Journal of Theoretical and Applied Information Technology
31st January 2016. Vol.83. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A NEW SOFT COMPUTING TECHNIQUE FOR EFFICIENT RULE MINING

SIJI P D [1] , DR.M.L.VALARMATHI1[2], S.MOHANA[3]

[1] Assistant Professor, Department of Computer Science, St. Josephs College Irinjalakuda, Thrissur, Kerala, India

[2] Associate Professor Department of CSE, Government College of Technology, Coimbatore, Tamil Nadu, India,

[3] PG Research Scholar, Department of CSE, Government College of Technology, Coimbatore, India

E-mail: [1] Srblessy@gmail.Com,[2]drmlv@gct.ac.in, [3]mohana.Apr7@gmail.Com

ABSTRACT

This paper proposed a prediction model based on Fuzzy Association Rule (FARs). Present a two model; it is the integration of the Fuzzy C-Means (FCM) and Multiple Support Apriori (MS apriori). Even though, enhancing the knowledge from the large dataset, give large amount of data. So reduce some amount of data through hybrid approach of Genetic algorithm (GA) with Particle Swarm Optimation (PSO), both useful for the fuzzy association rule. Extraction of knowledge process is performed thorough betathalesemia Dataset. FCM and MSapriori model is used to extract the FARs. experimental result proves the proposed model of hybrid feature reduction with FCM-MS apriori model is effectively predict the knowledge from the large database compared to other prediction model.

Keywords: *Association Mining, Fuzzy Rule, Genetic Algorithm*

1. INTRODUCTION:

Data mining is defined as knowledge discovery from databases; it can also be defined as the extraction of knowledge from the large database [1]. Several methods and techniques are needed to know about this different type of knowledge. Depend upon this knowledge, different classifications are handled in data mining tasks which are summarization, classification, clustering, association, and trend analysis [2].

Most common seen in data mining are extraction association rule from database [3, 4]. Association rule is mainly used for discovering the attributes between the huge dataset along with relationship among the dataset. This rule can be efficient providing results for uncovering unknown relationship based on the forecast and decision. Based on the above statement, association rule is one of the famous areas along with application and development in data mining [5].

Generally association rule is an extension of the Boolean association rule, many effective algorithms is proposed by various researchers, PUTAX [10], parallel algorithm [11], Partial-Match Retrieval method [12], multiple concept level [14], pruning method [13]. In early, implement a technique for mining association rule based on quantitative and categorical
attributes from the database [6]. For information process, mining fuzzy rule system is an important factor and can be used for various researches [7, 8, and 9].

This paper proposed the method for extraction of association rule from quantitative data using fuzzy clustering technique. Extracting the knowledge from the quantitative database based on Fuzzy Association Rule (FARs). In this paper also investigates the hybrid method of genetic algorithm (GA) with Particle Swarm Optimation (PSO) for feature reduction from the database. Prediction of the database and the extraction of the Fuzzy Association Rule are based on two methods, the first method is the combination of the Fuzzy C-Means (FCM) and the Apriori approach. Second method is based on FCM and a multiple support thresholds approach (FCM-MS Apriori).

The rest of this paper is organized as follows. The next section presents the literature about the mining of fuzzy association rule. Section 3 describes the feature reduction algorithm. Section 4 gives about the algorithm and prediction models.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2. LITERATURE SURVEY

In many research fields, fuzzy mining is widely used such as sequential pattern mining, intrusion detection, biological knowledge extraction etc., [15, 16, 17 and 18]. In these above research, membership function has to be advance. To derive a predefined number of membership function based on GA to obtain a maximum profit [19].

Extracting an accurate and compact fuzzy model based on the hybrid approach is utilized in [20]. In first stage to assign the suitable fitness for each individual member based on the GA. In second stage, to optimize the model parameter based on the Unscented Filter (UF).

Investigate the new fuzzy data mining technique with the help of genetic learning of the membership function for extracting the both fuzzy association rule and membership function from quantitative transactions. It is composed of two stages, first stage to learn the membership function by genetic process, second method to mine the fuzzy association rule [21].

