Journal of Computer Science and Cybernetics, V.34, N.1 (2018), 63–75
DOI 10.15625/1813-9663/34/1/10797

PARTITION FUZZY DOMAIN WITH MULTI-GRANULARITY REPRESENTATION OF DATA BASED ON HEDGE ALGEBRA APPROACH

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*

Abstract. This paper presents methods of dividing quantitative attributes into fuzzy domains with multi-granularity representation of data based on hedge algebra approach. According to this approach, more information is expressed from general to specific knowledge by explored association rules. As a result, this method brings a better response than the one using usual single-granularity representation of data. Furthermore, it meets the demand of the authors as the number of exploring rules is higher.

Keywords. Fuzzy association rule, algebra approach, multi-granularity, Data mining, membership functions

1. INTRODUCTION

In terms of exploring knowledge in the studies, the problem of determining of fuzzy domain of data is quantitative attributes are more and more significantly attracted. This is a considerably initial step for the whole process of information processing for most of later data mining problems, such as association rule mining, classification, identification, regression [2, 4, 3, 10, 14]. If we have a reasonable fuzzy partition, the knowledge discovered will better reflect the hidden rules in the information store. Vice versa, if there is no proper fuzzy partition at first, the knowledge which we explore may be subjective, imposing and not exactly. This is not a simple problem. First, it primarily relates to the perception of the individual and depends on the context. For example, in the attribute domain "distance", it is not easy to determine when it is called "far" or "relatively close". Moreover, fuzzy division much depends on the input data that we get. Some studies have hypotheses about the probability distribution function of the data or other hypotheses. However, the data is variable, assumptions are not always true and the amount of information is enormous. Therefore, it requires reliable but not too complicated methods to process information in acceptable time.

2. THE PROBLEM OF DIVIDING A DETERMINED FUZZY DOMAIN

It can be expressed that the problem of dividing the fuzzy domain is able to determine the quantitative attributes of an input data set. Particularly, if there exists a specified domain of an attribute (only quantitative attributes are considered), typically a numeric and continuous value, then our duty will be the division of the attribute domain into sets (discrete or intersecting) so that they can be processed in the next steps. Moreover, it is necessary to have this partition because the large amount of input information will be meaningless if we solve each record separately. As a result, it is impossible to derive hidden rules in the data since these rules show the relationship between the large number of attributes in the input data. The division may be discrete, but the general trend is to divide into well-defined or vague domains as it is more suitable. For example, with the attribute "distance", discrete division may be [0, 50km] as "near"; [51km, 100km] is "average"; [100km, 200km] is "far", but so the distance between 50km and 51km is very close to each other but they belong to two different distance labels, so this is not very reasonable. With fuzzy division, we consider the labels "near", "medium", "far" as fuzzy sets, where any value x of the value domain of the attribute "distance" will be converted into sets of the dependent degrees of "near" (x), $\mu_{medium}(x)$, $\mu_{far}(x)$. We will handle them on the dependent degree of x on fuzzy sets instead of directly dealing with values x . At that time, the handling would be more costly but obviously much more flexible.

There are some methods for dividing determined fuzzy domain:

- *Randomly divided*: In this method, we choose a fixed number of domains to divide (usually divided into three fuzzy domains with membership functions of isosceles triangles, the same width of the bottom). This method is simple and is probably better when we have no other information, but obviously it does not meet the diversity of the data.

- *Divided by fuzzy clustering (unsupervised learning)*: Use clustering algorithms, such as k-mean, to clump data into fuzzy sets. This method takes into account the diversity of data distribution, but we have to take many times when running this algorithm type.

- *Division by dynamic constraints [14]*: In this method, the data is divided into fuzzy domains according to the constraints defined on the membership functions to ensure some criteria such as the number of fuzzy domains is suitable; MFs are quite distinguished and MFs (must cover well the value domain) must cover good domain value of attributes (at least one MF receives a value of $\beta > 0$ at any point in the value domain).

Specifically ([1, 6, 9]), assuming R_1, R_2,...,R_k are membership functions which divide fuzzy domain of the attribute I. To make it simple, let R_i $(i = 1, ..., k)$ be uniform isosceles triangles (Figure 1), then the criteria for overlapping and coverage can be considered in the following formulas

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

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.

