British Journal of Mathematics & Computer Science
4(23): 3294-3311, 2014
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Discovery of Novel Association Rules Based on Genetic Algorithms

Fadl Mutaher Ba-Alwi[1*]

[1] *Faculty of Computer and Information Technology, Sana'a University, P.O.Box 1247, Yemen.*

Article Information

DOI: 10.9734/BJMCS/2014/12256
Editor(s):
(1) Doina Bein, Applied Research Laboratory, The Pennsylvania State University, USA.
Reviewers:
(1) Anonymous, National Polytechnic Institute, Mexico.
(2) Anonymous, Mohammed V-Souissi University, Morocco.
(3) Anonymous, Chung-Hua University, Taiwan.
(4) Anonymous, University of Ljubljana, Slovenia.
(5) Anonymous, Hunan University, China.
Peer review History: http://www.sciencedomain.org/review-history.php?iid=669&id=6&aid=6201

Received: 24 June 2014
Accepted: 20 August 2014
Published: 24 September 2014

Original Research Article

Abstract

Association rule mining is a data mining task that attempts to discover interesting knowledge from huge databases. Data mining researchers have studied subjective measures of interestingness to reduce the volume of discovered rules to ultimately improve the overall efficiency of KDD process. Genetic algorithm (GA) based on evolution principles have found its strong base in mining association rules (ARs). In this paper, confidence and novelty measures have been pushed into a genetic algorithm in order to generate association rules form huge data and discover a novel and hence interesting knowledge to support decision makers. A hybrid approach that uses objective and subjective measures has been used in this paper to quantify novelty of association rules during generation process in terms of their confidence and deviations from the known rules.

The proposed approach has a flexible chromosome encoding involve Apriori algorithm where each chromosome should be compute its support and confidence values to performs prune process of week chromosomes. In addition each chromosome differs from another in terms of number of items and classes. The proposed approach has been experimented using real-life public datasets and tested using real life applications. The experimental results have been presented and quite promising.

General terms: Data mining, KDD, association rule, genetic algorithm.

*Corresponding author: dr.fadlbaalwi@gmail.com;

Keywords: Novelty measure, support, confidence, chromosome.

1 Introduction

In the last years information collection has become easier, but the effort required to retrieve relevant pieces of it has become significantly greater, especially in large-scale databases. Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration [1]. The term KDD refers to the overall process of knowledge discovery in databases. Data mining is a particular step in this process, involving the application of specific algorithms for extracting patterns (models) from data [2].

Discovering association rules is one of the most applications of data mining techniques. Mining for association rules between items in huge amounts of database of sales transactions has been recognized as an important area of database research. These rules can be effectively used to uncover unknown relationships, producing results that can provide a basis for forecasting and decision making. The original problem addressed by association rule mining was to find a correlation among sales of different products from the analysis of a large set of data [3].

Soft computing is a consortium of methodologies that works synergistically and provides, in one form or another, flexible information processing capability for handling real-life ambiguous situations [2]. Soft computing methodologies (involving fuzzy sets, neural networks, genetic algorithms, and rough sets) are most widely applied in the data mining step of the overall KDD process. Genetic algorithms (GAs) are involved in various optimization and search processes, like query optimization and template selection.

GA can encode potential solutions to a specific problem on a simple genes and chromosomes and apply recombination operators to these structures as to preserve critical information.

The genetic algorithms are important when discovering association rules because they work with global search to discover the set of items frequency and they are less complex than other algorithms often used in data mining. The genetic algorithms for discovery of association rules have been put into practice in real problems such as commercial databases, biology and fraud detection event sequential analysis [3].

In this paper, author present new approach based on genetic algorithms in order to improve the mining of association rule and to find comprehensible and novel knowledge from huge database transactions. Also, novelty of the discovered association rules as a subjective and objective measures of interestingness will be study. A huge database of shopping supermarket management will be used in this paper. First, the proposed approach will finding the frequent item set of items and after that operations of genetic algorithm will be applying on these frequent itemset. The proposed approach will determine that maximum frequent itemsets, generate association rules and applying the novelty measure based on specific threshold to find the novelty and interesting association rules in smaller size with acceptable degree of accuracy by which so many important decisions can be taken.

