Journal of Theoretical and Applied Information Technology
31st January 2021. Vol.99. No 2

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

INTERVAL TYPE-2 FUZZY LOGIC USING GENETIC ALGORITHM TO REDUCE REDUNDANT ASSOCIATION RULES

MAGDA M. MADBOULY, EMAN ABD EL REHEEM, SHAWKAT K. GUIRGUIS

Information Technology Department

Institute of Graduate Studies and Research - Alexandria University

163 Horreya Avenue, Shatby-P. O. Box: 832, Alexandria, Egypt

E-mail: {mmadbouly, eman.abdelreheem, shawkat_g}@alexu.edu.eg

ABSTRACT

Data mining (DM) is an analysis extensive data in order to gain the novel and hidden information. DM becomes vital to a lot of research domain like soft computing, artificial intelligence, statistics and machine learning. One of the important topics of DM is Association Rule Mining (ARM) in mega databases where it is used to discover frequent itemsets using statistical metrics such as support (Sup) and confidence (Conf) which is an essential process in ARM. Also ARM is practiced to produce association rules (ARs) from frequent itemsets. Such ARs reveal a link between items in the real world. Several algorithms have been submitted to achieve these ARs. However, these algorithms suffer from redundancy problems and a large number of derived ARs, which makes the algorithms ineffective and complicated them for the end users to understand the rules that were created. Because of these motives, this paper uses the type-2 fuzzy association rules mining technique (T2FARM) to achieve frequent itemsets and identify all relationships between items and ARs that achieve minimum support (min sup) and minimum confidence (min conf) in addition to pruning redundant rules. And also adapts genetic algorithm (GA) to improve non-redundant rules derived. Empirical evaluations display that the proposed technique improves redundant rules pruning of DM compared to traditional fuzzy association rules (FARs) and able to improve non-redundant rules by GA.

Keywords: *Association Rule, Apriori Algorithm, Type-2 Fuzzy linguistic, Redundancy of Fuzzy Association Rules, Genetic Algorithm.*

1. INTRODUCTION

With the continuous improvement of database techniques, there are a huge number of databases have been created in several areas, resulting difficult to the search for useful information from this mountain of raw data. Therefore, researchers found a solution in the DM [1]. DM intends to explore useful information in these data. It is the stage in the knowledge discovery process that attempts to identify new and key patterns in the data [2]. DM includes many techniques such as classification, clustering, ARM, etc. where classification classifies new data based on the class labels and training set in a classification attribute. Classification techniques include neural networks, decision trees, and rule-based systems. Clustering collects data objects in many cluster so that objects within the group (cluster) have high similarity, but dissimilar to objects in other groups such as K-means, and K-medoids. ARM uses to identify frequent patterns and associations from data in databases such as market basket data analysis [3] [4]. In this research, we use ARM to find associations between variables in our database.

In DM, ARs have become represent an importantly to Finding interesting association and tacit information among a wide range of data items. The goal of ARs is analysis of sales processes and mining patterns as well as informing the customer's buying behavior. For instance, from the transaction data of the supermarket, the rule {biscuits, cheeps} => {Coca-Cola} is created, which represents these customers who buy biscuits and cheeps also buy Coca-Cola [5] [6] The apriori algorithm is a famous algorithm for ARs. It applies to derive all the frequent itemsets from the database and produce ARs to discover knowledge using pre-defined threshold measures minsup and minconf [7].

Normally, ARM includes two main parts: first, minsup is calculated to discover all frequent itemsets in the database by applying apriori algorithm where uses the itemsets and creates candidate K-itemsets and next check it with the minsup, if candidate K-itemset is higher than minsup then it is frequent itemset, but if it is less than minsup then pruning it. Second, from these frequent itemsets, the ARs are produced which fit

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

both the minsup and minconf thresholds [8].

In ARM, a huge amount of rules can be obtained. Some of these rules may be redundant and do not provide any new information that impedes the effective use of the discovered knowledge. Moreover, many of the rules extracted do not provide any benefit to the user or can be changed by other rules; therefore it is redundant rules [9]. So, it is considered a major problem in ARM. To address this problem, some researchers presented the fuzzy set theory to ARM (called FARs).

Fuzzy systems can assist in reducing the key drawback current ARM suffers from depending on the fuzzy set concept. FARs is a stretch of the classical ARM by defining sup and conf of the fuzzy rule. It is utilized for converting quantitative data into fuzzy data [9] [10]. So, the FARs has a appropriate property in terms of quantization of numerical attributes in the database compared with Boolean quantized generalized ARM. Ultimately, the mining outcomes of FARs give linguistic terms rather than intervals, which are simple to understand and nearer to the user's mind [11].

GA is searching and optimization methods that are applied to obtain the optimal solutions to a given computational problem through the notation of the model in a chromosome. So, it is flexible due to the same GA can be done with various representations. It is essential when identifying ARs because it operates with the global search to define frequency items and to optimize the rules. Exploring ARs by the GA is utilized in real issues such as commercial databases and fraud detection [12].

The contribution of this article is to prune redundant rules resulting from the apriori algorithm for a real DM application and to get frequent itemsets by creating minsup and minconf fit to a database to be mined. The proposed system relies on the using of T2FARM to reduce the redundant rules derived for an actual application and get all ARs that meet minsup and minconf values in addition to applying GA to optimize the non-redundant rules that presented from T2FARM.

A type-2 fuzzy logic system (T2FLS) is useful in cases where it is difficult to determine an accurate certainty and measurement uncertainties. It is known that T2 fuzzy sets allow to model and decrease uncertainties in rule-based FLS. GA discovers the optimized non-redundant rules that produced from T2FARM and foretell strong rules. The results led to displays the improvement of redundant rules pruning in addition to improving non-redundant rules compared to traditional FARs.

This paper is ordered as follows: Section 2 briefly discusses the research related to ARs. The proposed model is introduced in Section 3. The experimental result and evaluation of the proposed system are displayed in section 4. Finally, the conclusion and future work are given in section 5.

2. LITERATURE SURVEY

DM became an essential research area over the last decade due to it represents a rich research area to obtain useful hidden information from the data and getting associations among lots of domains in massive relational databases. So, there are several works done in academia and industry on developing new approaches to DM. For instance, in [13], the authors presented a classification DM technique to aid in developing the quality of the educational system by evaluating student data to study the attributes that may affect the student performance in courses. This study used the decision tree as a classification method to extract rules for studied and evaluated. The extracted classification rules allow students to predict the final grade in a course under study.

