Class Association Rule Mining on A Temporal Database using Fuzzy Logic with Genetic Approach

K. Rajeswari [#1],
[#1]Computer Science and Engineering, Sree Sowdambika college of engineering, Aruppukottai

***Abstract*— In this paper we present a GA based fuzzy data mining approach for extracting class association rules from quantitative data with optimal membership function. It is not an easy task to know a priori the most appropriate membership function for mining class association rules. To find the optimal membership function a fuzzy based CHC genetic learning model is used. The learning model is based on the 2-tuple linguistic representation allowing us to adjust the context associated with the linguistic term membership function by considering only one parameter. This model is helpful in obtaining more suitable membership function by reducing the search space. The proposed method is applied on the ozone level detection database from UCI repository to predict ozone day.**

***Keywords*— Data Mining, Fuzzy Set Theory, Genetic algorithm, class association rule, 2-tuple linguistic representation.**

I. INTRODUCTION

Data mining is the process of extracting a hidden, predictive data from a very large database [1]. Data mining techniques support automatic exploration of data and attempts to source out the patterns and trends in data and also to infer rules from those patterns.

Association discovery is one of the most common data mining techniques used to extract interesting knowledge from a very large datasets. Association rule is an expression of the form X→Y, where X, Y are subsets of items and X∩Y=Ø. Once the frequent item sets from the transaction dataset has found, association rules can be generated. With the support threshold and confidence threshold, association rules will be generated [2], [3].

Class association rule is defined to be an implication with a target (class attribute) as its consequence.

In reality, database not only consists of binary attributes but also quantitative attributes. With quantitative attributes, partitioning of data will have unnatural boundaries which lead to overestimate or underestimate the values [4].

Fuzzy set theory has been used more frequently in intelligent systems because of its simplicity and similarity to human reasoning [5].It is used in data mining technique because of its crisp and simplicity nature. It overcomes the problem of class association rule by having natural boundary in partitioning of quantitative data. It facilitates the interpretation of rules in a most realistic way [6]-[8].

In this paper we present an approach to extract fuzzy class association rules from quantitative data with optimal membership function.

II. RELATED WORK

In class association rule mining, partitioning of quantitative data will be with unnatural boundary. If there is unnatural boundary in partitioning of data, it leads to overestimate/underestimate of values [4]. To overcome this drawback, it has been moved towards fuzzy logic, which will have natural boundary. It is very difficult to find out the optimal membership function for the given database [5]-[8]. Some approaches have also achieved learning or tuning of the MFs [9].These methods suffered from exponential growth of search space, while genetic learning, when the number of variables becomes high. This increased search space results in high space complexity to generate the class association rules. To overcome this problem, a new linguistic rule representation model has been proposed to perform genetic learning of MFs [10],[11].

This new approach is based on the 2-tuple linguistic representation [12] that allows the symbolic translation of a linguistic term by considering only one parameter CHC genetic learning model is used to find the optimal membership function to mine class association rules.

III. FRAME WORK FOR MINING CLASS ASSOCIATION RULES

Architecture for mining the Fuzzy class association rule with optimal membership function is given in Fig 1.The algorithm for the proposed approach

A. Dataset used

To evaluate the proposed approach, ozone level detection database which is taken from UCI repository [15] [16] is used. The description of dataset is as follows

Number of attributes: 74
Number of Instances: 2536

Dataset consists of temperature with 26 attributes (T0,……,T23, T_PK, T_AV), wind speed ratio with 26attributes(WSR0,…..WSR23,WSR_PK,WSR_AV),relative humidity with 3 attributes(RH85,RH70,RH50), height with 3 attributes (HT85, HT70, HT50) and directions with 6 attributes (U85, U70, U50, V85, V70, V50) and sea level pressure. Even though the temperature and wind speed ratio are taken for different period of time, the values are with only slight variations.

Input:
R - raw dataset (Ozone Level Database),
Ψ - minimum_support value,
β - minimum confidence value.

Output:
A set of Fuzzy Class Association Rules.

Method:

Step 1: Data Preprocessing
By Attribute selection,
R (a1, a2,·····.,a74)→ P(a1, a3, a6,a9)

Step 2: Optimal MF by CHC Genetic learning Algorithm.

Step 3: Fuzzification
Converting the attribute values into its corresponding linguistic terms using optimal MF
P(a1, a3, a6,a9)→ D(a1.low, a1.medium, a1.high, ..., a9.low, a9.medium, a9.high).