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

$$\text{Coverage\_factor}\,(C_{q_k}) = \frac{1}{\dfrac{\text{range}\,(R_1, \ldots, R_m)}{\max\,(I_k)}} \tag{2}$$

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

The goal of fuzzy partition is to have the set MF so that the overlap is minimal and the coverage is maximized (while satisfying other criteria, such as at least one MF taking the value $\beta > 0$ at any point on the value domain mentioned above).

Recently, the concept of strong fuzzy partition was used to construct the set MF [10, 16]. The concept is defined as follows: the set of MFs makes a strong fuzzy partition if they cover the domain of the attribute value and at any point on the specified domain, the total of fuzzy degrees of this point to all MFs in the partition gain the value of 1. Strong fuzzy partitioning also created MFs which are relatively well-distributed.

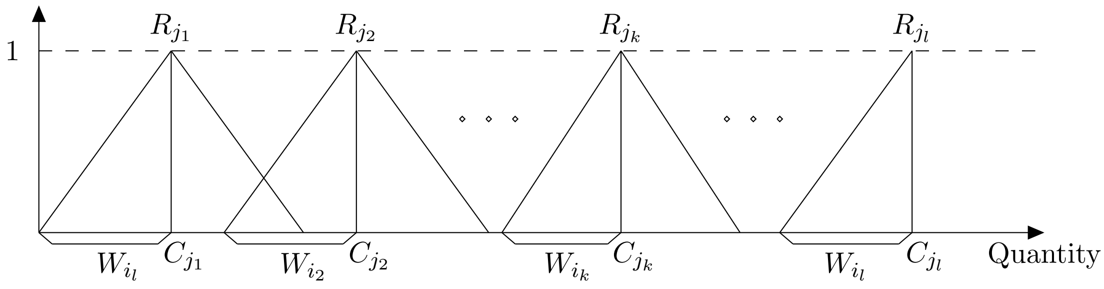


Figure 1. Membership functions of Item I_j

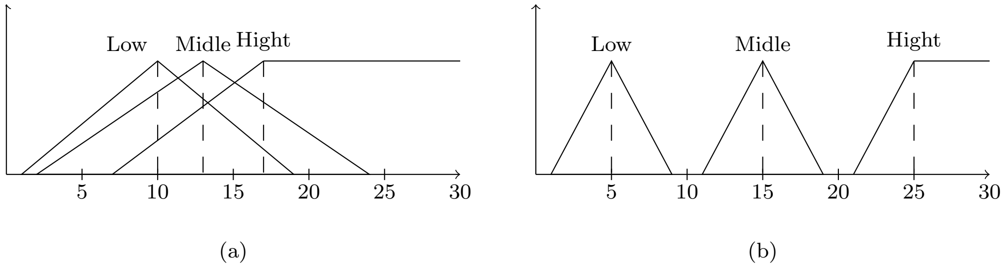


Figure 2. Two bad kinds of membership functions

With a good overlap_factor, we can exclude or limit the case (a) of Figure 2, when overlapping functions are far and not very specific. With good *coverge_factor*, it is possible to limit the case like (b) on Figure 2, when there is more space on the specified domain, not on any fuzzy set (fuzzy degree is 0). Go deeper into the field of knowledge mining problems, there will be other additional measures to optimize the sets of MFs such as the rule set constructed from MFs that will give the smallest classification error in the classification problem [4] or the minimum squares error is smallest in the regression problem [14]. In this paper, we focus on the association rule mining problem, so the additional measure, the $usage\_factor$, is the measure of the total of support degree of large 1-Itemsets. Remember that with the association rule $X \to Y$ (with support degree greater than minsup), the XY itemset is a large one. Then, any subset of a large itemset is also a large itemset. In particular, every subset with an attribute of XY must be a large itemset. Therefore, with a high level of support degree, it is hoped that we will receive many association rules. Although it is not sure like a consideration of all large itemsets, in return, the processing time will be less because only the frequent 1-Itemsets are considered. With such measurements, it is possible to use genetic algorithms to obtain optimal set MF, the balance between good system level and computational time are taken into account.

Partition of linguistic domain value based on hedge algebra's approach:

In the paper [15], we presented a method of partitioning the attribute value domain according to the hedge algebra's approach and demonstrate some advantages of this method with an illustrative example. In this approach, the MF sets are constructed from the quantitative linguistic values of the hedge algebra corresponding to the value domain of each attribute, namely triangles that represent the (dependent) membership functions of a fuzzy set with a vertex with the $coordinates((x_i), 1)$, the remaining two vertices are located on the domain value, with the corresponding coordinates $(v(x_{i-1}), 0), (v(x_{i+1})), 0)$, where $v(x_{i-1}), v(x_i), v(x_{i+1})$ are 3 consecutive quantitative linguistic values (see Figure 3).