In [14] the authors developed a new DM technique and software for medical data analysis. This approach named the hybrid prediction system. It consists of rough set theory and artificial neural network for dispensation medical data. The hybrid prediction system incorporates rough set theory and artificial neural network to make proficient data analysis and indicative predictions. The authors in [15] compared the performing of various DM techniques in e-learning systems to predict the marks that university students will obtain in the final exam of a course. Many classification methods have been adopted, like decision trees, fuzzy rules, and neural networks. There are three steps designed for the mining process: First, pre-processed to convert the valuable data into KEEL data files. Then, DM algorithms performed to get hidden knowledge inside the data of concern for the instructor. Finally, post-processing models got saved into result files that necessity be understood by the teacher to produce decisions regarding the students. Their tests show not produces better classification accuracy, and there has not one single algorithm that achieves the best classification accuracy with every dataset.

ARM is a popular researched method for DM. It aids in discovering associations between items in addition to knowing strong rules discovered from databases. This technique serves users by enabling buying products relying on their preferences. The academics in [16] applied a generalized ARM which pruning methods for generalizing rules. This process produces rarer candidate itemsets, and a vast count of rules pruned by min conf. This study suffers from employing original frequent and ARs as opposed to rescanning the database. Though, it illustrations can prune a tremendous count of rules.

The study in [17] introduced an approach for optimization ARM. It based on mining interesting positive and negative rules from frequent and infrequent pattern sets. This algorithm resolves the problem of negative rule generation and also optimized the process of rule generation. The algorithm of mining interesting positive and negative ARs can be described as follows: 1) Take an input dataset that contains several attributes 2) Initialize the data with length of the itemsets and calculate sup and conf. 3) Generate all the frequent and infrequent itemsets 4) Generate positive ARs from frequent items sets and negative ARs from infrequent itemsets. 5) Initialize all the general parameters involved in GA. 6) generate the child chromosomes of the positive and negative ARs and calculate the fitness value of each child chromosomes. Compare the individual fitness value of each child with the average fitness value and regenerate positive and negative ARs. 7) Crossover and mutate the remaining child chromosomes and reinitialize the fitness value and recalculate and regenerate final positive and negative rules.

The researchers in [18] applied DM to detect positive and negative ARs. It studies the sup and conf methods, the explanation of the positive and negative ARs, and the conflicting rules problems in the positive and negative ARM and the solutions of these conflicting rules problems. This research can implement in several applications to detect robust patterns and produces all varieties of limited rules.

Some works that applied FLS for processing the uncertain information in ARs. For instance, the authors in [19] suggested a FARM, which employs adjusted difference analysis to identify interesting associations among attributes without using any user-supplied thresholds such as minsup and minconf. It also can discover both positive and negative associations, and it uses a weight of evidence as conf measure. The quantitative values can be inferred from FARs. This allows interesting associations between different quantitative values to be revealed. The work in [20] provided a fuzzy transaction DM algorithm for extracting interesting knowledge from transactions stored as quantitative values. This algorithm integrates fuzzy-set concepts with the apriori mining algorithm and uses the result to find interesting itemsets and fuzzy association rules in transaction data with quantitative values. It transforms quantitative values in transactions into linguistic terms and then filters them to find association rules by modifying the apriori mining algorithm.

The work in [1] advised a method to mine FAR with weighted items. It developed measures of interestingness of discovered rules, and uninteresting large itemsets are pruned during the mining process. The FARM algorithm is extended to Apriori Gen large itemset based algorithm. It can be expressed by the following steps: Step 1: Map crisp values and fuzzy terms of each categorical attribute into consecutive integers. Step 2: Use the extended apriori algorithm to discover frequent item vectors where a frequent item vector has at least minimum weighted sup Step 3: From the frequent item vectors, generate all ARs that have at least minimum weighted conf.

Another work in [21], a fuzzy association rule-based classification method for high-dimensional problems is suggested to obtain an accurate and compact fuzzy rule-based classifier with a low computational cost. This method is based on the following three stages: 1) FAR extraction for classification: A frequent fuzzy itemsets are generated from the search tree and then generate FARs for classification. 2) Candidate rule pre-screening: in this stage, the authors used subgroup discovery based on an improved weighted relative accuracy measure; to decrease the computational cost and to preselect the interesting rules. 3) Genetic rule selection and lateral tuning: both techniques selection and tuning is presented to select and tune a compact set of FARs with high classification accuracy.

The study in [11] applied FARs to discover rules from the database in addition to prune the redundant rules extracted. They describe redundancy of FARs and display theorems concerning the redundancy of FARs. But their algorithm shows limitation in terms of computational time, and non-redundant rules are unexpectedly deleted. So, the authors see that this algorithm must enhance by applying the other method. The researchers in [22] proposed rare ARs from an educational database by utilizing the fuzzy-based apriori algorithm. This way is employed to less frequent itemsets by applying measure 'maximum sup' for providing rare items and to eliminate rare items and measure 'Rank' is used to prune the outliers from the rare items produced. The ARs can produce after the rare items produced.

Several works have been done in ARs using GA For instance; the study in [23] presented a GA for mining quantitative ARs. The rules obtained by genetic ARs extraction methods maintain a high conf and good coverage of the database, providing the user with high-quality rules. This study can get a reduced set of ARs, although the number of rules is restricted by the population size. In [24] introduced a GA based clustering method, which adjusts the

Journal of Theoretical and Applied Information Technology
31st January 2021. Vol.99. No 2

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

fuzzy sets to provide maximum profit and finding interesting weighted ARs from quantitative transactions based on user-specified linguistic min sup and minconf terms. This method uses two different fitness functions, maximizing the number of large itemsets and the average of the conf intervals of the generated rules.

In [25] a model of an intrusion detection system using fuzzy logic and GA based on DM techniques is presented. This model is based on the evolutionary design of intrusion detection systems. A General approach for intrusion detection is presented as following: Step 1: Identify the three parameters or features of the problem statement. Step 2: Classify the parameters or features depending on their uncertainty or crisp nature. Step 3: Once the parameters are classified, use fuzzy logic for modeling the uncertain parameters or features. Step 4: The crisp values can be modeled using statistical distributions depending on their classifications. Step 5: Normalize these modeled values. Step 6: Use clustering techniques and association rules for grouping them suitably. Step7: Identify a suitable model based on mathematical technique genetic algorithms or any other mathematical techniques for solving a problem. Step 8: Solve the problem using the above mathematical technique.