Step 4: Frequent Item Set Generation using APRIORI algorithm.
- Candidate set (C) generation from D
- Frequent Item Set Generation from C using Ψ.

Step 5: Class Association rules generation.
Generate the rules using β .

Fig.1 Algorithm for proposed approach

B. Dataset used

To evaluate the proposed approach, ozone level detection database which is taken from UCI repository [15] [16] is used. The description of dataset is as follows

Number of attributes: 74
Number of Instances: 2536

Dataset consists of temperature with 26 attributes (T0,……,T23, T_PK, T_AV), wind speed ratio with 26attributes(WSR0,…..WSR23,WSR_PK,WSR_AV),relative humidity with 3 attributes(RH85,RH70,RH50), height with 3 attributes (HT85, HT70, HT50) and directions with 6 attributes (U85, U70, U50, V85, V70, V50) and sea level pressure. Even though the temperature and wind speed ratio are taken for different period of time, the values are with only slight variations.

C. Measures used

1) Support

An objective measure for association rules is the rule support, representing the percentage of transactions from a transaction database that the given rule satisfies.
Support value can be defined in [15] given as,

$$support(A \rightarrow C_j) = \frac{\sum_{x_p \in class\ C_j} \mu_A(x_p)}{|T|} \quad (1)$$

The support value for class Cj can be given by,

$$MinimumSupport_{C_j} = minSup * f_{C_j} \quad (2)$$

The fuzzy support of an itemset can be calculated as,

$$Support(A) = \frac{\sum_{x_p \in T} \mu_A(x_p)}{|T|} \quad (3)$$

2) Confidence

Another objective measure for association rules is confidence, which assesses the degree of certainty of the identified association. Confidence value can be defined in [15] given as,

$$Confidence(A \rightarrow C_j) = \frac{\sum_{x_p \in class\ C_j} \mu_A(x_p)}{\sum_{x_p \in T} \mu_A(x_p)} \quad (4)$$

3) Fitness Function

To evaluate a determined chromosome we will use the fitness function which was defined in [8] given as,

$$fitness(C_q) = \sum_{x \in L_1} \frac{fuzzy\ \_support\ (x)}{Suitability\ (C_q)} \quad (5)$$

4) Crossover Threshold

This incest prevention mechanism uses a predetermined threshold, *L* for crossover which was defined in[8] given as,

$$L = (\#Genes * BITSGENE)/4.0 \quad (6)$$

IV. ALGORITHMIC DESIGN CONCEPT

The algorithm for extracting the fuzzy class association rule is given in Fig 2.

A. Data Preprocessing

1) Attribute selection

Attribute selection is done based on cfs subset level algorithm and best first algorithm using weka tool to avoid the redundancy and repetition. to avoid the redundancy and repetition. Based on this algorithm, the selected attributes are WSRAV, T_PK, RH50, HT50 and V70 and class attribute. Dataset after pre-processing is given in Table I.

TABLE I
PRE PROCESSED INPUT DATASET

S.no	WSRAV	T_PK	RH50	HT50	V70	class
1	3.5	22.2	0.6	5790	4.42	0
2	3.2	19.7	0.4	5645	8.11	1
..	..	..	..	..	..	..
1838	4.3	20.1	0.8	4999	8.03	0

1) chromosome representation

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:

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

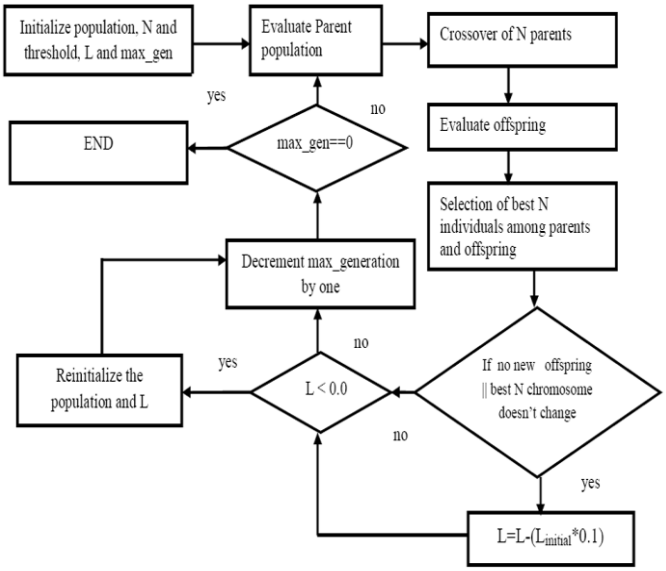

Fig. 3 CHC scheme

Fig 4 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.