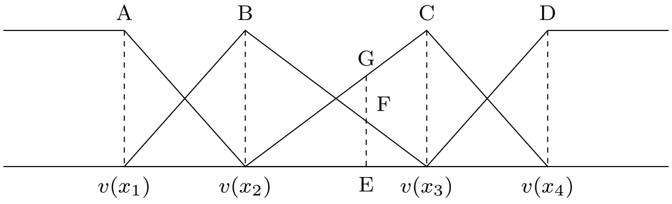


Figure 3. Building MF based on the HA's approach

The way to construct the membership functions or equivalent ones, the fuzzy sets that divide the domain value of the attribute according to the approach of the hedge algebra has the following advantages:

a) Because the construction of the hedge algebra is based on the sense that human beings feel, it is sensible that the membership functions built are quite reflective of the semantics of the fuzzy set it represents.

b) These MFs create a strong fuzzy partition as the above definition. It is easy to see that the cover of the membership functions is good (always covering the specified region domain value). Then, it can be seen that if we need to optimize the suitability of MFs, only optimizing the overlap and usability needed to be used. The optimization problem of the parameters of the hedge algebra according to the overlap and usefulness can be solved by a genetic algorithm (GA).

c) The parameters to be managed during construction are few (one for each parameter, the quantitative linguistic value), when changing the initial parameters of the hedge algebra, it is easy for MFs to be available and MF is maintained in terms of overlapping measures the same as the old ones. Therefore, this method is simple and reasonable.

3. SINGLE-GRANULAR AND MULTI-GRANULAR PRESENTATION OF DATA

The method of fuzzy domain partition according to the above approach of hedge algebra has the advantages as noted above, but there are still limitations related to the semantics of the data. According to the theory of the hedge algebra, the MFs created as above are based on a partition of the elements that have the same length. That means that the association rules that we explore only include the elements having the same length and that reduces the meaning of the explored rules. For example, the rules like ⟨If "very young" and "hard working" then "good future"⟩ and ⟨if "young" and "rather hard" then "good future"⟩ are two rules which are impossible to be simultaneously appeared in the exploration rule set because "young" and "very young" are two fuzzy labels of different lengths. If we do not care much about data semantics, merely dividing the domain that is almost machine-like (as most methods according to the past fuzzy set approach), the method in [7, 8] is pretty good. However, if the semantics of the data is taken into account, it is extremely important to have good knowledge in combining association rule - we must take a deeper approach. It is possible to construct semantic fuzzy spaces [11] to form partitions of different length elements, but this is not so standard since the generated partitions are not unique. It is also possible to use the extended hedge algebra with supplementary hedge h_0 [12] to construct a partition with different length elements. However, in this paper we have chosen an approach based on data representation of multi-granular structure.

3.1. Representation

Representation of data according to the multi-granular structure lies at the root of the problem of the Granular Computing (GrC), concept which has been a strong development trend in the past decade. The idea of GrC is that information is split into packets (granules) for processing. This division makes it not only easier to handle, but also helps us to better understand the information world because distributed packets are generalized. The information we receive can be divided into different ways, giving different views of the real world. Obviously, the more different perspectives on information we receive, the more knowledge we have about the problem of interest. That is why it is necessary to have a multi-granular representation for the data.

3.2. The reason why the multi-granular representation for the data should be used in mining associated rules

Ideally, the use of multi-granular representation, as noted, gives us a more diversified view of input information ("An advantage offered by a granular structure is the multilevel understanding and representation" [17]). The use of multi-granular representation helps us have a general overview as well as details in which we need. For example, in [5] the authors present an example of solving the problem of classifying elements of the Cone-Torus dataset. At level 1, the data is grouped into two-dimensional sets (by the Conditional Fuzzy C-Means Algorithm: CFCM), each dimension is separated by three fuzzy sets "low", "medium", "high". At the second level, in each dimension, data is further divided into fuzzy sets. For example, in context data clusters $x =$ "low" and $y =$ "low", data continues to be clustered (also by CFCM algorithm) into clusters by fuzzy sets $x =$ "is less than or equal to 1.1" and $y =$ "is greater than or equal to 3.7", $y =$ "is less than or equal to 1.0", $y =$ "about 2.6" and $y =$ " About 4.5 inches or more". Thanks to the fuzzy divisions at these two levels, the authors have come up with the rule set to classify data including general rules (e.g. ⟨IF x is LOW and y is LOW THEN P(class = 1) = 0.53, P(class = 2) = 0.38, P(class = 3) = 0.09, P(class = 3) = 0.29 ⟩) along with detailed rules (⟨ IF x is about 1.1 or less and y is about 2.6 THEN P(class = 1) = 0.31, P(class = 2) = 0.38, P(class = 3) = 0.01⟩). This system, according to the authors, has a high rate of classification and interpretability. In summary, the use of multi-granular representations gives us a high degree of general and well-defined knowledge that improves the performance of the method.