This is a useful feature as the volume of data keeps on escalating over the time and hence the user background knowledge is monotonically augmented. In addition, the proposed approach guarantees that a chromosome is pruned during constructing phase will certainly be pruned as the model is built and then pruned using novelty criterion. This strategy saves time and effort required to build the model.

2 Materials and Methods

2.1 Previous Work

There are many proposals that studied the novelty and interestingness rules have been proposed in the literature. We briefly review some of them below.

In [4], Authors propose an incremental association rules mining algorithm that integrates shocking interestingness criterion during the process of building the model. A new interesting measure called shocking measure is introduced.

One of the main features of the proposed approach is to capture the user background knowledge, which is monotonically augmented. The incremental model that reflects the changing data and the user beliefs is attractive in order to make the overall KDD process more effective and efficient.

In [5], presents an algorithm based on attribute information gain which can combine the subjective evaluation method and objective evaluation method together to discover interesting classification rules. The algorithm allows the users themselves to set the weight of each attribute's information gain, and the weights can reflect they preference, with different weights the algorithm can discover different interesting classification rules.

The disadvantage is that the accuracy of the interesting rules is not near to 1, and it costs too much time because the authors have to run the algorithm for each class especially for the large database.

In [6,7], The quantification of novelty is performed objectively and user involvement is sought for categorization of rules (as novel, unexpected, generalized, specialized, conformed) based on novelty measure, but the approaches used without mining and extracting new rules.

In [8], A new efficient type of genetic algorithm (GA) called uniform two-level GA is proposed as a search strategy to discover truly interesting, high-level prediction rules, Although the task of generalized rule induction requires a lot of computations, which is usually not satisfied with the normal algorithms, it was demonstrated that this method has coped with the problems of GAs such as divergence of genetic search process and remaining stuck on local solution of genetic search, and rapidly found interesting rules.

In [9], The authors proposed a framework to quantify the novelty in terms of computing the deviation of currently discovered knowledge with respect to domain knowledge and previously discovered knowledge. The approach presented is intuitive in nature and lays more emphasis on user involvement in quantification process by way of parameter specification. In the present work, the quantification is performed objectively and user involvement is sought in categorization of rules based on novelty index.

In [10], the novelty is estimated based on the lexical knowledge in WordNet. The proposed approach defines a measure of semantic distance between two words in WordNet and determined by the length of the shortest path between the two words (wi,wj). The novelty is defined as the average of this distance across all pairs of the words (wi,wj), where wi is a word in the antecedent and wj is a word in the consequent.

In [11,12,13,14], Generally, all studies build a model of training set that is selected to contain no examples of the important (i.e. novel) class. Subsequently, the mechanisms detect the deviation from this model by some way.

In [15], The use of GAs for rule discovery in the application of data mining has been studied some another studies, These algorithms are based on the Michigan approach in a way that each rule is encoded in a chromosome and the rule set is represented by the entire population.

In [16], a new algorithm has been proposed as a knowledge acquisition tool for classification problem and discovering significant IF-THEN rules. The proposed algorithm limited to generates set of classification rules and deletes the weak rules and selects only the significant rules.

In [17], a novel genetic based apriori algorithm has been proposed for web crawling. The proposed method yields promising results compared to the ordinary apriori algorithm and the authors present empirical results to substantiate this claim.

In [18], a new approach based GA has been proposed for discover classification rules. The proposed approach combined the GA with novelty measure to improve the discovered classification rules. The application of the proposed approach is limited on classification rules.

In [19], a new method based clustering has been proposed for mining fuzzy association rules. The proposed method used GA with other techniques for adjusts centroids of the clusters, which are to be handled later as midpoints of triangular membership functions. The proposed method not considered to improve any objective and subjective measures of patterns interestingness such as support, confidence, unexpectedness, novelty, and actionability.

In [20], a new genetic algorithm-based strategy has been developed for identifying association rules without specifying actual minimum support. The proposed algorithm does not require the minimum-support threshold but also didn’t consider any of objective or subjective measures of patterns interestingness.

In [21], a multi-objective based GA approach has been developed for optimizing fuzzy association rule mining in terms of three criteria which are strongness, interestingness and comprehensibility and then discovering these optimized rules. Novelty factor not considered by this approach.

In [22], a new algorithm has been proposed for mining quantitative association rules. The main idea in this algorithm is used GA for optimize both the support and the confidence in order to discover good intervals in association rules. The proposed algorithm didn’t contribute to improve of objective or subjective measures to discover interestingness association rules.