2) Initial Gene pool

To make use of the available information, the initial MF's 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).

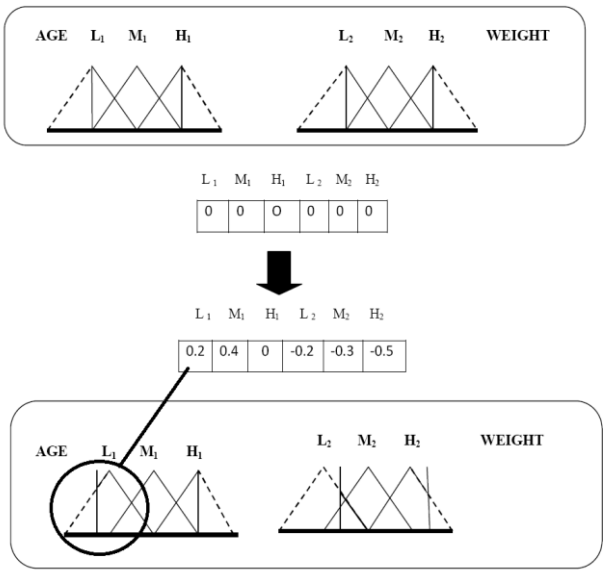

Fig. 4 Example of coding scheme

3) Chromosome evaluation

To evaluate a determined chromosome, use the fitness function. The fitness value of a chromosome C_q is given in equation (5).where L_1 is the set of large1-itemsets obtained by using the set of MFs in C_q , *fuzzy _support*(x) is the fuzzy support of the1-itemset x from the given transaction database by using the equation (3) 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 ,

$$Suitability(C_q) = \sum_{k=1}^{n}[overlap\_Factor(C_{qk}) + coverage\_factor(Cqk)] \quad (7)$$

where *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 of the MFs for an item I_k in the chromosome C_q is defined as,

$$overlap\_factor(C_{qk}) = \sum_{i=1}^{m}\sum_{i=1}^{m}\left[max\left(\frac{overlap\ (R_i,R_j)}{\min\left(spanR_{R_i}, spanR_{R_j}\right)}, 1\right) - 1\right] \quad (8)$$

where *overlap*(R_i , R_j) is the overlap length of R_i and R_j . $spanR_{Ri}$ is the right span of R_i (right extreme minus vertex) , $spanL_{Rj}$ is the left span of R_j (vertex minus left extreme). Notice that, in our case $spanR_{Ri}$ and $spanR_{Rj}$ are the same size. The coverage factor of the MFs for an item I_k in the chromosome C_q is defined as,

$$coverage\_factor(C_{qk}) = \frac{1}{\frac{range\ (R_1,\ldots,R_m)}{\max(I_k)}} \quad (9)$$

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 the entire domain, reducing the computation time. Thus, the suitability of the set of MFs in a chromosome C_q is therefore defined as

$$suitablity(C_q) = \sum_{k=1}^{n}[overlap\_factor(C_{qk}) + 1] \quad (10)$$

4) Crossover operator

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

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

- $O_2 = (o_{21} \cdot \cdot \cdot o_{2n})$, where o_{2i} is a randomly chosen number from the interval $[l^2_i , u^2_i]$, with $l^2_i = \max\{a_i , y_i - I_i \cdot \alpha \}$ and $u^2 i = \min\{b_i , y_i + I_i \cdot \alpha \}$.

5) Restart approach

To get away from local optima, this algorithm uses a restart approach 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 performs the restart procedure when a threshold value is reached or all the individuals coexisting in the population are very similar.

C. Fuzzification

Each quantitative value in the dataset which is specified in Table I, is updated with their membership grades of each and every linguistic term by using the optimal membership function obtained by CHC algorithm.

The membership grades are obtained by using the triangular membership function by using the equation (11). Dataset after fuzzification is specified in Table II.

$$triangle(x; a, b, c) = max\left(min\left(\frac{x-a}{b-a} , \frac{c-x}{c-b}\right), 0\right) \qquad (11)$$

D. Frequent itemset generation

Frequent itemsets are generated from fuzzified dataset whose support values were greater than the minimum support threshold of each class.The support value can be obtained by using the equation (1) . For each class the minimum support value can be fixed by using equation(2).

E. Class association rule generation

Fuzzy class association rules are generated from the frequent itemsets by using minimum confidence threshold. The confidence value is obtained by using the equation(4).