Mining a Fuzzy association rule in a probabilistic quantitative database by formulating a fuzzy–probabilistic database is implemented in [22]. Here author and his coworker discover Frequent Rule fuzzy–probabilistic Item sets (FRIs) and Fuzzy Association Rules (FARs) by an novel algorithm caller FARP (mining Fuzzy Association Rules from a Probabilistic quantitative data).

3. FEATURE REDUCTION METHOD USING HYBRID GA WITH PSO

Genetic Algorithms (GAs) are adaptive heuristic search algorithms based on the evolutionary ideas of natural selection and genetic processes. The functional principal of the genetic algorithm is first proposed by john Holland in 1975, inspired by the mechanism of natural selection where stronger individuals are likely the winners in a competing environment. GA is successfully applied to many fields such as optimization, machine learning, neural network, fuzzy logic controllers [23, 24]. Idea behind the GA is genetic and evolutionary theory. Three main operators in the genetic algorithm is crossover, mutation and selection operation.

The crossover operation generates offspring from two chosen individuals in the population, by exchanging some bits in the two individuals. The offspring thus inherit some characteristics from each parent.

The mutation operation generates offspring by randomly changing one or several bits in an individual. The offspring may thus possess different characteristics from their parents. Mutation prevents local searches of the search space and increases the probability of finding global optima.

The selection operation chooses some offspring for survival according to predefined rules. This keeps the population size within a fixed constant and puts good offspring into the next generation with a high probability.

Eberhart and Kennedy in 1995 developed a stochastic optimization technique; this is inspired by the social behavior of bird flocking or fish schooling [25]. In PSO, each particle particle keeps track of its coordinates in the problem space, which are associated with the best solution (fitness) it has achieved so far (The fitness value is also stored). This value is called pbest. Another ''best'' value that is tracked by the particle swarm optimizer is the best value, obtained so far by any particle in the neighbors of the particle. This location is called lbest. When a particle takes all the population as its topological neighbors, the best value is a global best and is called gbest.

The idea behind the PSO is changing the each particle velocity towards its pbest and lbest locations. Random term is used to weighted the acceleration, individual random numbers are generated for acceleration towards pbest and lbest. PSO also applied in many areas and research fields and it is one of the fastest optimation technique and also cheapest one [26].

Algorithm for hybrid method for feature reduction:

STEP 1: Start with a randomly generated particles/population.

STEP 2: Assign the variable and chrom the particles.

STEP 3: Maximum number is generated upto this number, following two steps is repeated.

STEP 4: Start PSO method.

4.1: for each particle, calculate the fitness, if the fitness value is better than the pbest value. Then the current value is assigned as new pbest

4.2: select the particle with best fitness value of all particles as the gBest.

4.3: Update all particle velocity and their position, until n particles.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.4: Finally save the best particle information.

STEP 5: Again start GA method, fitness calculation for each chromosome. This fitness calculation process is repeated until n offspring have been created.

STEP 6: Select a pair of parent chromosomes from the current population, the probability of selection being an increasing function of fitness. Selection is done with replacement, meaning that the same chromosome can be selected more than once to become a parent.

STEP 7: formation of two offspring with the help of probability of the crossover rate.

STEP 8: in the mutation rate, mutate the two offspring and kept the result in the new population.

STEP 9: replace the current population with new population.

4. ALGORITHM AND PREDICTION MODEL

A. The FCM-APRIORI Model:

To construct KB from the database, need to extract fuzzy rules. To obtain this, in this paper implemented the FCM-Apriori model.

Transformation of lot amount of data set in to the fuzzy sets is obtained by FCM. It is based on the objective functioning method and it also one of the fuzzy clustering algorithm, developed by Bezdek in 1981. Another approach such as Apriori is used to extract the fuzzy term sets from the fuzzy dataset. For this extraction purpose some measures are used, such as minsupp and minconf.