The academic in [26] presented multi-objective GA to generate high quality ARs. The authors combined four metrics for fitness function. These metrics are conf, completeness, interestingness and comprehensibility. The fitness function is calculated as the weighted average confidence, completeness, interestingness and comprehensibility. In [27] applied apriori hybrid and GA to mine the frequent patterns. Hybrid apriori to obtaining frequent patterns, and then ARs created by the apriori algorithm is optimized using GA. GA operators like selection, crossover and mutation applied to produce strong rules.

AIM OF THE WORK: Due to the vital challenge for the ARM algorithm, which is the tremendous number of derived rules that could be redundant and have not new information; the traditional fuzzy algorithm can face some difficulties. From this point, we use the study in [11] and implement the T2FARM to produce efficient ARs to prune redundant rules obtained from the apriori algorithm for DM application and identify frequent itemsets in addition to improve non-redundant rules to forecast strong rules by using GA. Results illustrated that such techniques could be worked effectively in DM, minimizing the effects of typical shortcomings of the FLS that is less able to pruning redundant rules.

3. PROPOSED SYSTEM

To overcome the above-mentioned problem this work integrated model that adopts T2FARM in addition to GA. our technique use T2FARM to reduce the redundant rules extracted from the apriori algorithm, to identify frequent itemsets that satisfy minsup and minconf values and also to enhance the non-redundant rules by applied GA. Our goal is to maximize the accuracy of the results that happened by combines an apriori algorithm with T2FLS and GA for a real DM application. The schema of the proposed system is shown in **Error! Reference source not found.**. Below, each phase is explained in details.

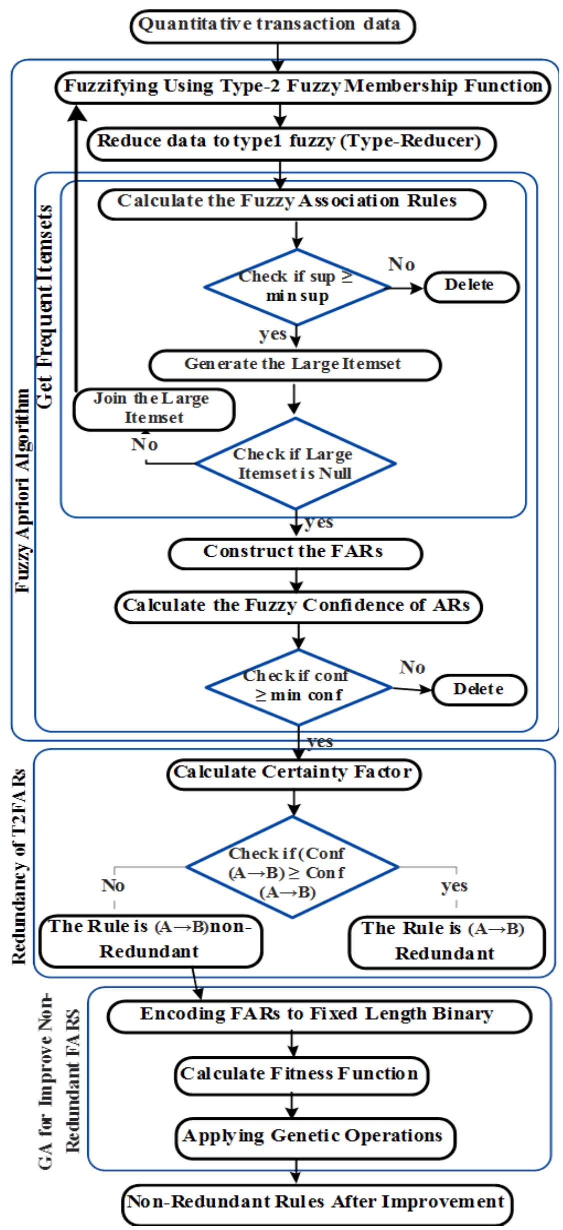

Figure 1: The Proposed System

3.1 Quantitative Transaction Data

We apply the dataset, popularly known as "Adult" data from the UCI machine learning repository. The Adult data set contains 32,659 records with 8 attributes (age, marital status, occupation, relationship, sex, native-country, education degree, and income). 60% of the data are used for training, and 40% of data are used for testing.

3.2 Fuzzifying Using Type-2 Fuzzy Membership Function

T2FLS model and minimizes the effects of uncertainties and vagueness in complex systems, because it has the capability of handling a higher level of uncertainty. T2 fuzzy set (FS) is characterized by a fuzzy membership function (MF) that is three dimensional. According to this type the membership value for each element of this set is a FS in [0, 1], instead of a crisp number like T1FS where representing uncertainty by membership value in [0, 1]. The three dimensional of T2 MF is (see Figure 2): upper membership functions (UMF), lower membership functions (LMF) and the area between these two functions is footprint of uncertainty (FOU), that give additional degrees of freedom that make it feasible to handle uncertainties. Therefore, T2FLS model and minimizes the effects of uncertainties which cannot be handled by T1FLS [28] [29].

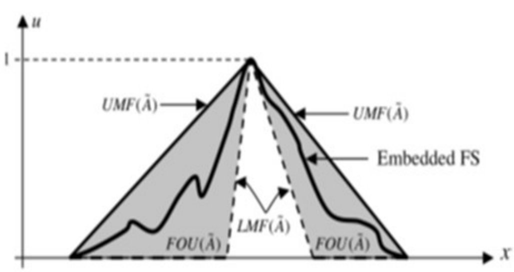

Figure 2: UMF and LMF representing the FOU

T2FS is concerned with quantifying using natural language in which words can have vague meanings. We operated to transforms the quantitative values v_{ij} of each transaction $T_i = (i = 1,2,\dots,n)$ for each item I_j into fuzzy values UMF f_{ijl}^{upper} and LMF f_{ijl}^{lower} using T2 MF for each R_{jl} where R_{jl} for each item is the l-th fuzzy region of item I_j (see equation 1, 2). These items are fuzzified using type-2 Triangular Fuzzy Number (TFN), provided to decide the degree of membership of items in the apriori algorithm as exposed in Figure 3. In this work, we utilize three attributes from database "age", "income", and "education degree" to apply T2FS on them and describe them by T2 Triangular MF that is defined as [30]. Each of these attributes is partitioned into four linguistic sets, the "age" variable takes values in database from [10, 70], it could be partitioned into four linguistic sets such as young [10, 30], youth [20, 40], middle age [30, 50], and old [50, 70] as presented in Table 1. Table 2 shows the four linguistic sets for variable "income" is low [0, 30], medium [10, 50], high [30, 70], and very high [50, 90]. Finally the variable "education degree" is divided into high school [0, 10], bachelor [5, 15], master [10, 20], and PhD [15, 25] as provided in Table 3.