The proposed approach will be use genetic algorithm as soft computing tool techniques for mining and reduce the volume of discovered knowledge. The proposed approach will be evaluate the discovered association rules in both objective and subjective Manners during mining process and depends on confidence and novelty measures.

2.2 The Proposed Approach

In this paper, genetic algorithm used to incorporate novelty criterion with data mining techniques in order to find the useful knowledge, was novelty measure developed by [8].

The next subsections describe several aspects of the proposed approach, namely Encoding and parameters settings, Compute support and confidence values in order to find the optimal class for each rule, Fitness evaluation as Novelty measure, Crossover and mutation process and reproduction process.

2.2.1 Framework of the proposed approach

The general framework of the proposed approach consist main five processes which are: Encoding and parameter settings, frequent itemset mining process, population and association rule generation process, crossover and mutation process and novelty as a fitness measure process. Each of those processes will be discussed in details in a separate sub section below.

The general framework of the proposed approach is illustrated in Fig. 1.

Fig. 1 involved some terms which are, **ARs** which refers to association rules, $\mathbf{F_k}$ is kth frequency itemsets, $\mathbf{N_{Di}}$ refers to ith novelty degree, and $\mathbf{CHRi_{PDNAR}}$ which refers to ith chromosome of previous discovered novel association rules.

2.2.2 Encoding and parameter settings

Let **T** be a set of transactions in dataset. **n** be a set of transactions available in database or dataset, $\boldsymbol{T} = \{ChrI_1, ChrI_2, \ldots, ChrI_{\boldsymbol{n}}\}$. **ChrI** represent a set of items (itemset) within each $ChrI_n$, **ChrI** = $\{item_1, item_2, \ldots, item_{\boldsymbol{k}}\}$, **k** represent the number of items (attributes) shall be represented in each chromosome (ChrI). Then, **CHRi** is composed of **m** genes of possible classes.

In the proposed approach, each proceed chromosome will be transformed into useful knowledge rule. So, A chromosome should be corresponds to the entire IF part of the association rule, and each gene corresponds to one item or class in this IF part. Chromosome also involve support, confidence, and the last gene contain the novelty value of the current rule in order to improve the reliability of the proposed approach and enhance the usefulness measure of discover knowledge as shown in Fig. 2.

