A Novel Approach for Discovery Quantitative Fuzzy Multi-Level Association Rules Mining Using Genetic Algorithm

Saad M. Darwish

Department of Information
Technology
Institute of Graduate Studies and
Research University of Alexandria
Alexandria, Egypt

Abeer A. Amer

Department of Computer and
Information Systems
Sadat Academy for Management
Sciences (SAMS)
Alexandria, Egypt

Sameh G. Taktak

Department of Computer and
Information Systems
Sadat Academy for Management
Sciences (SAMS)
Alexandria, Egypt

Abstract—Quantitative multilevel association rules mining is a central field to realize motivating associations among data components with multiple levels abstractions. The problem of expanding procedures to handle quantitative data has been attracting the attention of many researchers. The algorithms regularly discretize the attribute fields into sharp intervals, and then implement uncomplicated algorithms established for Boolean attributes. Fuzzy association rules mining approaches are intended to defeat such shortcomings based on the fuzzy set theory. Furthermore, most of the current algorithms in the direction of this topic are based on very tiring search methods to govern the ideal support and confidence thresholds that agonize from risky computational cost in searching association rules. To accelerate quantitative multilevel association rules searching and escape the extreme computation, in this paper, we propose a new genetic-based method with significant innovation to determine threshold values for frequent item sets. In this approach, a sophisticated coding method is settled, and the qualified confidence is employed as the fitness function. With the genetic algorithm, a comprehensive search can be achieved and system automation is applied, because our model does not need the user-specified threshold of minimum support. Experiment results indicate that the recommended algorithm can powerfully generate non-redundant fuzzy multilevel association rules.

Keywords—Quantitative Data Mining; Fuzzy Association Rule Mining; Multilevel Association rule; Optimization Algorithm

I. INTRODUCTION

In data-mining, discovering association rules in transaction databases is frequently examined. Association rules are widely offered and are beneficial for planning and marketing. For example, they can be managed to implicate supermarket officials of what products the customers have an inclination to purchase together. Taking market basket analysis as an example, the mining problem can be explained as given a database *D* of transactions, each transaction is a set of items; find all rules that relate the carriage of one set of items with that of another set of items [1].

The classical algorithms for mining association rules are formed on binary attributes databases, which have two weaknesses. Firstly, it cannot treat quantitative attributes; secondly, it handles each item with the same weight despite that strange item may have different importance. Also, a binary association rule bears from explicit boundary problems. Besides many real world transactions consist of quantitative attributes. That is why numerous researchers have been serving on the generation of association rules for quantitative data [2] [3].

Beginning approaches for quantitative association rule mining manages distinctive partitioning for transforming other attributes to binary ones which convey from a major problem that results in information damage because of sharp limits. In other words, modern algorithms neglect or exaggerate items beside the boundary. When distributing an attribute in the data into sets comprising individual ranges of values, the users are faced with the sharp boundary problem[2]. Quantitative attributes are discretized by joining the concept of hierarchies. This manner occurs before the event of mining. For example, a concept hierarchy for age may be adopted to reconstruct the initial numeric values of this attribute by ranges [2]. To surmount this problem researchers are laboring for mining association rules for quantitative attributes. They have contributed several algorithms that tackle quantitative algorithm and reveal how they dispense with quantitative data [3] [4].

In general, fuzzy technique overcomes the main drawback of the discretize technique. Fuzzy logic produces linguistic term instead of intervals which is more nearer to the human mind. The disadvantage is that although the loss of information is small but it exists. Furthermore, the needs for fuzzy membership function to be given by an expert, which is not always straightforward and can be biased. Despite that, fuzzy association rule mining approach appeared out of the requirement to mine quantitative data uniformly present in databases efficiently [2]. There are two essential basic criteria for association rules, support(*s*) and confidence(*c*). Since the database is large and users interest about only those frequently purchased items, usually thresholds of support and confidence are predetermined by users to separate those rules that are not so attractive or beneficial. The two thresholds are called minimal support and minimal confidence respectively [5].

Genetic Algorithm (GA) is a heuristic exploration that imitates the process of natural evolving. This heuristic is

routinely applied to produce valuable explications to optimization and search obstacles. Genetic Algorithm is based on conceptions of evolution hypothesis as a fundamental policy is that only the strongest beings remain. The genetic algorithms are significant when identifying association rules because they work with the global search to determine the set of items frequency and they are less complex than other algorithms frequently worked in data mining. The genetic algorithms for discovery of association rules have been settled into usage in real problems such as commercial databases and fraud detection [6].

Earlier investigations on data mining directed on locating association rules at a single-concept level. Mining association rules at multiple concept levels may guide to the discovery of more broad and significant knowledge from data. Related item taxonomies are normally predefined in real-world purposes and can be interpreted as hierarchy trees. Terminal nodes on the trees express actual items looking in transactions; internal nodes describe classes or concepts built from lower-level nodes [7]. A simple example is given in Fig. 1. Mining multi-level association rules are motivated by several purposes, such as: (a) the multi-level association rules are more reasonable and are more interpretable for users. (b) The multi-level association rules can supply us solutions for the undesirable and undesired rules. Encouraging applications involve spatial data analysis, emergency event analysis, and network data mining [8] [9].
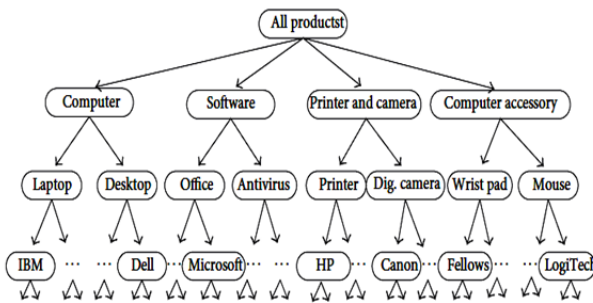

Fig. 1. The predefined taxonomy

A. Motivation & Rationale

If it arrives at quantitative association rule mining, a number of trials have been directed on performance (speed) and effectiveness (number of rules). Less effort has been converged on quality. Modern quantitative multilevel association rule mining algorithms depend on intense looks of the database to obtain regular exemplars beyond different abstraction levels [9]. However, these mining algorithms are often based on the postulate that users can blueprint the minimum support relevant to their databases [10]. Mining quantitative association rules is not a manageable enlargement of mining categorical association rules. Since the search space is unlimited, our aim is to detect a measurable set of exciting solutions (quantitative rules), near to the optimal answers. This illustrates why we have decided to solve this search problem with meta-heuristics routines, mainly genetic algorithms [11].