$$LMF = \begin{cases} h(x+a)/a, & \text{if } -a \le x \le 0 \\ h(a-x)/a, & \text{if } \quad 0 \le x \le a \\ 0, & \quad\text{otherwise} \end{cases} \tag{1}$$

$$UMF = \begin{cases} h(x+b)/b, & \text{if } -b \le x \le 0 \\ h(b-x)/b, & \text{if } \quad 0 \le x \le b \\ 0, & \quad\text{otherwise} \end{cases} \tag{2}$$

where, $0 \le a \le b$ and $0 \le h \le 1$. All of the MF' parameters are numerically identified rely on the experiences.

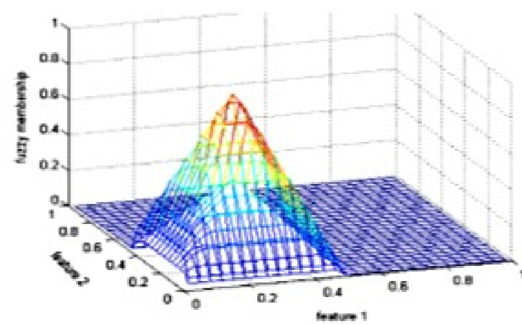

Figure 3: Triangular T2FMF.

Journal of Theoretical and Applied Information Technology
31st January 2021. Vol.99. No 2

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

Table 1: Age linguistic set and TFN

No.	Age linguistic Term	TFN
1	Young	(10,20,30)
2	Youth	(20,30,40)
3	middle age	(30,40,50)
4	Old	(50,60,70)

Table 2: Income linguistic set and TFN

No.	Income linguistic Term	TFN
1	Low	(0,10,30)
2	Medium	(10,30,50)
3	High	(30,50,70)
4	Very high	(50,70,90)

Table 3: Education degree linguistic set and TFN

No.	Education degree linguistic Term	TFN
1	High school	(0,5,10)
2	Bachelor	(5,10,15)
3	Master	(10,15,20)
4	PhD	(15,20,25)

In this stage, after produces a T1 FS output by COS type-reduction method, the apriori DM is related to finding frequent itemsets and discovers interesting ARs. The apriori algorithm requires two thresholds: minsup and minconf. These two thresholds identify the association that must hold before the rule will be mined. The steps of the apriori algorithm are given as follows:

3.4.1 Get frequent itemsets

The objective of this step is to find the itemsets with sup higher than or equal the minsup. These itemsets that meet the minsup threshold is a frequent itemset. To produce the frequent itemsets is as fol1ows [31]:

a. Compute the sup of ARs for each R_{jl} in the transactions. Thus sup_{jl} is the sum of f_{ijl} in transactions.

b. Check if sup_{il} ≥ minsup, If the value of sup_{il} is equal or larger than the minsup value, then this sup_{il} is a frequent and put it in the large 1-itemsets (L_1). If L_1 is not null, then do the next steps otherwise, exit the algorithm.

c. Set r = 1, where r is used to represent the number of items in the current itemsets to be processed.

d. Join the large r-itemsets L_r to generate the candidate (r +1) itemsets

e. Do the following sub steps for each newly formed (r +1) itemset

   1) Compute $f_{i\,il}^{upper}$ and $f_{i\,il}^{lower}$ by using type-2 MF
   2) Reduce each $f_{i\,il}^{upper}$ and $f_{i\,il}^{lower}$ of each item I_j to fuzzy value.
   3) Compute fuzzy value of each item by using the minimum operator for the intersection
   4) Compute the sup of ARs and then Check if sup_{il} ≥ minsup, If the value of sup_{il} is equal or larger than the minsup value, then put it in $Lr+1$, if $Lr+1$ is null then do the next step; otherwise, set r = r + 1 and repeat STEPs d to e.

3.4.2 Construct the FARs

After getting on fuzzy frequent itemsets, we capable of generates FARs. In our case, we generate 250 FARs from fuzzy frequent itemsets, examples of these FARs are:

- IF age is young, THEN income is high.
- IF age is youth, THEN income is low.
- IF income is high, THEN age is young.
- IF age is youth, and income is low THEN education degree is high school.

Where age, *income*, and *education degree* are sets of attribute and *young, youth, low, high school*, and *high* are linguistic labels.

3.4.3 Calculate the fuzzy confidence of ARs:

By getting FARs, we calculate the Conf of a rule A→B is defined as:

$$Conf_{(A\to B)} = \frac{supp_{fuzzy}(A \cup B)}{supp_{fuzzy}(A)} \quad (3)$$

where $\left(Conf_{(A\to B)}\right)$ is the percentage of the

Journal of Theoretical and Applied Information Technology
31st January 2021. Vol.99. No 2

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

fuzzy number of transactions that contain $A \cup B$ to the total fuzzy number of transactions that contain A. Then, we compare the conf of a rule with the minconf to get the satisfied rules that its conf is higher than or equals the minconf [5].

3.5 Redundancy of T2FARs

ARM finds an association among items based on the sup and the conf of the rule. However, these measures have many limitations as the database grows larger and larger. Thus the mined rules increase faster, and many redundant rules are extracted, and it becomes difficult to find interesting rules. For this reason, in this thesis, we use the certainty factor (CF) to measure the redundancy of extracted T2FARs. CF was developed to measure the uncertainty in the FARs. The value of the certainty factor is a number from -1 to 1. The CF of the rule $X \rightarrow Y$ is described as follows [8]:

$$CF(X \rightarrow Y) = \frac{Conf(X \rightarrow Y) - supp(Y)}{1 - supp(Y)}, \quad if\ Conf(X \rightarrow Y) > supp(Y) \tag{4}$$

$$CF(X \rightarrow Y) = \frac{Conf(X \rightarrow Y) - supp(Y)}{supp(Y)}, \quad if\ Conf(X \rightarrow Y) \leq supp(Y) \tag{5}$$

$$CF(X \rightarrow Y) = 1, \quad if\ supp(Y) = 1 \tag{6}$$

$$CF(X \rightarrow Y) = -1, \quad if\ supp(Y) = 0 \tag{7}$$

- **Theorem 1:** A→B and A`→B be two FARs:

$$\textbf{IF}\ Conf(A \rightarrow B) \geq Conf(A` \rightarrow B), \quad \textbf{THEN}\ Cf(A \rightarrow B) \geq Cf(A` \rightarrow B)$$