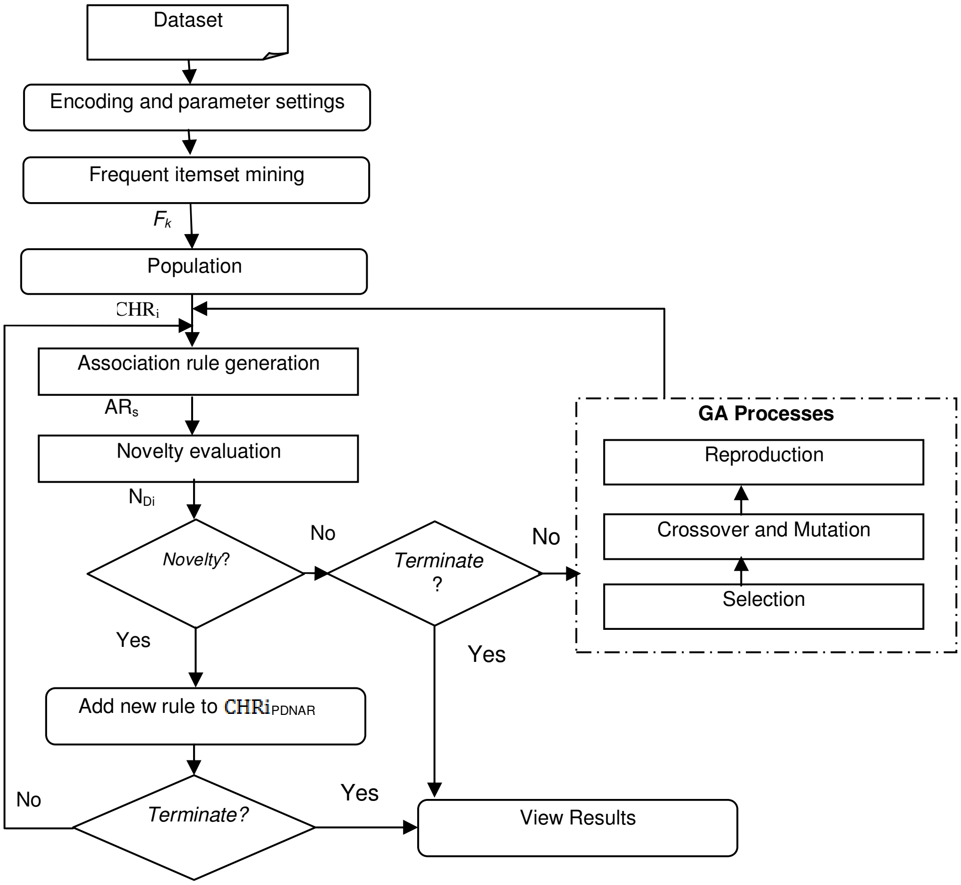


Fig. 1. General framework of the proposed approach.

item 1	...	...	item k	Class 1	...	Class m	Supp.	Conf.	N_D

Fig. 2. Chromosome encoding

Where,

- **k** = the level of frequent itemset - 1, which refers to number of possible items in the right side of rules.
- **m** = the level of frequent itemset – 1, which refers to number of possible classes of each rule.
- **T**: is a set of transactions in dataset.
- **CHRi**: refers to set of itemsets in T.
- **supp**.: refers to support value of **CHRi**.
- **conf**.: refers to confidence value of **CHRi** to select the optimal class of **CHRi**.
- **N_D =** refers to fitness value as novelty degree of **CHRi**.

So, before encoding process the proposed approach should be determine the some important parameters such as **n**, **T**, **k, m, CHRi, C_L , R_L, P_S , M_C, C_R, M_R, T_V , minsup, and minconf.**

Where:

- **n**: is a number of transactions available in given dataset.
- **C_L** = The chromosomes length = dynamic = [k + 4]
- **R_L** = The Rule length = dynamic [k] for both left (conditions attributes) and right sides (class attribute).
- **P_S** = The Population Size.
- **M_C** = The Maximum Cycles.
- **C_R**= The Crossover Rate.
- **M_R** = The Mutation Rate.
- **T_V** = The threshold value.
- **minsup**: value to find all rules that satisfy the user-specified minimum support.
- **minconf**: value to find all rules that satisfy the user-specified minimum confidence.

2.2.3 Frequent itemset mining process

A frequent itemset is an itemset whose support is greater than some user-specified minimum support. Frequent itemset generation is the essential process of the proposed approach. Apriori algorithm has been used in this approach to generate all frequent itemset. This algorithm requires the dataset transactions **T** and size of the itemset **k** as inputs. The Frequent itemset is output of this algorithm.

This process includes two sub processes which are generate candidate itemsets and find frequent itemsets.

In general, Apriori algorithm shall be use iteratively in this stage in order to generate initial candidate itemsets in C_1 and find all 1-item frequent itemsets and save them in L_1, then all 2-item frequent itemsets, and so on. In each iteration k, only consider itemsets that contain some k-1 frequent itemset.

Apriori algorithm is illustrate as the following [23,24].

Algorithm Apriori (*T, k*)

```
Begin
1.  q = 1,
2.  For each transaction $t \in T$ do
      - Find the frequent itemsets $C_q$ of size 1 and save them in $f_q$ using the Equation (1):
      -
          $$f_1 = \left\{ f \in C_q | \frac{c.count}{n} \geq minsup(C_q) \right\} \quad (1)$$
3.  For ($q = 2$; $F_{q-1} \neq \emptyset$; $q$++) do
          $C_q$= candidate-gen ($F_{q-1}$),
          for each transaction $t \in T$ do
          for each candidate $c \in C_q$ do
```

```
                    if $c$ is contained in $t$ then
                    $c.count$++,
                end
                end
        Find the frequent itemsets of $C_q$ and save them in $f_q$ using the Equation (2):
```

$$F_q = \left\{c \in C_q \mid \frac{c.count}{n} \geq minsup(C_q)\right\} \tag{2}$$

```
        end
    4.  return $F$.
End
```

Where,

- n: number of transactions in dataset T.
- C_q: refers to candidates of size q, those itemsets of size q that could be frequent.
- F_q: refers to itemsets that are actually frequent.