Regularly, when managing quantitative association rule mining several rules can be identified or inferred and confuse the user. But more importantly, some of these rules could be redundant and produce no new knowledge. Some attempt has been aimed at selling with redundant rules in flat datasets. However, datasets can have a hierarchy/taxonomy or multiple concept levels and thus redundancy in these datasets require to be adjusted. This subject is one of the phases of this research. Currently, the approach being taken is to resolve which rules are redundant and eliminate them, thus diminishing the number of rules a user has to deal with while not decreasing the information content [12]. This Paper offers an adapted version of the Apriori algorithm for mining fuzzy multi-level association rules in large databases for locating familiar item set at distinct levels of abstraction [13].

B. Research Problems

The purpose of this study is to offer to the field of data mining and in precise to multilevel quantitative fuzzy association rule mining. Hence to attain this distinct and well-defined investigation, difficulties are needed. By answering these it is likely to provide in an essential style. For our research, there are many central difficulties that we will converge on and strive to solve. These difficulties are:

- Boolean attributes can be studied as an exceptional instance of categorical attributes and it is an almost applicant to generalize the Boolean data mining algorithms. For qquantitative attributes, despite, the state is not so easy. We either have to somehow convert the quantitative association rules problem into Boolean one or to get different algorithms. Here we shall, in fact, produce an approach to discover quantitative association rules stemmed from a dataset with multiple concept levels.

- The number of rules expands exponentially with the number of items. But this complexity is undertaken with some advanced algorithms which can efficiently cut the search space [1][4]. The work picking this problem principally assists the user when scanning the rule set. Yet, the evolution of further valuable quality measures on the rules through employing genetic algorithm with fitness function (relative confidence of the association rule) to affirm the most intriguing association rules signifies the advanced aims to solve this problem [14].

- Without a priori knowledge, however, ascertaining the right intervals (discretize) for quantitative data mining can be a complex and intricate task. Moreover, these intervals may not be compact and acceptable enough for human experts to quickly gain nontrivial knowledge from those created rules. Fuzzy membership function can help to advise this problem.

- Completely and efficiently identifies no redundant association rules from datasets with a hierarchy.

C. Problem Statement

Market data in real-world usually involve quantitative values, so creating an advanced data-mining algorithm equipped to contract with quantitative data grants a challenge to workers in this study domain [4]. The multilevel association rules mining problem can be defined as follows: there are a

collection of items $I = \{i_1, i_2, ..., i_n\}$ and Γ is a classification tree that concisely clarifies the multilevel categorizing relations among items as the field awareness. i_1 is the ancestor of i_2 and i_2 is the descendant of i_1 if there is an edge in Γ from i_1 to i_2. Only leaf nodes are displayed in the database. D is a database of transactions where each transaction T in D is a set of items such that $T \subseteq I$. Each transaction is attached with an identifier T_{ID}. Let P indicates the set of positive integers, and I_v symbolizes the set $I \times p$. A couple $<x, v> \in I_v$ means the quantitative attribute x, with the related value $v.I_v$ $\{\forall <x, l, u> \in I \times p \times p / l \le u\}$, l is the lower limit and u is the upper limit of p. A triple $<x, l, u> \in I_v$ stands for a quantitative x with a value in the interval $[l, u]$. Note that a transaction T holds an item $x \in I$ if x is in T or x is an ancestor of some items in T. In addition, a transaction T involves $X \subseteq I$ if T bears every item in X.

A multilevel association rule is an inference of the form $X \Rightarrow Y$, where $X \subseteq I, Y \subseteq I$, and $X \cap Y = \phi$. No item in Y is an ancestor of any item in X; that is $Y \cap \text{ancestors}(x) = \phi$. This is because a rule of the style "$x \Rightarrow \text{ancestors}(x) = \phi$" is slightly true with 100% confidence, which is redundant. Both X and Y can contain items from any level of Γ [9][15][16].

However, there are still some limitations in quantitative association rule mining, such as [3][4][10]: (1) separation of the quantitative attribute, which is adopted in the design, is not accessible for all attribute and every user. (2) users, and even experts, regularly believe difficult to provide those thresholds like the minimum support, the interest level, and the minimum confidence.(3) the search space might be very large when we contact with quantitative attributes. Finally, (4) the rules declared by the algorithm might be too many to manage with.

II. LITERATURE REVIEW

In this section we compare the quantitative association rule mining algorithms taking into account the form of the rules and discuss each technique advantages, disadvantages and what kind of database can be used[17][18]:

1) Discretization:

The elementary intention of this routine is to transform quantities data to Boolean by examining a separation of the numerical attributes into collections of intervals. Then, an algorithm for detecting Boolean association rules can be handled to prepare quantities rules. Two main representations of partitions are included. A fixed partition, where the assortments of intervals are disjoint and another type, where the ends of intervals are overlaid with each other.

The principal benefit of this technique, beyond being the first work done on this track, is that manipulating both categorical and numerical data correspondingly. However, situations (disjoint or overlapped) yield problems; disjoint sets damage from *Min_Sup* and *Min_Conf* thresholds and

overlapped sets suffer from the cutting boundary problem. Using intervals rather than the real continuous data will inevitably result in a loss of information. The rules we make will be only an estimation of the best results. Another problem is the enlargement of the attributes dimension; the problem here is the need for more memory and time to treat these data.

2) Adjusted difference analysis:

This algorithm is based on engaging both adjusted difference analysis and discretization to discover rules between two attributes. The two attributes could be any mixture of numerical or categorical. This technique has the capability to identify positive and negative association rules and does not need any user thresholds (support and confidence). Its advantages are that it does not want any user thresholds and it has the talent to obtain a new significant objective measure of the association rules. The disadvantage of this technique is the problems of discretization as in the first technique. Also, this technique is obviously considered to be as generating a special case rule since the generated rules are always between two attributes only.