TABLE II

FUZZIFIED DATASET

S.no	WSR_AV			T_PK			....	class
	L	M	H	L	M	H		
1	0.3	0.1	0.0	0.1	0.0	0.0		0
2	0.3	0.6	0.0	0.0	0.1	0.9		1
..								
1838	0.6	0.4	0.0	0.0	0.8	0.1		0

V. EXPERIEMENTAL ANALYSIS

To evaluate the proposed approach we carried several experiments on a real world dataset named ozone dataset.

A. Analysis via support and confidence

Fig 5 depicts the relationship between the number of fuzzy class association rules and the minimum support. Fig. 6 depicts the relationship between the number of fuzzy class association rules and the confidence threshold.

In Classical approach, the rules generated may be discarded because of crisp boundary. So only it generates less number of rules. In Hong et al's approach, crisp boundary is overcome by making use of natural boundary which generates large number of rules. In the proposed approach extracts the best number of fuzzy association rules by making use of optimal membership function.

Finally, an example of classical class association rule based on Apriori approach and classical fuzzy class association rule based on Hong et al's approach and fuzzy class association rule based on the proposed approach with minimum support=0.001 and minimum confidence=0.01 is given below :

Classical Class Association rule-Without Fuzzy Logic

IF **WSR-AV** is Medium
then **day** is ozone
(Support=0.007 , Confidence=0.014
and Medium = 4 to 6.8)

Class association rule – With Classical Fuzzy method:

IF **WSR-AV** is Medium
then **day** is ozone
(Support=0.418 , Confidence=0.5
and Medium = 1.975 to 4.925)

Class association rule - With 2-tuple Fuzzy Linguistic representation :

IF **WSR-AV** is (Medium, -0.1)
then **day** is ozone
ie., IF **WSR-AV** is slighly lower than Medium
then **day** is ozone
(Support=0.016 , Confidence=0.5
and Medium = 1.375 to 2.255)

This example shows that the proposed approach, improves the confidence of the rules obtained and 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.

A. Analysis of time complexity

It can be easily seen from the Fig 7 that the reduction of the search space provided by the 2-tuple linguistic representation allows the proposed approach to decrease its runtime regarding Hong et al's approach.

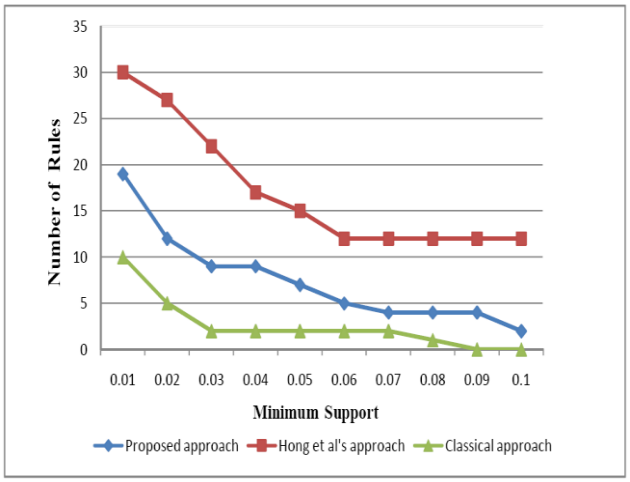

Fig. 5 Relationship between number of rules and minimum support with 0.01 for confidence threshold

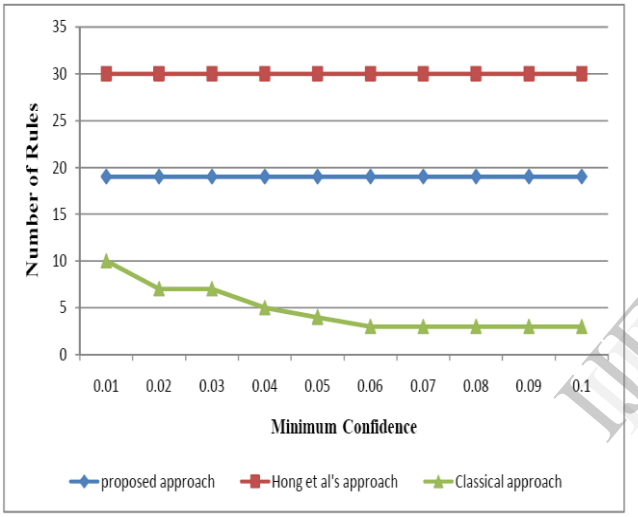

Fig. 6 Relationship between number of rules and minimum confidence with 0.01 for support threshold.