For fuzzy set theory (according to L.Zadeh), one of the limitations of methods of using multi-granular representations is that sometimes the selection of nonlinear functions is not easy since there are few reasons for defining membership functions of different levels and the relationship between them. Mostly, this determination is conducted only by experience, and in the above example we can also feel it. Simultaneously, carrying out calculations at different levels of data will entail complexity that costs much more in terms of time and memory. Even in recent studies [4], in the fuzzy rule-building application of the regression problem, the authors also use only single granularity presentation approach. In particular, using the evolutionary algorithm to construct the fuzzy rule set on the basis of optimizing fuzzy partition MF sets determines the properties of both the fuzzy domain division for each attribute and the other criteria mentioned above. Although the algorithm (performs) in [4] is better than existing ones as the number of fuzzy sets used to divide the domain attribute is not pre-predetermining but about semantics, it still does not allow the construction of different general and detailed rules in the same fuzzy rule system. On the contrary, with the hedge algebra, it is easy to identify fuzzy measurements at different levels of multi-granularity representation as it lies at the construction of the hedge algebra. In the hedge algebra's theory, it is only necessary to determine once the fuzzy measure values of the generating elements and the hedges, then we can determine the fuzzy range of all the elements based on the determined calculating formulas no matter how long this element is (i.e., how much this element is in the multi-granularity representation system). Decentralization, one of the main ways that GrC uses, is the way the hedge algebra is built. According to the theory of the hedge algebra, each of the element x of length k can be subdivided into elements $h_i x$ (where h_i is the hedge of hedge algebra that is being considered) with length $k + 1$. It can be said that the hedge algebra is a very suitable tool for multi-granularity computing. The example presented later will further clarify that.

3.3. MFs Codification and Initial Gene Pool

In this paper, we use structured HA as follows:
$AT = (X, G,\ H, \leq)$, $G = \{C^- = \{Low\} \cup C^+ = \{High\}\}$,
$H = \{H^- = \{Little\} \cup H^+ = \{Very\}\}$.

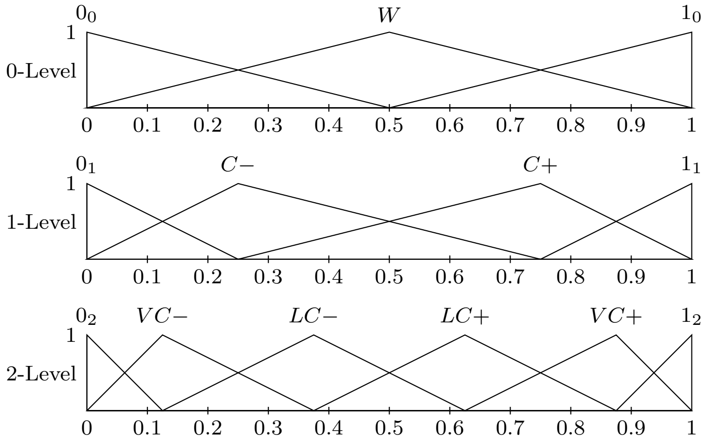


Figure 4. Building MF based on Multi-granular representation for an attribute

$\alpha = \mu\,(Little) = 1 - \mu\,(Very)$, $\beta = 1 - \alpha$, $w = fm\,(Low) = 1 - fm\,(High)$.

We performed a chromosome, a real number array size $n \times 2$ (where n is the number of items, 2 corresponds to the parameter α and w in each HA): $\{(\alpha_1, w_1), (\alpha_2, w_2), ..., (\alpha_n, w_n)\}$. 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].

Example: with $\alpha = 0.5$, $w = 0.5$, MFs is built as shown in Figure 4. Similarly, each attribute in the database will be built the MFs, as shown in Figure 4.