The candidate generation function required to run the apriori algorithm in order to generate potentially frequent itemset. The candidate generation function takes F_{q-1} as input and returns the candidates of the set of all frequent q-itemsets as output. This process can be done by two sub processes which are:

- **Join** process which aims to generate all possible candidate itemsets C_q of length q.
- **Prune** process which aims to remove those candidates in C_q that cannot be frequent.

The candidate generation function is executes as the following.

```
  Function candidate-gen($F_{q-1}$)
  Begin
   1.  $C_q = \varnothing$;
   2.  For all $f_1, f_2 \in F_{q-1}$
               with $f_1 = \{i_1, \dots, i_{q-2}, i_{q-1}\}$
               and $f_2 = \{i_1, \dots, i_{q-2}, i'_{q-1}\}$
               and $i_{q-1} < i'_{q-1}$ do
         $c = \{i_1, \dots, i_{q-1}, i'_{q-1}\}$,      // join $f_1$ and $f_2$
       $C_q = C_q \cup \{c\}$,
       for each ($q$-1) - subset $s$ of $c$ do
       if($s \notin F_{q-1}$) then
             delete $c$ from $C_q$,    // prune all candidate itemsets from $C_q$ where some (q-1)-
                   subset of the candidate itemset is not in the frequent itemset $F_{q-1}$
       end
       end
   3.  return $C_q$.
End
```

2.2.4 Population and association rule generation process

This process represent starting point to combined the genetic algorithm (GA) with data mining techniques in order to generate optimal association rules.

This process involves two main processes which are population generation and association rule generation. Each of those process will be explained in details in the following sub sections.

2.2.4.1 Population generation

The process of generating new populations based on the number of frequent itemsets $\boldsymbol{w}$ and exponential $\mathbf{O(2^q)}$, where $\mathbf{q}$ is the maximum number of items in $\boldsymbol{F}$. Find the probability number of populations and save them in $\boldsymbol{P_x}$ using the Equation (3):

$$P_x = (2^q - 2) * w \tag{3}$$

Programmatically we represent a chromosome as two-dimension array, that its rows represent the chromosomes in populations, and column represent genes that contains items and classes for each chromosome in populations.

This process involves two sub processes which are building chromosomes array and generate new populations. The following algorithms explained as the following.

```
Function build_chr_array(w, q)
Begin
        1. Read w and q.
        2. Find $P_x$.
        3. Build a chromosome array based on $P_x$.
        4. Call gen_population (q-1, 1)
End
```

Where

- P_x: is the number of probability populations.

```
Function gen_population (q, c)
Begin
  ▪ Read q and m,
  ▪ k = q – c,
  ▪ For each i^th frequent itemset in $F$,
     While $j \leq k$do,
        - $\mathbf{CHRi}[i,k] = F(k)$;
        - j++;
     end

     While $j \leq k + m$do,
        - $\mathbf{CHRi}[i,k+m] = F(k+m)$;
        - j++;
     end

     - Compute support value by Equation (4).
```

$$\mathbf{CHRi}[i,j+1] = \mathbf{supp}\left(\frac{item_{1..k} \cup class_{1..m}.count}{n}\right) \tag{4}$$

```
     - Compute confidence value by Equation (5).
```

$$\mathbf{CHRi}[i, j + 2] = \mathbf{conf}\left(\frac{item_{1..k} \cup class_{1..m}.count}{item_{1..k}.count}\right) \tag{5}$$

```
        ▪ End-for
End
```

Where

- q: is the maximum number of items in frequent itemsets.
- k: is the maximum number of possible items in the right side of rules.
- m: is the maximum number of possible classes of each rule.

2.2.4.2 Association Rules Generation

Context of each chromosome will be transform into useful knowledge as conditional rules. This process requires chromosomes context as input, the output is useful association rules which its support and confidence values are equal to or greater than minimum support and confidence. Generating association rules algorithm explained as the following.

Algorithm Ass_rule_gen (*CHRi*)

```
  Begin
    1. Call build_chr_array(w, q)
    2. For each pair of population do
         While m < k do
              For each m^th classes of association rule do
              o Call gen_population (k,m)
              o Compute support value
              o Compute confidence value
              o $item_{1..k} \rightarrow class_{1..m}$is an association rule if
                  - confidence(CHRi) ≥ minconf do
                  - add $\mathrm{CHRi}_{\mathbf{CDNRE}}$ to $\mathrm{CHRi}_{\mathbf{PDNAR}}$
              end-for
              m++
       end-while
       end-for
  End
```

Where

- $\mathrm{CHRi}_{\mathbf{CDNAR}}$: refers to current discovered useful knowledge.
- $\mathrm{CHRi}_{\mathbf{PDNAR}}$: refers to previous discovered useful knowledge.