3) Fuzzy Approach based on integrating fuzzy set and fuzzy logic concepts with Apriori algorith:

It reforms numerical data into fuzzy member between [0,1] with membership function; then operate with the fuzzy member with an adjusted Apriori technique that can manipulate comfortably the extracted rules, which are stated in linguistic terms. These approaches are based on the fuzzy additions to the classical association rules mining by establishing support and confidence of the fuzzy rule. While the mining results are straightforward to interpret by human operators, two shortcomings still insist on implementing such fuzzy approaches to the original problems. One is the computational time for mining from the database, and the other is the precision of deduced rules. More formal description, as well as a survey of the existing methods of quantitative association rule mining, can be found in [20].

In the literature, several researchers have concentrated on fuzzy multilevel association rules mining [3][21-25]. Some of these methods evoked multilevel membership functions by ant colony systems and genetic algorithm without stipulating the actual minimum support. To improve the performance of computing, setting the functions for each item followed by calculating minimum supports is engaged. Other work carried benefits of the OLAP and data mining technology which conducted efficiency and adaptability [9].

Up-to-date, there exist only a few algorithms for quantitative multilevel fuzzy association rule mining (*QMLFRL*). For examples, in [4] the authors advised an *QMLFRL* based on the idea that the minimum support for an item at a higher taxonomic concept is valued as the minimum of the minimum supports of the items pertaining to it and an item minimum support for an itemset is established as the maximum of the minimum supports of the items enclosed in the itemset. Under this limitation, the characteristic of downward-closure is conserved, such that the original Apriori algorithm can be simply prolonged to find fuzzy large item sets.

With the same purpose, the authors in [26] suggested a new method of quantitative association rule extraction that can quantize the attribute by applying a clustering algorithm and learn rules simultaneously. They implemented clustering using all attributes at the same time in advance and deduced the rules from the clusters in the aspect of "association". Based on the numerical experiments, the authors have confirmed that their algorithm outperformed the conventional algorithm based on Cartesian product type quantization in terms of total precision of quantization and rule extraction.

Extra relevant work introduced in [6] to create rules based on the quantitative dataset, utilizing the notion of threshold - frequent item sets that are produced using the genetic algorithm. In this illustrations, crossover & mutation are involved to create numerous unification of the rule and can recognize co-occurrence of item sets. Here three objectives are studied: comprehensibility, interestingness, confidence, so produced rules are established as multi-objective association rules. These objectives serve to decrease search space for fitness function. Finally, optimal rules are formed that is based on distribution approach for the numeric-valued attribute (Right-hand side of a rule reveals the distribution of the values of numeric attributes such as the mean or variance).

The benefit of the preceding systems is that they carry linguistic expressions which make created rules to be much normal for human experts; but they may generate a large number of interesting association rules. Still, for many purposes, it is not periodically simple to ascertain effective association rules (matching the minimum support and confidence) among data items at low (primitive) levels of abstraction due to the sparsity of data in multidimensional space. Other associated problems cover: (1) the shortage of sufficient support for dynamically needed hierarchies; (2) algorithm efficiency cannot meet real application specifications; (3) the association between different concepts levels may be dropped; (4) Their approach enabled users to stipulate various minimum supports to different items. [4][27].

A. Research Contribution

The idea developed in this paper is partly inspired by the existing work on *QMLFRL*, buts it utilizes the genetic algorithm to compute the minimum support and minimum confidence for each level in the Taxonomy regardless of the nature of the data; thus making automatic system. Prior studies have completely investigated single-level association rules mining with GA, such as mining single objective rules and mining multi-objective rules. However, in the big data analysis setting, powerful association rules are regularly in multilevel forms and mining multilevel association rules in big data demands more efficient methods. The GA-based multilevel association rules mining method recommended in this paper is one effort to efficiently discover multilevel association rules in big data.

III.   THE PROPOSED MODEL

The advanced mining algorithm combines fuzzy set notions, data mining, and multiple-level taxonomy to determine fuzzy association rules in a given transaction data set deposited as quantitative values. The knowledge obtained is described by fuzzy linguistic terms, and thus simply readable by human beings. This system utilizes a top-down progressively deepening strategy to locate large itemsets [4]. In this paper, we made our primary intention toward automatically detecting minimum support and minimum confidence of each taxonomy' level by constructing a genetic algorithm based heuristic method for practical multilevel association rules mining in big datasets. By using the advantage of the genetic algorithm, which can efficiently ascertain multiple solutions concurrently in a large multidimensional problem without conducting exhaustive searches, our offered method can enhance the mining performance while preserving the wanted accuracy but bypassing the exhausting list of association rule candidates. Definitions linked to multilevel association rules are presented as follows [9][15][16]:

Definition 1: an item set, X , is a set of data items $\{X_i, X_j\}$, where $X_i, X_j \in I$. The support of an item set X in a set S, $\sigma(X/S)$ is the number of transactions (in *S*) which covers X against the total number of transactions in *S*. The confidence of $X \Rightarrow Y$ in S, $\varphi(X \Rightarrow Y/S)$, is the fraction of $\sigma(X \cup Y/S)$ in competition with $\sigma(X/S)$, i.e., the possibility that item set Y takes place in *S* when item set occurs X in *S*.

Definition 2: An item set X is large in set *S* at level L if the support of X is no less than its matching minimum support threshold σ'_L . The confidence of a rule $X \Rightarrow Y/S$ is high at level L if its confidence is no less than its equivalent minimum confidence threshold φ'_L .

Definition 3: a rule $X \Rightarrow Y/S$ is strong if $X \cup Y/S$ is large at the existing level and the confidence of $X \Rightarrow Y/S$ is high at the current level.

Definition 4: A fuzzy transaction denoted by *T* is given by:

$$\overline{\overline{T}} = \{(x, \mu(x)) \mid \forall x \in I\} , \ 0 \le \mu(x) \le 1, \mu : I \to [0,1], \overline{\overline{T}} \subseteq T$$

where *T* is a general set of transactions, and $\mu(x)$ is degree of membership of *x*.