Following steps are used in this paper

Step 1: Getting the data from the database.

Step 2: Feature reduction is used to reduce the data

Step 3: FCM, transformation of data set into the fuzzy set.

Step 4: Extraction of fuzzy rule based on the Apriori method.

Step 5: Fuzzy rule is saved in the KB.

Step 6: For prediction, Fuzzy Inference System is used to command the KB.

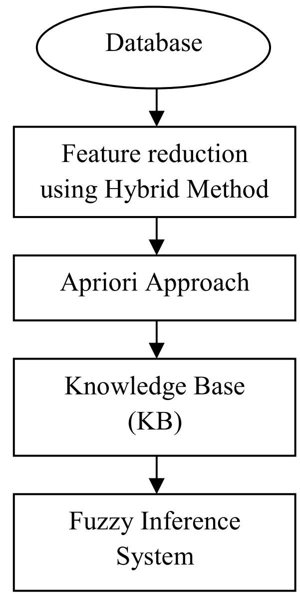

Figure 1: Block Diagram for proposed framework

The following steps are used for proposed FCM-Apriori Model:

Step 1: Cluster the data into terms by using FCM, from the data find out the center of the fuzzy set, minimum and maximum value for each field of the input data set.

Step 2: Using triangular and trapezoid membership function, convert the dataset tin to the fuzzy data set.

Step 3: In each term for all records calculated the support value by summing the fuzzy membership value, it is obtained by using below equation

$$Sum_{if} = \sum_{i=1}^{n} \mu(X)_{if} \quad (1)$$

Step 4: finally this summation value is stored in the candidate termset C_1.

Step 5: If the term is greater than or equal to minsupp are removed to L_1.

Step 6: Terms are joined and produce the fuzzy tern, in this term $c[1], c[2], \dots, c[n]$, here $c[1]$ indicate the first fuzzy term and $c[n]$ represent the last fuzzy term.

Step 7: Each termset does not belong to the same field., that is $c[1] \cap c[i] = \emptyset, c[1] \cap c[i+1] = \emptyset \dots . c[1] \cap c[n] = \emptyset$. if the every termset is kept their values in the candidate termset C_2, using minimum operator support vector value is calculated. At last summation result is stored in candidate termset C_2.

Step 8: if the termset is greater than minsupp are moved to L_2.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Step 9: again terms are joined and combined as $L_2 = p\ join\ L_2 = q$, here $p.term_1 = q.term_1 \ldots .p.term_{k-2} = q.term_{k-2}, p.term_{k-1} \neq q.term_{k-2}, p.term_{k-1} \neq q.term_{k-1}$. Depend on the every sub-termset of the candidate, this combination is formed.

Step 9: Again termset is calculated and stored in the candidate termset C_3.

Step10: if the support value and term set in C_3 is greater than or equal to minsupp are moved to L_3.

Step 11: Again joined and combined processes are happened until L_n is empty.

Step 12: Termsets are pruned by selection of the termsets including the target attribute. As a consequence, termsets are phrased in IF–Then form, then the Confidence Value (CV) is calculated based on below equation. The rules that are greater than or equal to minconf are accepted. Then the contradiction rules are removed, based on the CV.

$$CV = \frac{\sum[(IF) \cap (Then)]}{\sum(min(IF))}$$

B. The FCM-MSapriori model:

Suppose assumed that all the items in the database have the same frequency, if using the single minsupp for whole database. Even though, in real applications, high frequency and low frequency item sets are obtained in the database. So needed to find out the frequency item set, for this finding process to set the minsupp for specific value. if it is set as a too high, then it extract a low number of frequent item sets. When minsupp is low, then it extracts the high number of frequent item set. In this condition some of the frequent itemsets are uninteresting or insignificant. in the low number of frequent itemset, rate item problem is appeared and is called as dilemma. To prevent or overcome the dilemma, Msapriroi approach is introduced. This approach is based on a multiple minimum support thresholds, here the number is generated is depend on the control parameter. MSapriori approach is used for extracting FARs of rarely and highly frequent termsets from fuzzy data sets.