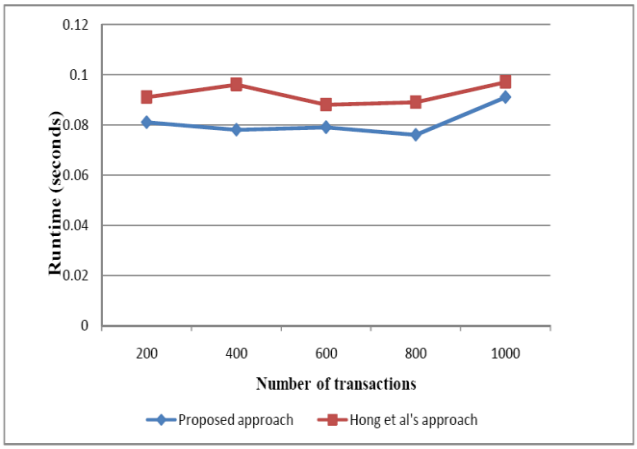

Fig.7 Relationship between number of transactions and runtime

VI. Conclusion

In this paper, we have proposed a GA based fuzzy data-mining algorithm for extracting fuzzy class association rules with optimal membership function. The CHC genetic learning model is used to obtain optimal membership function by adjusting the linguistic term membership function contexts with the help of 2-tuple linguistic representation with reduced search space. This optimal membership function obtained is used to generate frequent class association rules to predict ozone day. The proposed work is compared with hong et al's approach which uses static membership function and Apriori based approach without using fuzzy and found that the proposed work generates less number of more confidence rules with less time complexity.

In future, the proposed work can be extended to work with similar type of environmental datasets, which is crucial to demonstrating the impact of this research.

References

[1] J. Han and M. Kamber,: *"Data mining: concepts and techniques."* Second Edition, Morgan Kaufmann , 2006

[2] R. Agrawal and R. Srikant, *"Fast algorithms for mining association rules in large databases".* Proceedings of 20th International Conference on Very Large Data Bases, pages. 478–499,1994

[3] R. Agrawal, T. Imielinski, and A. Swami, "*Mining Association Rules between Sets of Items in Large Databases*". Proceedings of ACM SIGMOD, pages 207–216,1993

[4] R. Srikant and R. Agarwal,"*Mining Quantitative Association Rules in Large Relational Tables"* . Proceedings of ACM-SIGMOD Conference on Management of Data ,1996

[5] H.J. Zimmermann, "*Fuzzy Set Theory and its Applications"*. Kluwer Academic publisher Boston (1991)

[6] T.Hong, C.Kuo, S. Chi, "*Trade-off between computation time and number of rules for fuzzy mining from quantitative data"*, Int.J. Uncertain. Fuzziness knowledge based systems Vol. 9, No. 5.pp 587–604 ,2001

[7] J.H lee, H.L Kwang,*" An Extension of association rules using fuzzy sets".* Proceedings of 7th international fuzzy systems ,1997

[8] T.Dhanya and D.Nagesh Kumar, *"Data Mining for Evolving Fuzzy Association Rules for Predicting Monsoon Rainfall of India"* . Journal of intelligent systems 2009.

[9] T. Hong, C. Chen, Y. Wu, Y. Lee, (2006) , *"A GA-based fuzzy mining approach to achieve a trade-off between number of rules and suitability of membership functions "*- Soft Computing, Springer-verlang, 2006 Vol. 10, No. 11 pages 1091–1101.

[10] Alcalá , J. Alcalá-Fdez, F. Herrera , *"A proposal for the genetic lateral tuning of linguistic fuzzy systems and its interaction with rule selection",* IEEE Trans. Fuzzy Systems Vol. 15 , No. 4, pp 616–635,2007.

[11] J. Alcal´a-Fdez, R. Alcal´a, M. 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. 905–9219 ,2009

[12] F. Herrera and L. Martnez, *" A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Transactions on Fuzzy Systems"*, vol. 8, no. 6, pp. 746–752 , 2000.

[13] Kun Zhang, Wei Fan, *"Forecasting skewed biased stochastic ozone days: analyses, solutions and beyond, Knowledge and Information Systems",* Vol. 14, No. 3,2008.

[14] UCI Repository of machine learning databases, http://archieve.ics.uci.edu

[15]. Jesus Alcala-Fdez, Rafael Alcala, and Francisco Herrera, *"A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning".* IEEE Transactions, vol: 19, 2011