Definition 5: A soft quantitative transaction set that is symbolized by T'_q . Let (*F, E*) is a soft set over the universe *U* and $X \subseteq E$, *F* means the fuzzy power set of *U*, and *E* is a set of parameters. A set of attributes *X* is said to be supported by a transaction if:

$$T'_q = \left\{(< x, l, u >, e) \parallel \forall < x, l, u > \in I \times p \times p \mid l \le u, e \in E\right\}$$

Mining of association rules essentially focalizes at a single conceptual level. There are applications which lack to locate associations at multiple abstract planes. In a large database of transactions, where each transaction consists of a set of items and a taxonomy (is-a hierarchy) on items, it is expected to find out associations between items at any level of taxonomy. To investigate multilevel association rule mining, anyone wants to afford data at multiple-level association at multiple levels of abstraction and efficient methods for multiple level rule

mining. The first specification can be accomplished by producing concept taxonomies from the primitive level concepts to higher levels. The second condition dictates efficient methods for multilevel rule mining [15].

One modification of Apriori to multi-level datasets is the ML_T2L1 procedure [15][24]. The ML_T2L1 algorithm manages a transaction table that has the hierarchy information encoded into it. Each level in the dataset is treated separately. Firstly, level 1 (the highest level in the hierarchy) is examined for large 1-itemsets using Apriori. The list of level 1 large 1-item sets is then employed to refine and clip the transaction dataset of any item that does not have an ancestor in the level 1 large 1-itemset list and eliminate any transaction which has no common items (thus comprises only infrequent items when evaluated using the level 1 large 1-itemset list). From the level 1 large 1-itemset list, level 1 large 2-itemsets are concluded (using the cleaned dataset). Then level 1 large 3- item sets are inferred and so on until there are no more frequent item sets to find at level 1. Since ML_T2L1 specifies that only the items that are descendant from frequent items at level 1 (essentially they must descend from level 1 large 1-itemsets) can be frequent themselves, the level 2 item sets are concluded from the refined transaction table. For level 2, the large 1-itemsets are created, from which the large 2-itemsets are determined and then large 3-itemsets etc. After all the frequent itemsets are found at level 2, the level 3 large 1-itemsets are located (from the same purified dataset) and so on. ML_T2L1 reforms until either all levels are explored using Apriori or no large 1-itemsets are exposed at a level. The principal steps of the proposed system are as follows [8-10][14][24][29-31]:
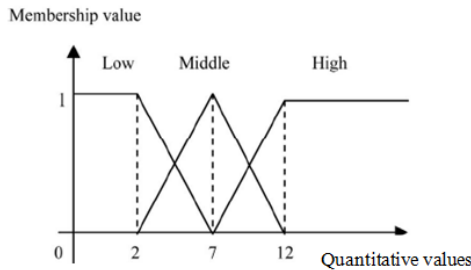

Fig. 2. The membership functions of items in Γ

Input: A group of N quantitative transaction data D, a predetermined catalog Γ with the original items $\{i_1, i_2, ..., i_n\}$, a set of membership functions for each item in deferent levels. In our case, all the membership functions have the same style as shown in Fig. 2; but the x-axis is determined for each element in Γ based on the higher quantitative value associated with it. Finally, the parameter set minimum support α_k and minimum confidence λ_k that are acquired by genetic algorithm.

Output: A collection of fuzzy multiple-level association rules below the restrictions of optimal minimum support and confidence.

Step 1: Translate the predefined taxonomy using an arrangement of numbers and the symbol "*" by the formula, $C = \rho * 10 + i$, where i is the position number of the node at

current level l, C signifies the code for the ith node at current level and ρ is the code of parent of the ith node at the present level.

Step 2: Interpret the item terms in the transaction data agreeing to the encoding scheme. Then set $k = 1$, $r = 1$ where k, $1 \leq k \leq x$ is the recent level number, x is the number of level in a given taxonomy and r denotes the number of items kept in the current frequent item sets.

Step 3: Cluster the items with the same first k digits in each transaction D_i, and add the quantities of the items in the similar sets in D_i. Symbolize the total of the j-th group I_j^k for D_i as v_{ij}^k.

Step 4: We explored several membership function for various data items for that each data item has its own features and its own membership function, then transform the value v_{ij}^k of each transaction D_i for each encoded group I_j^k into a fuzzy set f_{ijl}^k (Eq.1) by plotting v_{ij}^k on the specified membership function, where I_j^k is the j-th item on level k, v_{ij}^k is the quantitative value of I_j^k in D_i, h_j^k is the number of fuzzy areas for I_j^k, R_{jl}^k $(1 \leq l \leq h_j^k)$ is the l-th fuzzy region of I_j^k, f_{ijl}^k is v_{ij}^k fuzzy membership value in R_{jl}^k

$$\left(\frac{f_{ij1}^k}{R_{j1}^k} + \frac{f_{ij2}^k}{R_{j2}^k} + ... + \frac{f_{ijh}^k}{R_{jh}^k} \right). \tag{1}$$

Step 5: Assemble the fuzzy regions (linguistic terms) with membership values larger than zero to create the candidate set C_1^k; Calculate the scalar cardinality S_{jl}^k of each fuzzy region R_{jl}^k in the transaction data as $S_{jl}^k = \sum_{i=1}^{n} f_{ijl}^k$.

Step 6: Investigate if the value S_{jl}^k of each region R_{jl}^k in C_1^k is larger than or equals to the threshold α_k which is the optimal minimum support for level k obtained from implementing the genetic algorithm to the set of transactions included in this level according to Γ (see algorithm 1). If R_{jl}^k matches the threshold, place it into the large 1-itemset L_1^k for level k. That is:

$$L_1^k = \left\{ R_{jl}^k \big| S_{jl}^k \geq \alpha_k, R_{jl}^k \in C_1^k \right\}. \tag{2}$$

Step 7: if L_1^k is null, let $k = k + 1$ and go to step 3; else, create the applicant set C_2^K from $L_1^1, L_1^2, ..., L_1^K$ to catch "level-crossing" large itemsets. The created applicant set C_2^K has to fulfill the following conditions: (1) Each 2-itemset in C_2^K must comprise at least one item in L_1^k. (2)The two regions in a 2-itemset may not have the same item name. (3) The two item names in a 2-itemset may not be with the hierarchy relation in the taxonomy. (4) Both of the support values of the two large

1–itemsets including a candidate 2-itemset must be larger than or equal to the minimum support $\alpha_{k=2}$.