4. PROPOSED MINING ALGORITHM

In this section, our approach used partition fuzzy domain with multi-granularity representation of data, a proposed algorithm for mining MFs and association rules is described in detail.

Input: Transaction database 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 its associated set of MFs.

Phase 1: Learning the MFs.

In this paper, we use a multi-granularity approach. Each attribute in the database will be built by MFs, as shown in Figure 4. The MFs is a string encryption as described in Section 3.3. Using the algorithm in [15], we obtain a set of MFs to use for Phase 2.

Phase 2: Mining fuzzy association rules.

The set of the best MFs is then applied in mining fuzzy association rules from the given transaction database using the algorithm proposed in [13].

5. EXPERIMENTAL RESULTS

In this part, we present the experimental results of the proposed method for a particular database. The source of the data is taken from the FAM95 database, conducted by the Bureau of Statistics 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 of personal income, family income, taxable income for head, federal tax for head, final sampling weight for weight and March supplement income and tax [1, 6, 9].

Table 1. Relationship between the number of itemset and the minimum support (%)

	Min support (%)						
	20	30	40	50	60	70	80
1-itemset	59	50	38	29	26	22	17
2-itemset	974	675	465	371	285	187	78
3-itemset	8890	4806	3111	2660	2518	772	150
4-itemset	50242	20719	13095	11890	4708	1774	167
5-itemset	187379	57461	36432	34995	9506	2528	167

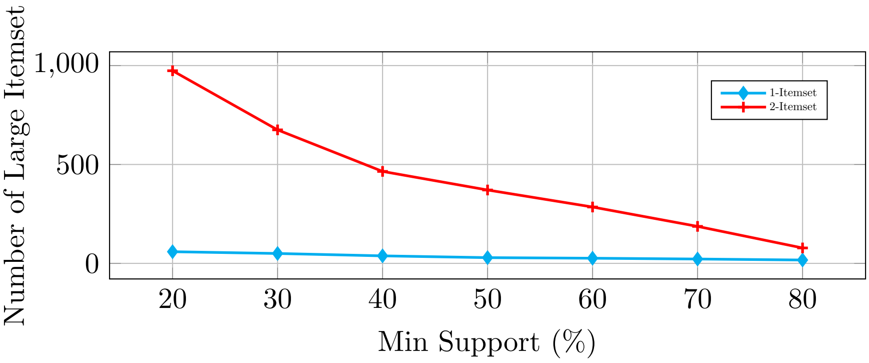


Figure 5. Relationship between the number of Large itemset and the minimum support

The results compared with other methods are listed in the below Table 2: Herrera's method proposed in [1], the method of using HA and sign-granularity was proposed in [20]. Here, (listing properties that use comparative form: overlay, overlap as the table of the previous paper), and methods for comparison are performed through single-particle representation. As given in the introduction, there hasn't been results regarding the fuzzy association rule mining using multinomial manifests due to the complexity of the experiment. (The latest article [18] only mentions an experiment that uses the multi-granularity representation of regression problems). It can be seen that multi-granularity representation will bring better results. In addition, as discussed above, in terms of semantics, using multi-granularity representation will give us rules with different linguistic labels, for example (e.g., 2 fuzzy rules whose linguistic elements have the length of 1, 2). In order to have similar rules, based on the above methods, we must divide each of the above attributes into at least nine fuzzy sets. We also tested Herrera's method with such partition; although it increases in terms of the index (Table 2), it is still poor in terms of suggested method (Fig. 5). It should be emphasized that, with our method, the computation involved in multi-granularity representation significantly increases in complexity as well as in time, while the results are far better.

Table 2. Relationship between large 1-itemsets and minimum support (%) with 9 linguistic terms

	Min support (%)							
	20	30	40	50	60	70	80	90
Proposed Approach	54	46	35	27	23	14	12	5
The method proposed in [15]	21	17	13	8	7	6	3	1
Herrera et al's Approach	25	21	15	10	5	3	2	0

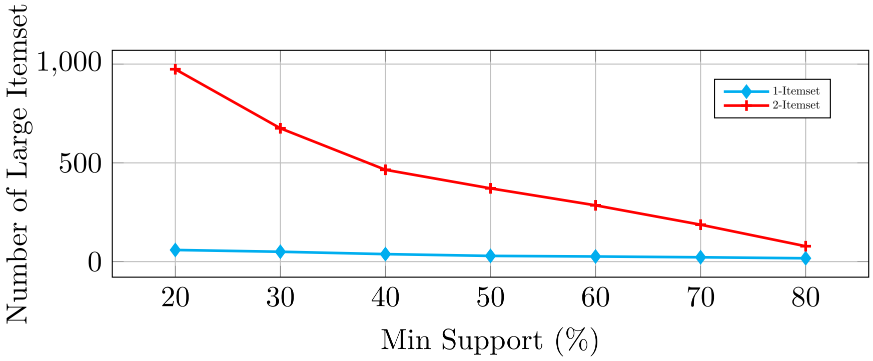


Figure 6. A two-degree-of-freedom manipulator (pan-tilt) with a camera on a wheeled mobile robot