Ex: IF *Conf (age is young → income is low)* ≥ *Conf (education-degree is high school → income is low*) THEN *Cf (age is young → income is low)* ≥ *Cf (education-degree is high school → income is low).*

- **Theorem 2:** combine FARs: we consider A→C, B→C and A,B→C be two FARs, where A, B,

  and C are fuzzy itemsets.

$$\textbf{IF}\ \max\left(Conf(A \rightarrow C), Conf(B \rightarrow C)\right) \geq Conf(A, B \rightarrow C),$$
$$\textbf{THEN}\ \max Cf\left((A \rightarrow C), Cf(B \rightarrow C)\right) \geq Cf(A, B \rightarrow C)$$

low), Conf (education degree is high school → income is low)) ≥ *Conf (age is young & education degree is high school → income is low*) THEN max (*Cf (age is young → income is low), Cf (education degree is high school → income is low))* ≥ *Cf (age is young & education degree is high school → income is low*).

- **Theorem 3:** we consider A→B FAR, where A and B are fuzzy itemsets. Let fuzzy itemset

-

- family Q,

$$\textbf{IF}\ \max_{Y \in Q}\left(Conf(Y \rightarrow B)\right) \geq Conf(A \rightarrow B),$$
$$\textbf{THEN}\ \max_{Y \in Q}\left(Cf(Y \rightarrow B)\right) \geq Cf(A \rightarrow B)$$

degree is master) → income is low) ≥ *Conf (age is young → income is low)* THEN max (*Cf (age is young & education degree is high school*) ,or (*age is youth & education degree is master*) *→ income is low)* ≥ *Cf (age is young & income is low).*

From these theorems, we now able to extract redundant rule and non-redundant of FARs in terms of a joining of FARs to prune redundant rules by considering $A \rightarrow B$ FAR, where A and B are fuzzy itemsets based on $Q = 2^x - X - \emptyset$.

$$\textbf{IF}\ \max_{Y \in Q}\left(Conf(Y \rightarrow B)\right) \geq Conf(A \rightarrow B), \textbf{THEN}\ \text{the rule } A \rightarrow B \text{ is a redundant rule.}$$

r

r

a

$$\textbf{IF}\ \max_{Y \in Q}\left(Conf(Y \rightarrow B)\right) < Conf(A \rightarrow B), \textbf{THEN}\ \text{the rule } A \rightarrow B \text{ is non-redundant rule.}$$

is stronger than any corresponding redundant rule.

3.6 Genetic Algorithm

In this phase, after applying the previous theorems, we now utilize GA to improve the non-redundant FARs. GA is global search techniques with the ability to investigate a large search space for suitable solutions only requiring a performance measure. In addition to their capacity to discover near optimal solutions in complex search spaces, the generic code structure and independent performance features of GAs qualifies them to incorporate a priori knowledge.

GA starts with implement encoding to non-redundant FARs as a bit-string by converts each non-redundant FAR into an intermediary representation and then utilizes GA operators to choose a bit-string rule for "mating" to enhance FARs.

3.6.1 Encoding FARs:

In this stage, non-redundant FARs is encoding to produce an initial population. First, each FAR is translated into intermediary representation to retain its syntactic and semantic constraints. Each intermediary representation is comprised of N attributes in the antecedent part and one attribute in the consequent part. Each attribute in the antecedent part is encoded into a fixed-length binary to make an

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

individual in the initial population where each bit denotes a possible linguistic set. Similarity, the attributes in the consequent part is encoded into a fixed-length binary with each bit indicating a possible linguistic set. The advantage of binary encoding is the simple encoding and decoding operation. If some attributes in the attribute set are not used by non-redundant FARs, dummy tests are applied in the antecedent part to display these attributes by put underlined under these attributes.

Each non-redundant FAR denotes one individual in the initial population. For instance, as we mentioned earlier that, the age, income, and education degree are sets of attribute. The "age" attribute has four linguistic sets {young, youth, middle-age, old}, four linguistic sets for variable "income" {low, medium, high, very high}, and four linguistic sets for variable "education degree" {high School, bachelor, master, PhD}. Assume some non-redundant FARs is expressed as the following:

- FAR1: If *income is medium* THEN *age is youth.*

- FAR2: If *age is young* THEN *education degree is high School.*

The intermediary representations of these rules are be formed as follows:

- FAR1`: If *income is medium, and (education degree is High School, or education degree is Bachelor, or education degree is Master, education degree is PhD*) THEN *age is youth.*

- FAR2`: If *age is young, and (income is low, or income is medium, or income is high, or income is very high)* THEN *education degree is high School.*

Then we encode each attribute in these rules into a fixed-length binary string where each bit describes to a possible linguistic set. The encoded of rules declared in Table 4 and Table 5:

Table 4: The encoded of non-redundant FAR_1`

Non-redundant FAR_1`	Antecedent part		Consequent part
Attributes	Income	Education degree	Age
linguistic set	0100	1111	0100
Resulted rule	010011110100		

Table 5: The encoded of non-redundant FAR_2`

non-redundant FAR_2`	Antecedent part		Consequent part
Attributes	age	income	Education degree
linguistic set	1000	1111	1000
Resulted rule	100011111000		

3.6.2 Fitness Function

The main goal of a fitness function is to measure the quality of the non-redundant FARs. In this work, each chromosome in the population is assessed by the predictive accuracy to assess the quality of the rules. It is based on a confusion matrix to present correctly and incorrectly prediction as displayed in Table 6.

Table 6: Confusion matrix

	Actual Positive	Actual Negative
Positive Prediction	True Positive (TP)	False Positive (FP)
Negative Prediction	False Negative (FN)	True Negative (TN)
	Total Positive	Total Negative

For example, assume the rule: If x then y applied to measure Predictive accuracy to discover percentage of records having predicted consequent y that is covered by rule antecedent x as given in the following form:

$$Accuracy = \frac{number\ of\ correct\ prediction}{total\ number\ of\ prediction} = \frac{TP+TN}{TP+FP+TN+FN} \quad (8)$$

y,

FP: Number of records satisfactory x but not y;
TN: Number of neither records nor satisfactory x nor y;
FN: Number of records not satisfying x but satisfactory y.