Step 8: If L_1^k is null, then increase k by one, $r =1$ and go to step3 else set $r=r+1$.

(a) If $r = 2$ create the candidate set C_2^K, where C_2^K is the set of candidate itemset with 2 items on level k from $L_1^1, L_1^2, L_1^3,..., L_1^k$ to learn "level-crossing" of frequent itemset. Each 2-itemset in C_2^k must contain at least one item in the L_1^k and the next item should not be its ancestor in the taxonomy. All possible 2-itemsets are composed in C_2^k.

(b) If $r > 2$, produce the candidate set C_r^k, where C_r^k is the set of candidate itemset with r-items on level k from L_{r-1}^k in a way similar to that in the preceding steps.

Step 9: For each acquired candidate r-itemset S with items $(S_1, S_2,....S_r)$ in C_r^k:

a) Calculate the fuzzy value of S in each transaction datum D_i by the minimum operator as $f_{is} = \min(f_{is1}, f_{is2},..., f_{isr})$

b) Estimate the scalar cardinality of S in all the transaction data as $count_s = \sum_{i=1}^{n} f_{is}$.

c) If $count_s$ is larger than or equals to the pre-defined minimum support α_k place S into L_r^k.

Step 10: If L_r^k equal null then increase K by one and go to the next step; if not increase r by one and go to step 8.

Step 11: If $k > x$ then go to the next step, else set $r =1$ and go to step 3.

Step 12: create the fuzzy association rules for all frequent r- itemset including $S = (S_1, S_2,...., S_r)$, $r >2$ as follows:

- Catch all the rules $A \rightarrow B$ where $A \subset S$, $B \subset S$ and $A \cap B = \phi$, $A \cup B = S$.

- Calculate the confidence value of all association rules by $\dfrac{\sum_{i=1}^{n} min(f_{is})}{\sum_{i=1}^{n} min(f_{iA})}$.

Step 13: Choice the rules that have confidence values not less than predefined confidence threshold λ_k, where λ_k is the predefined minimum confidence value for level k found by applying genetic algorithm.

Step 14: eliminate redundant rules from multi-level datasets. Herein, Rule R_1 is redundant to rule R_2 if (1) the itemset X_1 is made up of items where at least one item in X_1 is descendant from the items in X_2 and (2) the item set X_2 is entirely made up of items where at least one item in X_2 is an ancestor of the items in X_1 and (3) the other non-ancestor items

in X_2 are all present in item set X_1. The additional state (4) the confidence of $R_1(C_1)$ is less than or equal to the confidence of $R_2(C_2)$.

A. Parameters extraction using genetic algorithm

A genetic algorithm is a class of investigating algorithm that is employed to automatically set the optimal minimum support and minimum confidence for each taxonomy's level. It explores a solution space for an optimal answer to a problem [28]. The algorithm generates a "population" of feasible solutions to the problem and makes them "evolve" over many generations to locate valid and better solution. The algorithm begins with a collection of solutions (represented by chromosomes) called a population. Solutions from one population are selected and managed to establish a new population. The framework of the basic genetic algorithm is as follows (see Fig. 3).

```
Procedure genetic algorithm
begin (1)
    t = 0;
    inicialize P(t);
    evaluate P(t);
    While (Not termination-condition) do
    begin (2)
        t = t + 1;
        select P(t) from P(t − 1);
        recombine P(t);
        evaluate P(t);
    end (2)
end (1)
```

Fig. 3.   Structure of the genetic algorithm [29]

1) [Start] create arbitrary samples of n chromosomes (appropriate results for the problem)

2) [Fitness] assess the fitness (qualification) function f(x) of every chromosome x in the population

3) [New population] generate a new resident by iterating the subsequent steps until the new population is complete.

Selection: pick two parent chromosomes from a population according to their fitness (the better fitness, the higher possibility to be chosen)

Crossover: with a crossover probability, crossover the parents to produce a new generation (children). If no crossover was conducted, offspring is an accurate reflection of parents.

Mutation: with a mutation probability, the GA mutates a new generation at each location (site on the chromosome).

Accepting: store distinct generation in a new population.

4) [Replace] manage recently produced population for a more route of the algorithm

5) [Test] if the end condition is satisfied, stops, and returns the best solution in current population

6) [Loop] go to step 2.

Through repetition t, the GA keeps a population $p(t)$ of results $r_1^t,...., r_N^t$, where r_i^t characterizes rule set that is

arbitrarily created for each level. Each solution r_i^t is gauged by the function $E(\bullet)$ and $E(r_i^t)$ is a degree of fitness of the solution. The fitness value determines the relevant power of an individual to remain and create offspring in the next production. In the next iteration $(t+1)$ a new resident is designed on the foundation of the operations (2) and (3) [29].

B. Data Encoding

Given a randomly generated association rules for each level, the system uses the Michigan approach for encoding, in which a chromosome is a collection of all used rules; here the population consists of many rule collections. Coding in the Michigan method is binary coding, in which "1" means that a knowledge base rule will be in a knowledge base, whereas "0" means it will not be used. The key benefit of this technique is that the entire rule base is coded; therefore, it is not necessary to do the quantitative analysis of indispensable rules to see if the method functions properly, because, unlike the Pitts method, all possible rules take part in the working time of the genetic algorithm. The considerable size of the chromosome is a disadvantage. The dimension of the chromosome is dependent on the volume of the rule base and it increases exponentially depending on the number of itemsets [10][14][30].

C. Generic Operators

The frequently employed genetic operators are reproduction, crossover, and mutation. To achieve genetic operators, one must pick individuals in the population to be worked on. The collection plan is mainly based on the fitness level of the individuals exhibited in the population. For election; the system manages roulette wheel sampling fashion. In this procedure, the parents for crossover and mutations are chosen based on their fitness, i.e. if a candidate has more fitness function value more will be its opportunity to get elected. The implementation of roulette wheel sampling is performed by first normalizing the values of all applicants so that, their chances sprawl between 0 and 1, and then by applying random number function, a random number is estimated, and then matching to this value and the fitness normalized value, the candidate is elected [14].