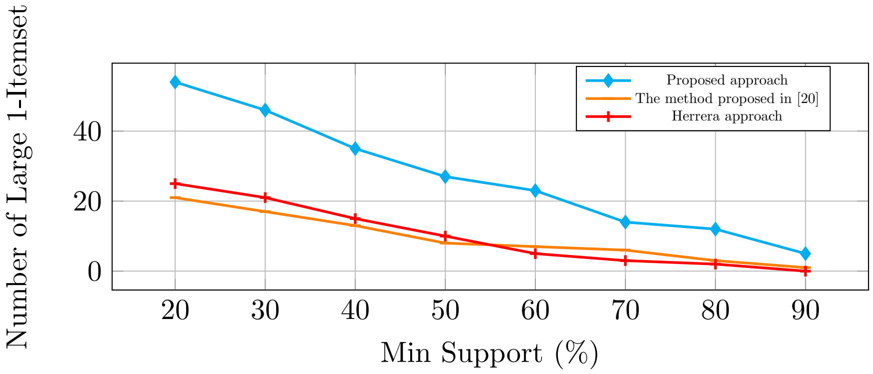
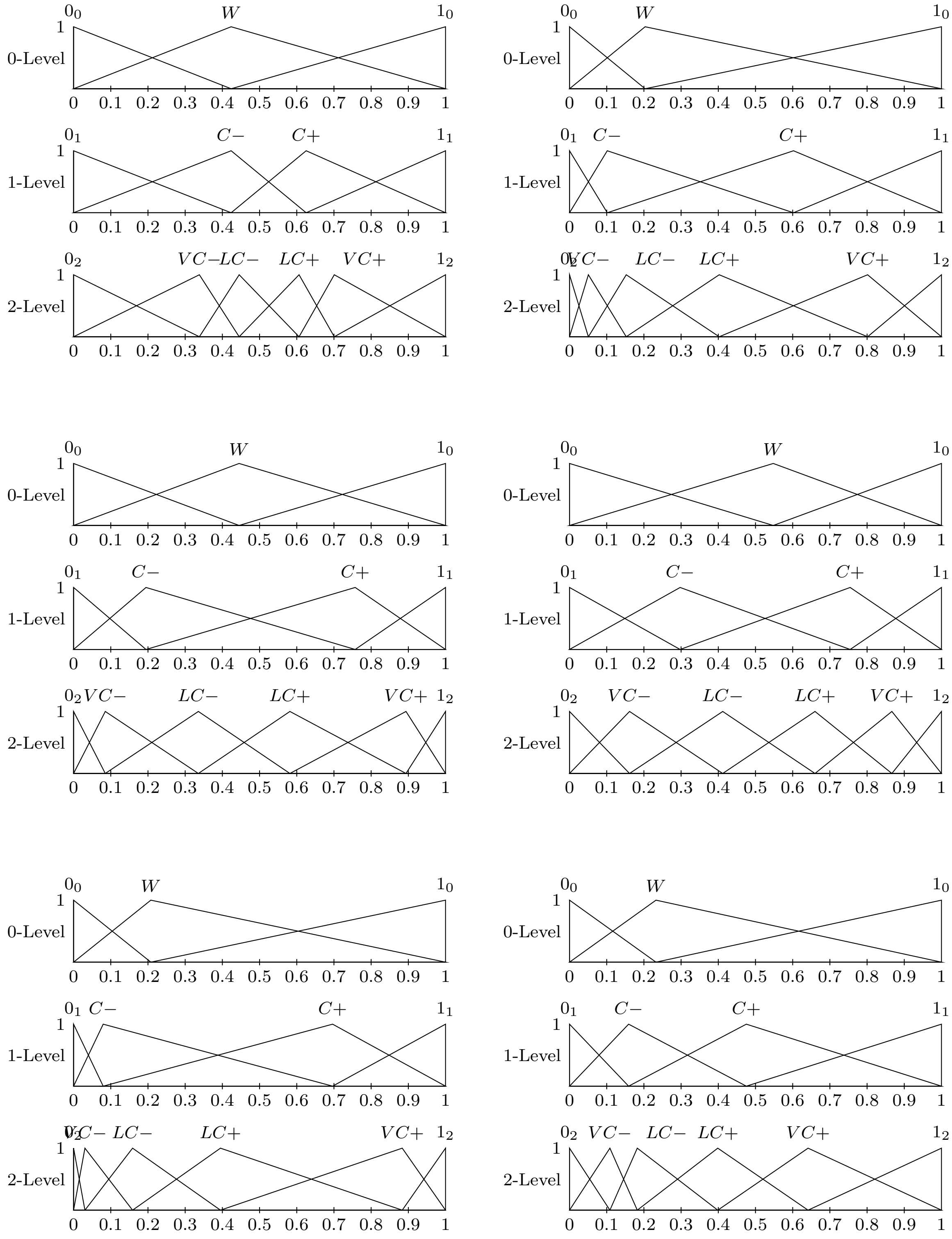