3.6.3 Genetic Operators

- **Selection**: GA organizes the non-redundant FARs based on this fitness function, i.e., the rule owning the highest accuracy is at the top. Then the GA selects the top rules to do crossover and mutation.
- **Crossover**: is utilized to transfer the genetic information of two parents to produce new offspring (new generation). In this work, we utilize a single point crossover. It starts randomly picking the same points for both parent chromosomes, and then crossover is applied to copy these points from one parent to the second parent. For illustration, we used above non-redundant FARs is displayed as following:

- FAR1: If *income is medium* THEN *age is youth*.
  - ➢ Chromosome1: 010011110100

- FAR2: If *income is low* THEN *education degree is high School.*

- Chromosome2: 100011111000

After crossover:

New chromosome1: 010011111000

New chromosome2: 100011110100

Finally, we can get on new rules by decoding the new off-springs that obtain from crossover operator as described as the following:

- New FAR1: If *income is medium* then *education degree is high School.*
- New FAR2: If *income is low* then *age is youth*.

- **Mutation**: After a crossover is performed, mutation is implemented as displayed in the next form. It randomly changes chosen bits from 0 to 1 or from 1 to 0. The purpose is to maintain diversity within the population. Mutation changes randomly the new offspring.

- New FAR1: If *income is medium* then *education degree is high School.*

After mutation:

New chromosome1: 010011111000

New chromosome`1: 000111110100

Finally, after applying GA, we can get the non-redundant rules after improvement.

4. RESULTS AND DISCUSSION

In this section, we do some experiments that have been carried out to check the performance of the proposed system and assert improvements over the traditional approach. A program is implemented using MATLAB to assess the performance of the proposed system. We used “Adult” database from the UCI machine learning.

In the first experiment, we display the capacity of the proposed technique in eliminating redundant ARs compared to traditional FARs by using different minconf from 0.1 to 0.9 along with minsup 0.2 [32] due to the conf threshold value has a larger effect on the number of FARs when smaller minsup values are used. We can see that the number of the extracted rules dependent on the minconf whenever the number of ARs decreases with the increase of the minconf.

Table 7 show the relationship between all extracted rules and number of deleted redundant rules by traditional FARs. The results presented that the traditional FARs extracted many rules and many non-redundant rules (see Figure 4). So, our issue is to limit the extraction of redundant rules.

Table 7: Extracted rules and eliminating redundant rules by traditional FARs

	Minconf	No. of Extracted Rules	No. of Non-Redundant Rules
Traditional FARs	0.1	250	244
	0.2	244	240
	0.3	238	230
	0.4	230	223
	0.5	227	220
	0.6	222	216
	0.7	216	210
	0.8	200	195
	0.9	197	190

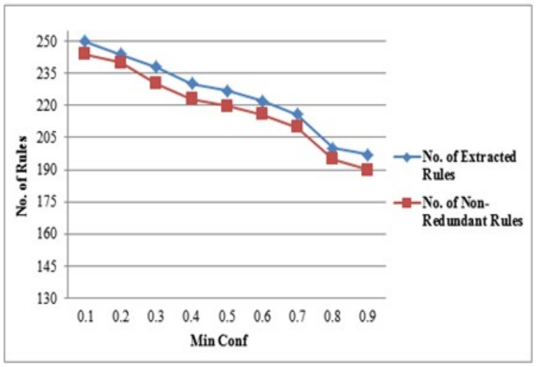

Figure 4: Extracted rules and eliminating redundant rules by traditional FARs

Figure 5 and Table 8 present all extracted rules from proposed system and number of deleted redundant rules by the proposed system. The results presented the proposed system extracts more interesting rules, most of the extracted rules can satisfy with large values of minconf and in addition to proposed technique ability to delete redundant rules compared to traditional system.

Table 8: Capacity of the proposed system in eliminating redundant rules

	Minconf	No. of Extracted Rules	No. of Non-Redundant Rules
Proposed System	0.1	248	237
	0.2	240	228
	0.3	236	215
	0.4	228	210
	0.5	221	200
	0.6	218	188
	0.7	211	180
	0.8	195	168
	0.9	182	150

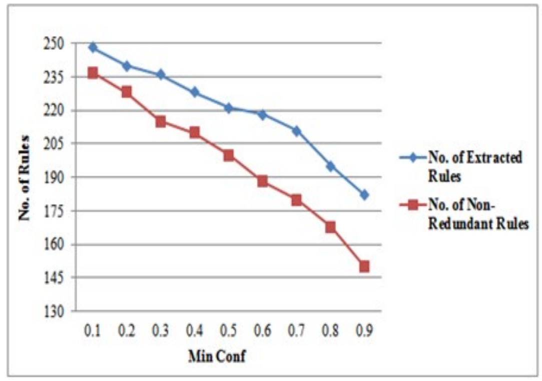

Figure 5: Capacity of the proposed system in eliminating redundant rules

The system with traditional fuzzy in terms of pruning redundant ARs. Results demonstrated that many redundant rules are extracted from traditional FARs as well as the redundant rules are effectively deleted by the proposed algorithm. So, the proposed system is more accurate to remove redundant rule than a traditional FARs (see Figure 6).

Table 8: Capacity of the proposed technique in eliminating redundant rules compared to traditional FARs

Minconf	Traditional FARs	Proposed System
0.1	244	237
0.2	240	228
0.3	230	215
0.4	223	210
0.5	220	200
0.6	216	188
0.7	210	180
0.8	195	168
0.9	190	150

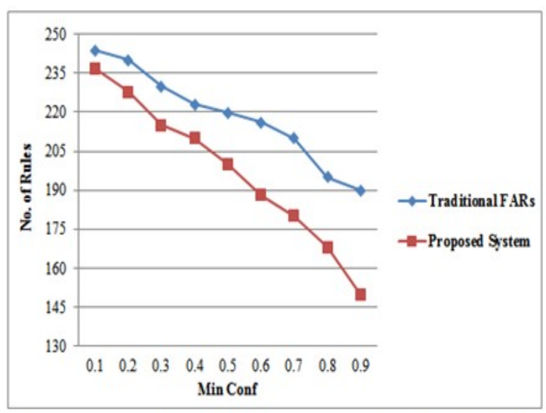

Figure 6: Capacity of the proposed technique in eliminating redundant rules compared to traditional FARs

In the second experimental, the execution time is used to compute time for the proposed system and traditional FAR as displayed in Table 9 by using high minconf 0.9 where most of the extracted rules can satisfy with large values of minconf. We can see that the result of the proposed system is corresponds to highest time. This is because the computations of T2FLS are highly complicated computations but it able to delete the redundant rules better than traditional FAR as shown in first experiment.