As an individual is picked, reproduction operators only imitate it from the current population into the new population (i.e., the new generation) without transposition. The crossover operator begins with two selected individuals and then the crossover point (an integer between 1 and L-1, where L is the length of strings) is picked arbitrarily. The third genetic operator, mutation, offers random variations in the arrangements in the population, and it may irregularly have useful results: departing from a local optimum. In our GA, mutation is just to oppose every bit of the strings, i.e., changes a 1 to 0 and vice versa, with probability p_m [30].

The algorithm stops fulfilling when the decay situation is reached − i.e. when the best and worst producing chromosome in the population disagrees by less than 0.1%. It also ends execution when the total number of generations defined by the user has arrived. Besides, the algorithm bypasses forming the initial population completely randomly because it may appear in rules that will include no training data instance whereby having very low fitness. Furthermore, a population with rules that are insured to comprise at least one training instance can lead to over−fitting the data. It was shown that employing non−random initialization can reach to an elevation in the quality of the solution and can drastically decrease the runtime [24]. We, therefore, devised an initialization method which involves picking a training instance to serve as a "seed" for rule generation based on the alteration of itemsets within each level [31].

In general, genetic operator assists in controlling the heterogeneity of the population and also in blocking early concurrence to local optima [14]. Our intention is to explore fascinating association rules. Consequently, the fitness function is vital for ascertaining the interestingness of chromosome, and it does influence the convergence of the genetic algorithm. In this case, the proposed system examines two different fitness functions. The first one considers the identical confidence of the corresponding association rule as illustrated in Eq. 3, whereas the second fitness function joins the support (*sup*) and *confidence* (*conf*) attributes, which are required to define an association rule (see Eq. 4) [9][10][14]. Parameters α and β are the significant factors to equilibrium the weight of the *support* and *confidence* in the fitness function, $\alpha + \beta = 1$. To mine confirmed association rules from the big database with our GA approach, the threshold of the fitness function has to be predefined; in our case, $\alpha = \beta = 0.5$.

$$f_1 : rconf(X \rightarrow Y) = \frac{\sup(X \cup Y) - \sup(X) \times \sup(Y)}{\sup(X)(1 - \sup(Y))} \qquad (3)$$

$$f_2(x \rightarrow y) = \alpha \times \sup(x \rightarrow y) + \beta \times conf(x \rightarrow y) \qquad (4)$$

By adopting the recommended system, rather of producing an untold number of interesting rules in conventional mining models, only the most interesting rules are declared according to the interestingness measure determined by the fitness function. The main motivation for using GA in the learning of high-level prediction rules is that they conduct a global search and cope better with attribute cooperation than the greedy rule selection algorithms [14].

In brief, the proposed evolutionary method for quantitative association rule mining is particularly prompted by (1) partition of quantitative attribute is not accessible for every attribute and every user, (2) users, and even experts, usually feel tedious to define the minimum-support, (3) the search space might be very large when we face quantitative attributes, and (4) the rules passed might be too many to deal with [10]. However, mining association rules are not adequate of benefits; it has some defects too, first of all, the algorithmic complexity. The number of rules increases exponentially with the number of items. But this complexity is undertaken with some advanced algorithms which can efficiently clip the search space. Secondly, the obstacle of attaining rules from rules, i.e. selecting interesting rules from the set of rules.

The suggested work undertakes the second problem that essentially assist the user when scanning the rule set, and valuable quality measures on the rules are adopted based on genetic algorithm. Usually, when handling association rule mining many rules can be found or inferred and confuse the

user. But more importantly, some of these rules could be redundant and yield no new knowledge. Some attempts have been pointed at dealing with redundant rules in flat datasets, however, datasets can have a hierarchy/taxonomy or compound concept levels and thus redundancy in these datasets require to be concentrated on. This issue is one of the features of this study.

IV. EXPERIMENTS AND RESULTS

In this section, we perform some experiments that have been conducted out to examine the performance of the proposed approach and confirm enhancements over the traditional method without optimization. The experiment was conducted with MATLAB software. All the experiments are handled on a laptop computer with the following specifications: Processor Intel(R) Core (TM) i5-2520MCPU@2.50GHz 2.50GHz, memory 4.00GB, and System type 64-bit operating system, x64-based processor.

A. Dataset

The dataset was hired as in [8] and can be viewed as a benchmark because it is used for comparison. This is a market basket dataset that consists of the items and amounts of items marketed in every purchasing container. This dataset consists of 1000 sales receipts of a food material repository based on the predefined taxonomy from 7 items (10000 transactions). The predefined taxonomy in the first level holds 7 nodes that describe the items worked in the test, the second level comprises 14 nodes that describe the taste or different types of a particular stock and in the third level, it also consists of 48 nodes that express the manufacturing companies and factories. The database transactions carry the name of the product and the quantity of such purchased merchandise. One item may not be employed twice in one transaction.

B. Methodology

A comparison was made between the conventional approach for mining multi-level fuzzy quantitative association rules [8] and the approach proposed in this paper that uses GA optimization technique. The objective was to find new detailed knowledge by (1) an enhancement in pronouncing multi-level optimal support and confidence that is employed to obtain interesting rules. (2) eliminating redundant rules that were encountered in the traditional approach. Both algorithms utilize a top-down progressively deepening approach to infer large itemsets and also consolidate fuzzy boundaries instead of explicit boundary intervals.

The mined rules from the proposed system are closer to the reality, and it gives the ability to mine association rules at different levels based on the optimal re-calculated mining parameters (*min_sup*, *min_conf*); unlike the traditional method that depends on the experts to determine these parameters manually. Employing GA to find these parameters makes the proposed system is context-independent and more general. In the experiments, thresholds for *min-sup* and *min-conf* were set at 0.28 and 1.7 respectively for the traditional algorithm for each taxonomy level.

In the first experiment, we examine if correct association rules can be specified with a fixed number of initial generations

and in a bounded time period. Using our dataset, the initial population size ranges from 30 to 100. The results are displayed in Table 1. With a limited population, most strong association rules could be inferred 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, despite we can get full association rules immediately, but 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 valid association rules have been mined. Therefore, we choose 50 as the default population for the dataset, which works well in our algorithm.