Figure 7. Relationship between the number of Large 1-itemset and the minimum support


0_0
W
1_0
1
0-Level
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
0_1
$C-$
$C+$
1_1
1
1-Level
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
0_2
$VC-$
$LC-$
$LC+$
$VC+$
1_2
1
2-Level
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

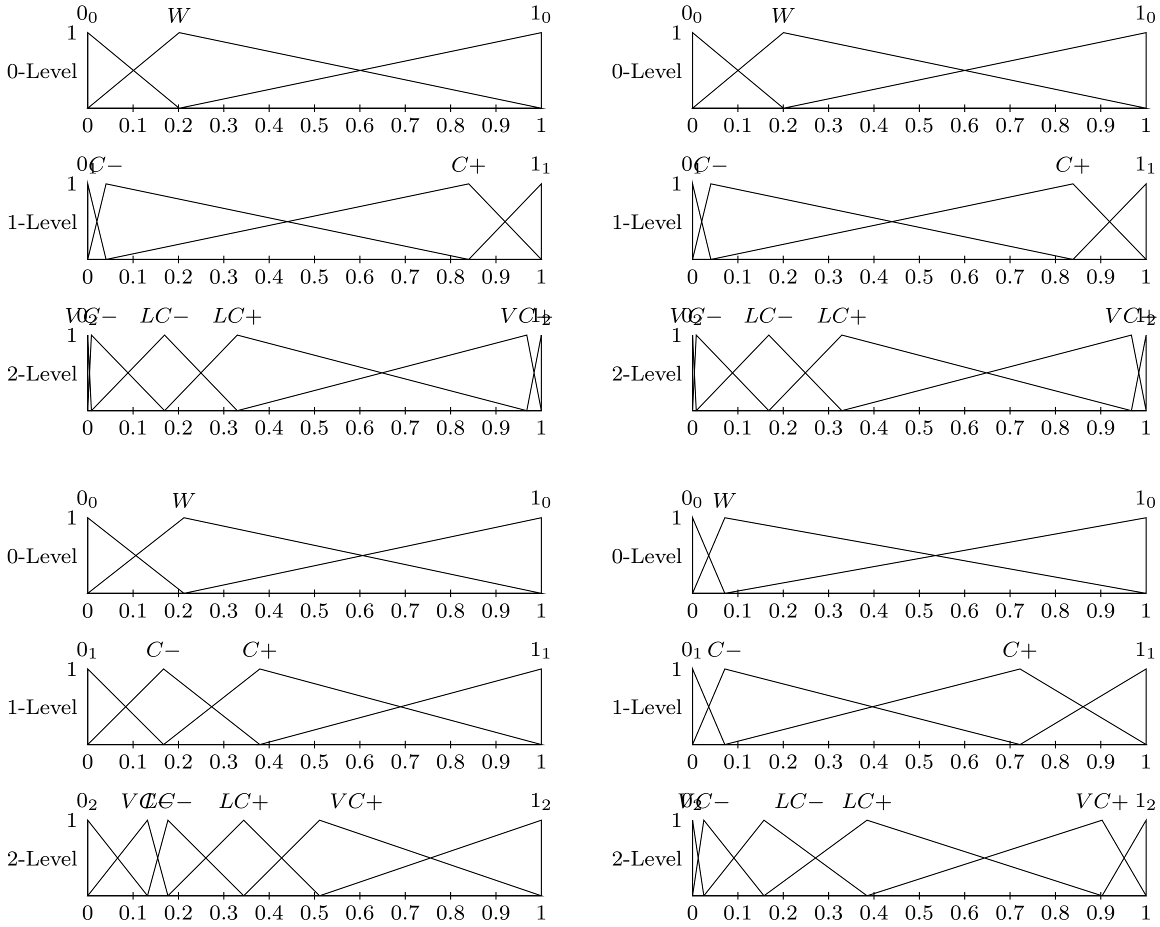


Figure 8. MFs obtained after using GA for optimization

6. CONCLUSIONS

The paper presents the method of mining the association rule according to the hedge algebra's approach based on dividing the fuzzy domain of the attribute values according to the multi-granularity representation. Experimental results based on the database of the US Census in 1995 showed us the advantage of this method. Firstly, it provides a fairly simple but effective way of constructing fuzzy sets and dividing value domain of attributes. Moreover, these fuzzy sets not only ensure the criteria for the fuzzy division system but also provide a good response in terms of semantics to the explored rules. It means that the mining rules include both highly generalized and detailed rules, depending on the data representation layer in the multi-granularity structure we construct based on hedge algebra.

ACKNOWLEDGMENT

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

REFERENCES

[1] J. Alcalá-Fdez, R. Alcalá, M. J. Gacto, and F. Herrera, "Learning the membership function contexts for mining fuzzy association rules by using genetic algorithms," *Fuzzy Sets and Systems*, vol. 160, no. 7, pp. 905–921, 2009.

[2] J. Alcala-Fdez, R. Alcala, and F. Herrera, "A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 5, pp. 857–872, 2011.

[3] M. Antonelli, P. Ducange, B. Lazzerini, and F. Marcelloni, "Learning concurrently data and rule bases of mamdani fuzzy rule-based systems by exploiting a novel interpretability index," *Soft Computing*, vol. 15, no. 10, pp. 1981–1998, 2011.

[4] ——, "Multi-objective evolutionary design of granular rule-based classifiers," *Granular Computing*, vol. 1, no. 1, pp. 37–58, 2016.

