–283, 1991.

[3] C. M. Kuok, A. Fu, and M. H. Wong, Mining fuzzy association rules in databases, ACM Sigmod Record, vol. 27, no. 1, pp. 41–46, 1998.

[4] A. Gyenesei, A fuzzy approach for mining quantitative association rules, Acta Cybern., vol. 15, no. 2, pp. 305–320, 2001.

[5] Herrera, Martinez, Learning the Membership Function Contexts for Mining Fuzzy Association Rules by Using Genetic Algorithms, Fuzzy Set and System, 905-921, 2009.

[6] Li-Xing Wang and J.M.Mendel, Generating Fuzzy Rules by Lerning from Examples, IEEE Trans. SMC, 1, 1992.

[7] C. Chen, T. Hong, Vincent S. T. and L. Chen, Multi-objective genetic-fuzzy data mining. International Journal of Innovative Computing, Information and Control, 8, 2012.

[8] M.J. Gacto, R. Alcal, F. Herrera, Interpretability of linguistic fuzzy rule-basedsystems: An overview of interpretability measures, Information Sciences, 8, 2011.

[9] M. Antonelli, P. Ducange, F. Marcelloni, Genetic Training Instance Selection in Multiobjective Evolutionary Fuzzy Systems: A coevolutionary Approach, IEEE Trans. on Fuzzy Systems, 20, 276–290, 2012.

[10] Alcala-Fdez, Jesús and Alcala, Rafael and Herrera, Francisco, A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning, IEEE Transactions on Fuzzy Systems, 19, 5, 857–872, 2011.

[11] P.Pulkkinen and H.Koivisto, A Dynamically Constrained Multiobjective Genetic Fuzzy System for Regression Problems, IEEE Tran. on Fuzzy Systems, 18, 857–872, 2010.

[12] Corrado Mencar, Marco Lucarelli, Ciro Castiello, Anna M. Fanelli, Design of Strong Fuzzy Partitions from Cuts, Conference of the European Society for Fuzzy Logic and Technology, 2013.

[13] Tanaka, H, Uejima, S, and Asia, K., Linear regression analysis with Fuzzy model, IEEE Trans. Systems.Man.Cybernet, 12, 903–07, 1982.

[14] N. C. Ho, T.T. Son, D.L. Long, Hedge Algebras approach to fuzzy classification, Journal of Computer Science and Cybernetics, 25, 53–68, 2009.

[15] N. C. Ho, T. T. Son, and P. D. Phong, Modeling of a semantics core of linguistic terms based on an extension of hedge algebra semantics and its application, Knowledge-Based Systems, vol. 67, pp. 244–262, 2014.

[16] N. C. Ho, W. Pedrycz, D. T. Long et al., A genetic design of linguistic terms for fuzzy rule based classifiers, International Journal of Approximate Reasoning, vol. 54, no. 1, pp. 1–21, 2012.

[17] N. C. Ho, T. T. Son, T. D. Khang et al., Fuzziness Measure, Quantified Sematic Mapping and Interpolative Method of Approximate Reasoning in Medical Expert Systems, Journal of Computer Science and Cybernetics, vol. 18, no. 3, pp. 237–252, 2002.

[18] N.C. Ho and Wechler, Wolfgang, Hedge algebras: an algebraic approach to structure of sets of linguistic truth values, Fuzzy sets and systems, 35, 3, 281–293, 1990.

[19] N.C. Ho, Wechler, Wolfgang, Extended hedge algebras and their application to fuzzy logic, Fuzzy sets and systems, 52, 3, 259–281, 1992.

[20] N. C. Ho, and N. V. Long, Fuzziness measure on complete hedge algebras and quantifying semantics of terms in linear hedge algebras, Fuzzy Sets and Systems, vol. 158, no. 4, pp. 452–471, 2007.

[21] Jiawei Han, Data Mining: Concepts and Techniques: University of Illinois at Urbana-Champaign, Micheline Kamber, 2012.

[22] R. Agrawal, T. Imielinski, A. Swami, Fast algorithms for mining association rules, The International Conference on Very Large Database, 487–499, 1994.

[23] D. L. Olson, and D. Delen, "Advanced data mining techniques", Springer Science & Business Media, 2008.

[24] C.-H. Chen, T.-P. Hong, Y.-C. Lee et al., Finding Active Membership Functions for Genetic-Fuzzy Data Mining, International Journal of Information Technology & Decision Making, vol. 14, no. 06, pp. 1215–1242, 2015.

[25] N. C. Ho, T. D. Khang, H. V. Nam et al., Hedge algebras, linguistic-value logic and their application to fuzzy reasoning, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 7, no. 04, pp. 347–361, 1999.

[26] N. C. Ho, V. N. Lan, and L. X. Viet, Quantifying hedge algebras, interpolate reasoning method and its application to some problems of fuzzy control, WSEAS Transactions on Computers, vol. 5, no. 11, pp. 2519–2529, 2006.

Received on November 11 - 2017
Revised on July 01 - 2017