5. Experimental Results

For analysis and validation purpose, betathalesemia Dataset is used. We used Java Programming and MySQL for getting the result. The study is helpful to determine the regional and ethnic distribution of beta thalassemia mutation and the effect of consanguinity in patients referred for prenatal diagnosis of beta b-thalassemia and to target the high risk population for screening. In populations with high consanguinity rates and common inherited blood disorders, community programs for premarital screening to detect carriers of hemoglobinopathies such as thalassemia and sickle cell anemia are in progress. Simulation results prove the proposed model through some illustration.

Table 1: Part of C_1

Term	Support
1.000000000000 000	18.50000000000 0004
2.000000000000 000	35.99999999999 9993
3.000000000000 000	35.99999999999 9993
4.000000000000 000	35.99999999999 9993
5.000000000000 000	18.50000000000 0000
6.000000000000 000	36.00000000000 0000
7.000000000000 000	35.99999999999 9986
8.000000000000 000	36.00000000000 0007

The above table gives the candidate term of C_1 after the calculation of the support value for each term.

Table 2: Part of L_1

Term	Support
1.000000000000	18.50000000000
2.000000000000	35.99999999999
3.000000000000 000	35.99999999999 9993
4.000000000000 000	35.99999999999 9993
5.000000000000 000	18.50000000000 0000
6.000000000000 000	36.00000000000 0000
7.000000000000 000	35.99999999999 9986
8.000000000000 000	36.00000000000 0007
9.000000000000	18.50000000000

Journal of Theoretical and Applied Information Technology
31st January 2016. Vol.83. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

000	0000
10.000000000000000	35.999999999999993
11.000000000000000	36.000000000000000

Table 2 gives the termset, here the support values that are greater than or equal to minsupp. The above termset are the part of the large termset of L_1.

Table 3: Part of C_3

Term	Support
[1,7]	18.5
[1,8]	9
[1,13]	18.5
[1,14]	9
[1,19]	18.5
[1,20]	9
[2,7]	8.9
[2,8]	35.9
[2,9]	9
[2,13]	8.9
[2,14]	35.9
[2,15]	9
[2,19]	9
[2,20]	35.9

The above table 3 gives the value for candidate termset of term1 and candidate terms of C_2 respectively.

Table 4: Part of L_2

Term
[1,5]
[2,6]
[2,10]
[2,13]
[6,1]
[10,11]
[12,11]
[12,21]
[12,24]

The above table 4 gives the value for large termset L_2 respectively.

Table 5: Part of C_3

Term	Support
[1,5,6]	0
[1,6,7]	0
[1,7,8]	9
[1,8,9]	0
[1,9,10]	0
[1,11,12]	0
[1,12,13]	0
[1,13,14]	8.9
[1,14,15]	0

The above table 5 gives the storage of support value result in the candidate term of C_3.

Table 6: Part of L_3

Term	Support
[1,7,8]	9
[1,13,14]	8.9
[1,19,20]	9
[2,7,8]	9
[2,13,14]	8.9
[2,19,20]	9
[3,7,8]	9
[3,13,14]	8.9
[3,19,20]	9
[4,7,8]	9
[4,13,14]	8.9
[4,19,20]	9

The above table 6 gives the large termset value of L_3 with their support value.