TABLE I. RELATION BETWEEN THE NUMBER OF INITIAL GA POPULATION AND THE NUMBER OF MULTILEVEL ASSOCIATION RULES (USING F1 WITH MICHIGAN ENCODING, GENERATION NO. =10)

No. of Initial population	30	50	70	100
No. of association rule (Redundant)	3706	607	607	607
No. of association rule (non-redundant)	1733	349	349	349

One reason of the stability of the number of extracted rules with initial population contains 50 chromosomes is that the proposed system utilizes Michigan approach for rule coding in which a chromosome is a collection of all used rules. So, the lowest number of initial populations will contain specific rules permutations. In Michigan approach, each chromosome contains a comprehensive representation of the rules.

In the next collection of experiments, we confirmed how worthy the extracted association rules are from either the GA-based algorithm or the traditional algorithm without GA. To measure its value with the 10000 transactions, we use the formula of the fitness function f_1 using the following configuration mutation rate = 0.1, crossover rate = 0.9, generation No.=10,and initial population=50. The results in dataset are shown in Table 2.

TABLE II. COMPARATIVE STUDY

Methods	No. of non-redundant rules	Calculated *min-sup*	Calculated *min-conf*	Time (Sec)
Proposed Method with GA	349	L1=0.95 L2=0.67 L3=0.27	L1=1.12 L2=1.45 L3=1.92	500
Traditional method without GA [8]	2281	0.28 (L=1 to 3)	1.7 (L=1 to 3)	400

TABLE III. COMPARATIVE STUDY BETWEEN THE TWO FITNESS FUNCTIONS

Fitness Function	levels	Computed Fitness	No. of association rules (non-redundant)
f_1 (Eq. 3)	level 1	1.90	349
	level 2	1.97	
	level 3	2.06	
f_2 (Eq.4) $\alpha = \beta = 0.5$	level 1	1.07	2249
	level 2	1.26	
	level 3	1.47	

TABLE IV.   PROPOSED SYSTEM EVALUATION UNDER DIFFERENT PARAMETERS OF GA USING F_1

Parameters ratio	No. of association rules (redundant)	No. of association rules (non-redundant)
Mutation= 0.9 crossover= 0.1	751	419
Mutation=0.8 crossover=0.2	751	419
Mutation=0.7 crossover=0.3	607	349

Compatible with the outcomes above, with the advance of the time boundary, GA-based approach can catch high relevant association rules with a little more time than traditional method (25% increases in time). But in terms of quality, the proposed system extracts more interesting rules; only about 17 % of the total number of rules extracted from the comparative system. In general, a large number of extracted rules inside market basket analysis will hamper the decision-maker. The proposed system offers the most interesting rules subject to the fitness function, which is accountable for acting the assessment that imitates how optimal the solution is: the higher the number, the better the solution.

The third set of experiments was executed to compare how relevant the mined association rules are from either the fitness function that considers the relative confidence of the corresponding association rule (Eq. 3) or the fitness function that considers both the support and confidence attributes (Eq. 4). The experiment is conducted under the previous configuration for GA. The results in Table 3 reveal that the use of f_1 generates a further mined association rules rate improvement of 83% reduction in the number of extracted rules. From this experiment, we realized that the fitness function represents a critical issue in the success of genetic algorithm; this is clearly shown in the case of f_2. Using f_2 did not bring any advantages to the GA; thus, we get the same number of extracted rules that have been obtained from the traditional method (about 2249).

The performance improvement of GA using f_1 comes from the correct extraction of interesting rules; because of calculating the support of union for items inside each rule in addition to the support of each item separately. Unlike the second function that uses the support and confidence of each rule, which represents the standard case used by many of the existing mining algorithms (e.g. Apriori).

Having compared our system to different multi-level quantitative mining algorithm, we will next explore the influence that GA factor settings have on our system, which incorporates both mutation rate and crossover rate. To retain the number of factor setting blending small, we will only fluctuate the setting for one parameter at a time while holding the setting for another parameter to its default value. In Table 4, we vary mutation rate from 0.7 to 0.9 and look at the number of association rules (non-redundant) achieved by our system on the used dataset. From the table, we can see that decreasing mutation rate will diminish the number of extracted rule (17% lower in the number of rules). This decreasing is noticeable. This decrease is due to the fact that mutation is managed to preserve genetic heterogeneity from one generation of a

population to the next. In GA, mutation operators are frequently employed to give exploration and crossover operators are extensively employed to supervise the population to focalize on one the good solutions encounter so far (exploitation). Consequently, while crossover attempts to concentrate to a special point in the landscape, the mutation does its best to evade convergence and investigate more areas.

V.   CONCLUSION

Really, mining quantitative association rules is an optimization obstacle rather than being an uncomplicated discretization one. In this paper, we have introduced a new genetic-based algorithm to mine multilevel association rules in big quantitative date sets that deals with quantitative attributes by accurate fuzzification the values -partitioning the values of the attribute. The proposed system uses fuzzy set concepts, multi-Level taxonomy, different pre-calculated minimum supports for each level and different membership function for each item to discover fuzzy association rules in a given transaction data set.

In our algorithm, the minimum supports and minimum confidences for each level for fuzzy quantitative association rule are defined by the genetic algorithm optimization. In this case, the employed GA combines a population initialization technique that guarantees the production of high- quality individuals; individually planned breeding operators that confirm the removal of inadequate genotypes; an adaptive mutation probability to ensure genetic heterogeneity of the population; and uniqueness testing based on both support and confidence that is employed to hold only high quality and interesting rules.

The proposed system gives the user with rules according to two interestingness metrics, which can quickly be extended if need by changing the fitness function. The results report that: In terms of mining of association rules, the proposed method keep higher precision compared with the traditional methods and the extracted rules are more close to reality. This is because of adopting various membership functions for every individual item, optimized minimum supports, and minimum confidences, and finally, the non-redundant algorithm to enhance the quality and application of the rules. Future work includes employing GA to tune the fuzzy membership function for each item.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques," Elsevier Inc., USA, 2012.

[2] T. Hong, Y. Tung, S. Wang, Y. L. Wu, and M. T. Wu, "A multi-level ant colony mining algorithm for membership functions, "Information Sciences, vol. 182, no. 1, 2012, pp. 3-14.

[3] A. Gosain, and M. Bhugra, "A comprehensive survey of association rules on quantitative data in data mining, " the IEEE Conference on Information and Communication Technologies, India, 2013, pp.1003-1008.