Table 9: Computation time

AR System Approaches	Time in Sec.
Fuzzy-approach	0.099
Proposed System	1.194

In third experiment, the proposed system displays its capacity to improve the non-redundant T2FAR by using GA. we use the formula of the fitness function using the following configuration mutation rate = 0.4, crossover rate = 0.9 [33], the initial population size ranges from 30 to 100, generation no. ranges from 10 to 30. The results are displayed in Table 10

Table 10: Proposed system evaluation to improve the non-redundant FAR by using GA using fitness function

Generation No. \ Initial population size	30	50	70	100
10	74.76%	84.41%	84.41%	84.41%
20	89.54%	92.23%	92.23%	92.23%
30	96.30%	98.52%	98.52%	98.52%

The results shows the ability of the proposed system in improving the non-redundant T2FAR according to the fitness function, the fitness function is used to select individuals in the population to generate new individuals. Among the new individuals and the original individuals in the population, those with high fitness values are passed to the new generation. As shown in Table 4.5, with a limited population, most non-redundant T2FAR could be improved in our dataset. We can decide that if the population is too small, the realization of the GA-based algorithm will be similar to the random algorithm. But if the population is too large, the computational complexity rises fast. However, as we can see, there is a good stability that, with a restricted population and a limited time period, most non-redundant T2FAR have been improved. Therefore, we choose 50 as the default population for the dataset, which works well in our algorithm. Also as we shown, when the number of generations increased, the resulting fitness value also increased. The iterations are stopped, when the highest fitness values is reached. The best solution in our proposed system to improve the non-redundant T2FAR by using GA achieved a fitness value 98.52% after 30 generation.

Difference form prior work:

- The problem of traditional FARs is to find a tremendous amount of extracted rules; some of these rules might be redundant, have not new information and it becomes hard to find interesting rules. So, our issue is to limit the extraction of redundant rules.
- To overcome the above problem this work adopts T2FARM in addition to GA. Our technique use T2FARM to reduce the redundant rules extracted from the apriori algorithm and also to enhance the non-redundant rules by applied GA.
- The proposed system extracts more interesting rules addition to proposed technique ability to delete redundant rules compared to traditional system. So, proposed system is more accurate to prune redundant rule than a traditional FARs.
- So, our results presented that many redundant rules are extracted from traditional FARs as well as the redundant rules are effectively deleted by the proposed algorithm. So, the proposed system is present good performance and more accurate to remove redundant rule than a traditional FARs.
- Finally, the proposed system enhances non-redundant FARs using GA according to the fitness function.
- But, the proposed system takes highest time that's because the computations of T2FLS are highly complicated computations and GA takes a long time to optimize machining rules compared to traditional FARs.

5. CONCLUSION

In this article, The submitted approach is based on the hybridization between ARs, and T2FLS to know frequent itemsets and identify ARs from these frequent itemsets that achieve minsup and minconf values in addition to minimize the redundant rules mined from the apriori algorithm. Moreover, we apply GA to develop non-redundant ARs derived from T2FARs.

T2FLS is beneficial in cases where it is incapable of determining an exact certainty and measurement uncertainties. It represents inT2 fuzzy sets that permit to fuzzifying the utilized data to discover the FARs and then we applied CF and redundancy theorem to measure the redundancy of extracted T2FARs and to reduce the redundant FARs. Finally, we also used GA to optimize non-redundant rules that generated from T2FARM by implement encoding to non-redundant FARs as a bit-string by converts each non-redundant FAR into an intermediary representation and then utilized GA operators to choose a bit-string rule for "mating" to enhance FARs and foretell strong rules.

Our goal from the suggested method is to maximize the accuracy of the results by pruning redundant rule and improve non-redundant rules that happened by combines an apriori algorithm with T2FLS and GA. The experimental results show that the proposed approach provides better performance in eliminate redundant rules when compared with traditional FARs. Our experimental results on a real-world data set confirm that the proposed approach has great potential in terms of accuracy where it is more accurate in remove redundant rule and enhance non-redundant FARs using GA achieved a fitness value 98.52% after 30 generation but it takes highest time around 2.988 second that's because the computations of T2FLS are highly complicated computations and GA takes a long time to optimize machining rules.

REFRENCES

[1] J. Shu Yue, E. Tsang, and E. Tsang, "Mining Fuzzy Association Rules with Weighted Items", *proceeding of 2000 IEEE International Conference on Systems, Man, and Cybernetics*, USA, 8-11 Oct. 2000, pp. 1906- 1911.

[2] K.S. Deepashri, A. Kamath, "Survey on Techniques of Data Mining and it's Applications", *International Journal of Emerging Research in Management &Technology,* Vol. 6, Issue 2, 2017, pp. 198-201.

[3] M. Delgado, N. Marín, D. Sánchez and M. Vila, "Fuzzy Association Rules: General Model and Applications", *IEEE Transactions on Fuzzy Systems,* Vol. 11, No. 2, April 2003, pp.214- 225.

[4] Y. Zhao and T. Zhang, "Discovery of Temporal Association Rules in Multivariate Time Series", *proceeding of International Conference on Mathematics, Modeling and Simulation Technologies and Applications (MMSTA 2017),* 2017, pp. 294-300.

[5] *B*. L. Helm, "Fuzzy Association Rules an Implementation in R", *Master's thesis, Vienna University of Economics and Business Administration,* 2007.

[6] S. M. Darwish, A. A. Amer, and S. G. Taktak, "A Novel Approach for Discovery Quantitative Fuzzy Multi-Level Association Rules Mining Using Genetic Algorithm", *International Journal of Advanced Research in Artificial Intelligence (IJARAI),* Vol. 5, No.6, 2016, pp. 35-44.

[7] B. Sowan, K. Dahal, M. A. Hossain, L. Zhang, and L. Spencer, "Fuzzy Association Rule Mining Approaches for Enhancing Prediction Performance", *Expert Systems with Applications,* Vol. 40, Issue 17, Dec. 2013, pp. 6928-6937.