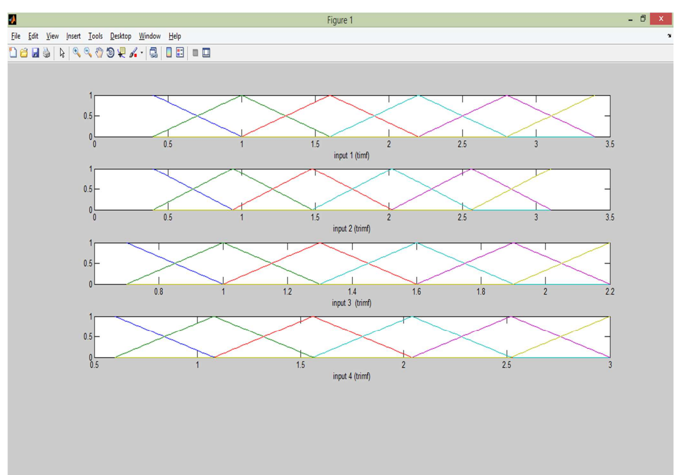

Figure 1: Membership Function

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 1 shows the membership function for the fuzzy set. This membership function is formed for four attribute.

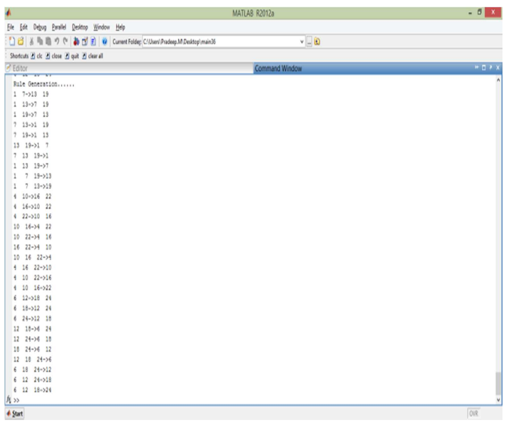

Figure 2: Result Of Rule Generation

Figure 2 shows rule generation result in MATLAB command window.

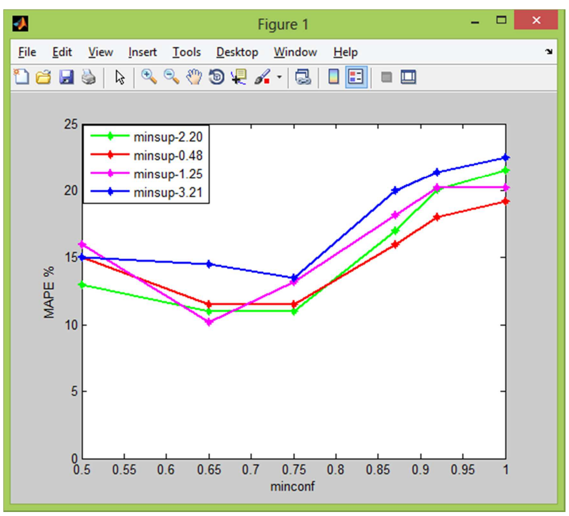

Figure 3: The MAPE Over Different Mincoff And Minsupp

The above figure 3 shows the minimum MAPE produced with minsupp=1.25 and minconf=0.95. When the minsupp value is lower than the 1.25, then the MAPE will increase for most of the mincoff values. Because large number of rules are generated by theses low minsupp values, due to this value noise will be produced in the fuzzy inference system. If the number of rule generation is low means, on that stage minsupp value is large.

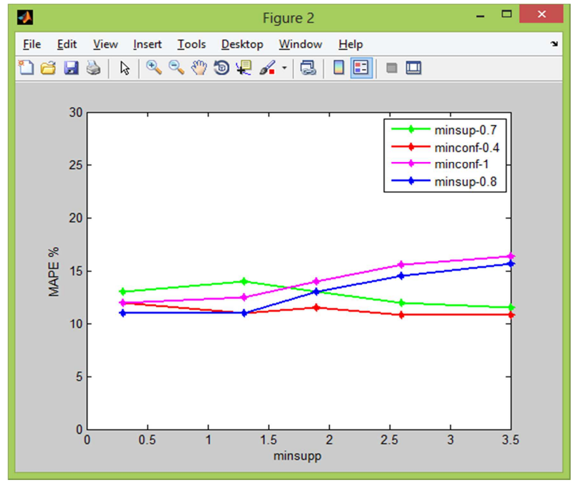

Figure 4: The MAPE Over Different Minsupp And Minconf

The above figure 4 shows the similar sensitivity analysis for the large dataset. The MAPE value for the large dataset size is less than the small data set value. Here, if the rule generation is large means, in that condition minconf value is low. If it is greater than the smaller number of rules are generated.