[4] Y. C. Lee, T. P. Hong, and T. C. Wang, "Mining fuzzy multiple-level association rules with multiple minimum supports, " Expert Systems with Applications, vol. 34, no. 1, 2008, pp.459-468.

[5] K. Poornamala, and R. Lawrance "A general survey on frequent pattern mining using genetic algorithm, " ICTACT Journal on Soft Computing, vol. 3, issue 1, 2012, pp.440-444.

[6]   D. Kanani, and S. Mishra, "An optimize association rule mining using genetic algorithm," International Journal of Computer Applications, vol. 119, issue 14, 2015, pp.11-15.

[7]   E. Mahmoudi, E. Sabetnia, M. Torshiz, M. Jalal, and G. Tabrizi, "multi-level fuzzy association rules mining via determining minimum supports and membership functions, " the Second International Conference on Intelligent Systems Modeling, and Simulation, Iran, 2013pp.55-61.

[8]   A. Kousari, S. Mirabedini, and E. Ghasemkhani, "Improvement of mining fuzzy multiple-level association rules from quantitative data," Journal of Software Engineering and Applications, vol. 5, no. 3, 2009,pp. 190-199.

[9]   X. Yang, M. Zeng, Q. Liu, and X. Wang, "A genetic algorithm based multilevel association rules mining for big datasets" Mathematical Problems in Engineering, Vol. 2014, 2014, pp.1-10.

[10] X. Yan, C. Zhang, and S. Zhang, "genetic algorithm-based strategy for identifying association rules without specifying actual minimum support, " Expert Systems with Applications vol. 36, no. 2, 2009, pp. 3066-3076.

[11] S. Aouissi, A. Vrain, and C. Nortet, "Quantminer: a genetic algorithm for mining quantitative association rules", the 20th International Conference on Artificial Intelligence , 2007, pp. 1035–1040.

[12] G. Shaw, Y. Xu, and S. Geva, "Utilizing non-redundant association rules from multi-level datasets" the IEEE International Conference on Web Intelligence and Intelligent Agent Technology, vol. 03, Australia, 2008, pp.681-684.

[13] P. Gautam, and K. R. Pardasani, "Algorithm for efficient multilevel association rule mining", International Journal on Computer Science and Engineering, vol. 2, no.5, 2010, pp.1700-1704

[14] S. Manish, A. Agrawal, and A. Lad, "Optimization of association rule mining using improved genetic algorithms," the International Conference on Systems, Man and Cybernetics, vol. 4, USA, 2004, pp.3725-3729.

[15] V. Ramana, M. Rathnamma, and A. Reddy, "Methods for mining cross level association rule in taxonomy data structures" International Journal of Computer Applications, vol. 7, no.3, 2010, pp.28-35.

[16] S. Saraf, N. Adlakha, and S. Sharma, " Soft set approach for mining quantitative fuzzy association patterns in databases, " International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, issue 11, 2013, pp.359-369.

[17] M. Delgado, N. Manín, M. J. Martín-Bautista, D. Sánchez, and M. -A. Vila" Mining fuzzy association rules: an overview,"Studies in Fuzziness and Soft Computing, Springer Berlin Heidelberg , vol. 164, 2005, pp.351-373.

[18] G. Attila, "A fuzzy approach for mining quantitative association rules, " ACTA CYBERN, vol. 15, no.2, 2001, pp.305-320.

[19] W. Toshihiko, and H. Takahashi, "A study on quantitative association rules mining algorithm based on clustering algorithm, " International Journal of Biomedical Soft Computing and Human Sciences, vol. 16, no.2, 2010, pp.59-67.

[20] R. Sridevi, and E. Ramaraj, "A general survey on multidimensional and quantitative association rule mining algorithms, " International Journal of Engineering Research and Applications, vol. 3, issue 4, 2013, pp.1442-1448.

[21] P. Gautam, N. Khare, and K. R. Pardasani," A model for mining multilevel fuzzy association rule in database, " Journal of Computing, vol. 2, issue 1, 2010, pp.58-64.

[22] O. Oladipupo, C. Ayo, and C.Uwadia, " a fuzzy association rule mining expert-driven Approach to Knowledge Acquisition," African Journal of Computing and ICT, vol. 5, no. 5, 2012,pp.53-60.

[23] S. Prakash, M. Vijayakumar, and R.Parvathi, " A Novel Method of Mining Association Rule with Multilevel Concept Hierarchy", International Journal of Computer Applications, vol. 5, no. 5, 2011, pp.26-29.

[24] G. Shaw, Y. Xu, and S. Geva, "Eliminating redundant association rules in multi-level datasets, " the 4th International Conference on Data Mining, USA, 2008pp.1-8.

[25] E. Ayetiran, and A. Adeyemo, "A data mining-based response model for target selection in direct marketing," International Journal Information Technology and Computer Science, vol. 4, no. 1, 2012, pp. 9-18.

[26] W. Toshihiko, and H. Takahashi, "A study on quantitative association rules mining algorithm based on clustering algorithm," Journal of the Biomedical Fuzzy Systems Association, vol.16, no.2, 2010, pp.59-67.

[27] Y. Wan, Y. Liang, and L. Ding, "mining multilevel association rules from primitive frequent itemsets," Journal of Macau University of Science and Technology, vol.3, issue 1, 2009,pp.10-19.

[28] R. Haldulakar, J. Agrawal, "Optimization of Association Rule Mining through Genetic Algorithm," International Journal on Computer Science and Engineering, vol. 3, no. 3, 2011, pp.1252-1259.

[29] M. Kayaa, R. Alhajj, "Genetic algorithm based framework for mining fuzzy association rules, " Fuzzy Sets and Systems, vol. 152, 2005, pp. 587–601.

[30] S. Tiwari, M. K. Rao" Optimization in association rule mining using distance weight vector and genetic algorithm , "International Journal of Advanced Technology & Engineering Research, vol. 4, Issue 1, 2014, pp.79-84.

[31] P. Wakabi-Waiswa,   and V. Baryamureeba, "Mining high quality association rules using genetic algorithms", In Proceedings of the twenty second Midwest Artificial Intelligence and Cognitive Science Conference, USA, 2009, pp. 73-78.