[8] G. Suganya, and K. J. Paulraj Ananth, "Analysis of Association Rule Mining Algorithms to Generate Frequent Itemset", *International Journal of Innovative Research in Science, Engineering and Technology,* Vol. 6, Issue 8, Aug. 2017, pp. 15571- 15577.

[9] Y. Xu, Y. Li, and G. Shaw, "Concise Representations for Approximate Association Rules", *proceeding of IEEE International Conference on Systems, Man and Cybernetics,* Singapore, 2008, pp. 94- 101.

[10] T. Watanabe, "Mining Fuzzy Association Rules of Specified Output Field", *proceeding of IEEE International Conference on Systems, Man and Cybemetics.* 2004, pp. 5754- 5759.

[11] T. Watanabe, "Fuzzy Association Rules Mining Algorithm Based on Output Specification and Redundancy of Rules", *proceeding of IEEE International Conference on Systems, Man and Cybemetics,* 2011, pp. 283- 289.

[12] F. Herrera, "Genetic Fuzzy Systems: Status, Critical Considerations and Future Directions", *International Journal of Computational Intelligence Research*, Vol.1, No.1, 2005, pp. 59-67.

[13] Q. A. Al-Radaideh, E. M. Al-Shawakfa, and M. I. Al-Najjar, "Mining Student Data Using Decision Trees", *International Arab Conference on Information Technology (ACIT'2006)*, 2006, pp.1-5.

[14] M. Durairaj, and K. Meena, "A Hybrid Prediction System Using Rough Sets and Artificial Neural Networks", *International*

Journal of Innovative Technology & Creative Engineering, Vol. 1, No. 7, July 2011, pp. 16-23.

[15] C. Romero, P. G. Espejo, A. Zafra, J. R. Romero, and S. Ventura, “Web Usage Mining for Predicting Final Marks of Students That Use Moodle Courses”, *Journal of Computer Applications in Engineering Education,* Vol.27, Issue 3, 2013, pp. 135- 146.

[16] Y. Huang, and Ch. Wu, “Mining Generalized Association Rules Using Pruning Techniques”, *proceeding of IEEE International Conference on Data Mining*, 2002, pp.227-234.

[17] N. S. Rai, S. Jain, and A. Jain, “Mining Interesting Positive and Negative Association Rule Based on Improved Genetic Algorithm (MIPNAR_GA)”, *(IJACSA) International Journal of Advanced Computer Science and Applications*, Vol. 5, No. 1, 2014, pp. 160-165.

[18] L. Antonie, and O. R. Zaïane “Research and Application of Algorithm for Mining Positive and Negative Association Rules”, *proceeding of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2004.

[19] W. Au, and K. C.C. Chan, “FARM: A Data Mining System for Discovering Fuzzy Association Rules”, *Proceeding of 1999 IEEE International Fuzzy Systems (FUZZ-IEEE'99)*, South Korea, pp. 1-7.

[20] T. Hong, Ch. Sh. Kuo, and Sh. Ch. Chi, “Mining association rules from quantitative data”, *Journal of Intelligent Data Analysis*, Vol. 3, Issue 5, Nov.1999, pp. 363-376.

[21] J. 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, Issue 5, Oct. 2011, pp. 857 – 872.

[22] A. M. Rajeswari, M.Sridevi, and C.D. Chelliah, “Outliers Detection on Educational Data using Fuzzy Association Rule Mining”, *Int. Conf. on Adv. in Comp., Comm., and Inf. Sci. (ACCIS-14)*, 2014, pp. 1-9.

[23] G. Shankar, and L. Bargadiya, “A New Improved Apriori Algorithm for Association Rules Mining”, *International Journal of Engineering Research & Technology (IJERT)*, Vol. 2 Issue 6, June – 2013, pp. 2523- 2527.

[24] M. Kaya, and R. Alhajj, “Utilizing Genetic Algorithms to Optimize Membership Functions for Fuzzy Weighted Association Rules Mining”, *Applied Intelligence*, Vol. 24, No. 1, Feb. 2006, pp. 7-15.

[25] G. V. S. N. R. V. Prasad, Y. Dhanalakshmi, Dr. V. V. Kumar, and Dr I. R. Babu, “Modeling an Intrusion Detection System Using Data Mining and Genetic Algorithms Based On Fuzzy Logic”, *IJCSNS International Journal of Computer Science and Network Security*, Vol.8 No.7, July 2008, pp. 319- 325.

[26] J.Malar Vizhi and Dr. T.Bhuvaneswari, “Data Quality Measurement With Threshold Using Genetic Algorith”", *International Journal of Engineering Research and Applications*, Vol. 2, Issue 4, July-August 2012, pp. 1197- 120.

[27] Ch. Babi, Dr. M. Venkateswara Rao, and Dr. V. Venkateswara Rao, “Mining Frequent Patterns from Big Data Sets using Genetic Algorith”, *International Journal of Engineering Research and Technology*.Vol. 11, No. 2, 2018, pp. 287-306.

[28] Y. Chen, and L. Yao, “Robust type-2 fuzzy control of an automatic guided vehicle for wall-following”, *Proceedings of the International Conference of Soft Computing and Pattern Recognition*, 2009, pp. 172-177.

[29] S. M. Darwish, T. F. Mabrouk, and Y F. Mokhtar, “Enriching Vague Queries by Type-2 Fuzzy Orderings”, *Lecture Notes on Information Theory*, Vol. 2, No. 2, June 2014, pp. 177-185.

[30] M. Jerry, “Type-2 Fuzzy Sets and Systems: an Overview”, *IEEE Computational Intelligence Magazine*, Vol. 2, No. 1, Feb. 2007, pp. 20-29.

[31] Ch. H. Chen, T. P. Hong, and Y. Li, “Fuzzy Association Rule Mining with Type-2 Membership Functions”, *Asian Conference on Intelligent Information and Database Systems*, March 2015, pp.128-133.

[32] J. Fdez, R. Alcala, M. Gacto, and F. Herrera, “Learning the membership function contexts for mining fuzzy association rules by using genetic algorithms”, *Journal of Fuzzy Sets and Systems*, Vol. 160, 2009, pp. 905-921.

[33] Ch. H. Wang, T. P. Hongb, Sh. Sh. Tseng, “Integrating membership functions and fuzzy rule sets from multiple knowledge sources”, *Fuzzy Sets and Systems*, Vol. 112, 2000, pp. 141-154.