Table 7: Calculation Of MAPE For Betathalesemia Dataset

Methods	MAPE Value
FCM and Apriori	13.4
FCM and MS apriori	12.9
FCM and MSMM Apriori	11.5
Proposed method	9.7

Table 7 provides the result of MAPE for betathalesemia Dataset. It presents the value from using FCM-Apriori, FCM-MSapriori, FCM-MSMMapriori and proposed method. The first approach of FCM-apriori algorithm gives the MAPE value of 13.4, MAPE of 12.9 is obtained from FCM-MSapriori, 11.5 is noted from the FCM-MSMMapriori model, while our model feature reduction with FM-MSapriori model is equal to 9.7. From these result concluded that the proposed model of hybrid feature reduction with FCM-MSapriori model provides better performance compared to other prediction models. These different MAPE value is illustrated in figure 5.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

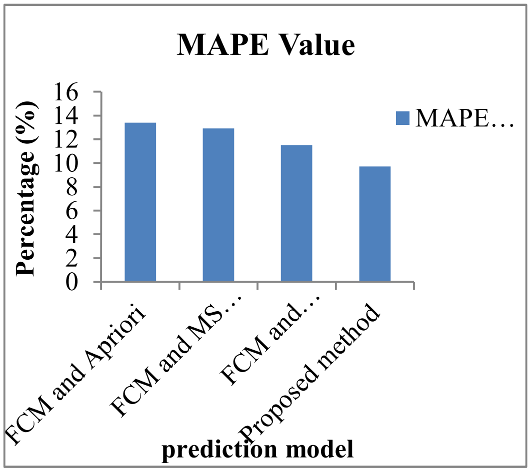

Figure 5: MAPE Value Through Different Prediction Model

6. CONCLUSION:

This paper proposed two type of prediction model to enhance the knowledge from the large database. This prediction model is used by a novel approach of fuzzy association rule. This rule is one of the mining approaches. In this paper integration of the FCM and MSapriori approach is used to generate the dominating FARs. This model used betathalesemia Dataset. Feature reduction is implemented in this paper to reduce the data from the database; it is obtained by using hybrid approach of GA with PSO. Also observed from, that the large data set offered lower MAPE value, it is noted from the difference of minsupp and minconf value. FM-MSapriori model is used to prevent or overcome the existing real application drawbacks, which some dataset contain high frequency and some of item set having lower frequency. the proposed approach proves that it have improvement in the extraction of FARs compared to the existing approaches. Efficient rule mining with the implementation of rough set and better performance analysis can be done in future to get efficient and relevant rule.

REFERENCES

[1] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, A. Zanasi, *Discovering Data Mining from Concept to Implementation*, Englewood Cliffs, Prentice-Hall, 1997

[2] M.S. Chen, J. Han, P.S. Yu, *Data mining: an overview from a database perspective*, IEEE Transactions on Knowledge and Data Engineering (1996) 866–883.

[3] R. Agrawal, T. Imielinksi, A. Swami, *Mining association rules between sets of items in large database*, in: The 1993 ACM SIGMOD Conference, Washington, DC, USA, 1993, pp. 1–10.

[4] R. Agrawal, T. Imielinksi, A. Swami, *Database mining: a performance perspective*, IEEE Transactions on Knowledge Data Engineering (1993) 914–925.

[5] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, Los Altos, 2000.

[6] Srikant, R., & Agrawal, R. (1996). *Mining quantitative association rules in large relational tables*. In The 1996 ACM SIGMOD international conference on management of data (pp. 1–12).

[7] Z. Pei, X. Liu, L. Zou, Extracting association rules based on intuitionistic fuzzy sets, Int. J. Inno. Comput. Inf. Control 6 (6) (2010) 2567–2580.