[5] G. Castellano, A. M. Fanelli, and C. Mencar, "Fuzzy information granulation with multiple levels of granularity," in *Granular Computing and Intelligent Systems.* Springer, 2011, pp. 185–202.

[6] C.-H. Chen, T. Hong, V. S. Tseng, L.-C. Chen *et al.*, "Multi-objective genetic-fuzzy data mining," *International Journal of Innovative Computing*, 2012.

[7] N. C. Ho, T. T. Son, N. D. Khang, and L. X. Viet, "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.

[8] N. C. Ho and N. Van 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.

[9] T.-P. Hong, C.-H. Chen, Y.-C. Lee, and Y.-L. Wu, "Genetic-fuzzy data mining with divide-and-conquer strategy," *IEEE Transactions on Evolutionary Computation*, vol. 12, no. 2, pp. 252–265, 2008.

[10] C. Mencar, M. Lucarelli, C. Castiello, and A. M. Fanelli, "Design of strong fuzzy partitions from cuts." in *EUSFLAT Conf.*, 2013.

[11] C. H. Nguyen, W. Pedrycz, T. L. Duong, and T. S. Tran, "A genetic design of linguistic terms for fuzzy rule based classifiers," *International Journal of Approximate Reasoning*, vol. 54, no. 1, pp. 1–21, 2013.

[12] C. H. Nguyen, T. S. Tran, 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.

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

[14] P. Pulkkinen and H. Koivisto, "A dynamically constrained multiobjective genetic fuzzy system for regression problems," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 1, pp. 161–177, 2010.

[15] N. T. A. Tran Thai Son, "Hedges algebras and fuzzy partition problem for qualitative attributes," *ournal of Computer Science and Cybernetics*, vol. 32, no. 4, 2016.

[16] D. Wijayasekara and M. Manic, “Data driven fuzzy membership function generation for increased understandability,” in *Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on*. IEEE, 2014, pp. 133–140.

[17] Y. Yao, “A triarchic theory of granular computing,” *Granular Computing*, vol. 1, no. 2, pp. 145–157, 2016.

[18] L. A. Zadeh, “The concept of a linguistic variable and its application to approximate reasoningi,” *Information sciences*, vol. 8, no. 3, pp. 199–249, 1975.

Received on October 10, 2017
Revised on April 20, 2018