[8] M. Ragulskis, K. Lukoseviciute, Non-uniformat tractor embedding for time series forecasting by fuzzy inference systems, Neurocomputing 72 (2009) 2618–2626.

[9] G. Raschia, N. Mouaddib, SAINTETIQ: *a fuzzy set-based approach to database summarization*, Fuzzy Sets Syst. 129 (2002) 137–162.

[10] Hipp, J., Myka, A., Wirth, R., & Guntzer, U. (1998). *A new algorithm for faster mining of generalized association rules*. In PKDD'98 (pp. 74–82).

[11] Sung, S. Y., Wang, K. & Chua, L. (1996). *Data mining in a large database environment*. In Proceedings of the IEEE international conference on systems, man, and cybernetics (pp. 988–993).

[12] Matsumoto, K., Hayase, T., & Ikeda, N. (1999). *Data mining of generalized association rules using a method of partial-match retrieva*l. In Discovery science (pp. 160–171.

[13] Huang, Y.-F., & Wu, C.-M. (2002). *Mining generalized association rules using pruning techniques*. In Proceedings of the IEEE international conference on data mining (pp. 227–234).

[14] Gaul, W., & Schmidt-Thieme, L. (2001). *Mining generalized association rules for sequential and path data*. In ICDM (pp. 593–596).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[15] Romsaiyud1, W., & Premchaiswadi, W. (2011). *Applying mining fuzzy sequential patterns technique to predict the leadership in social networks*. In Ninth international conference on ICT and knowledge engineering (pp. 134–137).

[16] Wang, M. H., Su, X. B., Liu, F. M., & Cai, R. C. (2012). *A cancer classification method based on association rules. In 9th International conference on fuzzy systems and knowledge*, discovery (pp. 1094–1098).

[17] Watanabe, T., & Fujioka, R. (2012). *Fuzzy association rules mining algorithm based on equivalence redundancy of items*. In IEEE international conference on systems, man, and, cybernetics (pp. 1960–1965.

[18] Tajbakhsh, A., Rahmati, M., & Mirzaei, A. (2009). *Intrusion detection using fuzzy association rules.* Applied Soft Computing, 9(2), 462–469.

[19] Kaya, M., & Alhajj, R. (2003). *A clustering algorithm with genetically optimized membership functions for fuzzy association rules mining*. In The IEEE international conference on fuzzy systems (pp. 881–886).

[20] M. Eftekhari , S.D. Katebi, *"Extracting compact fuzzy rules for nonlinear system modeling using subtractive clustering, GA and unscented filte*r", Applied Mathematical Modelling 32 (2008) 2634‑2651.

[21] Jesús Alcalá-Fdez, RafaelAlcalá, María José Gacto, Francisco Herrera, "Learning the membership function contexts for mining fuzzy association rules by using genetic algorithms", Fuzzy Sets and Systems 160 (2009) 905–921.

[22] Bin Pei, Suyun Zhao, Hong Chen, Xuan Zhou, Dingjie Chen, *"FARP: Mining fuzzy association rules from a probabilistic quantitative database*", Information Sciences 237 (2013)242–260.

[23] Alcala, R., Alcala-Fdez, J., Gacto, M. J., & Herrera, F. (2007). *Genetic learning of membership functions for mining fuzzy association rules.* In The IEEE international conference on fuzzy systems (pp. 1–6).

[24] Gautam, P., Khare, N., & Pardasani, K. R. (2010). *A model for mining multilevel fuzzy association rule in database*, CoRR abs/1001.3488.

[25] R.C. Eberhart, J. Kennedy, *A new optimizer using particle swarm theory, in: Proceedings of the Sixth International Symposium on Micromachine and Human Science*, Nagoya, Japan, 1995, pp. 39–43.

[26] P.J. Angeline, *Evolutionary optimization versus particle swarm optimization: philosophy and performance differences*, in: Evolutionary Programming VII, Lecture Notes in Computer Science, vol. 1447, Springer, 1998, pp. 601–610.