2.2.5 Crossover and mutation process

The chromosomes which selected by selection process habituation to dealing with it as new population, then crossover occurs to exchange information between randomly selected parent chromosomes by recombining parts of their genetic materials.

We used one-point crossover, with crossover probability = 100%, and one point (gene) is selected at random locations from both items or classes sides of two chromosomes and interchanged.

Note: in the proposed approach should be set support and confidence values by zero for each child chromosome before execute the crossover operation between two chromosome.

When occur one point-crossover randomly between Pc1 and Pc2, then Assuming that, the two children are illustrated in Figs. 3 and 4 and formulated as:

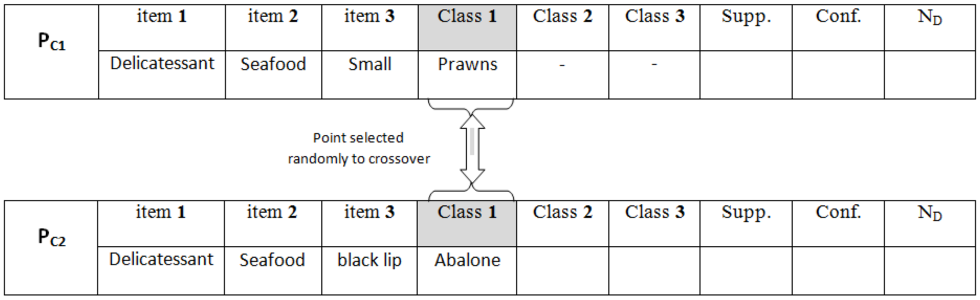
C$_{C1}$ = {Gene item 1, … , Gene item k, *Gene class 1*, … , Gene class m}

And

C$_{C2}$ = {Gene item 1, … , Gene item k, *Gene class 1*, … , Gene class m}

Where,

P$_{C1}$ = parent chromosome 1
P$_{C2}$ = parent chromosome 2
C$_{C1}$ = Child chromosome 1
C$_{C1}$ = Child chromosome 2

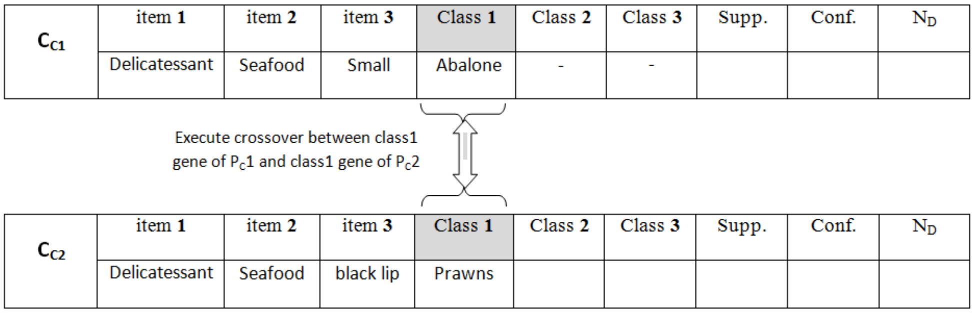


P_{C1}	item **1**	item **2**	item **3**	Class **1**	Class **2**	Class **3**	Supp.	Conf.	N_D
	Delicatessant	Seafood	Small	Prawns	-	-			

Point selected randomly to crossover

P_{C2}	item **1**	item **2**	item **3**	Class **1**	Class **2**	Class **3**	Supp.	Conf.	N_D
	Delicatessant	Seafood	black lip	Abalone					

Fig. 3. Chromosomes before one-point crossover process using real life data.



C_{C1}	item **1**	item **2**	item **3**	Class **1**	Class **2**	Class **3**	Supp.	Conf.	N_D
	Delicatessant	Seafood	Small	Abalone	-	-			

Execute crossover between class1 gene of P_c1 and class1 gene of P_c2

C_{C2}	item **1**	item **2**	item **3**	Class **1**	Class **2**	Class **3**	Supp.	Conf.	N_D
	Delicatessant	Seafood	black lip	Prawns					

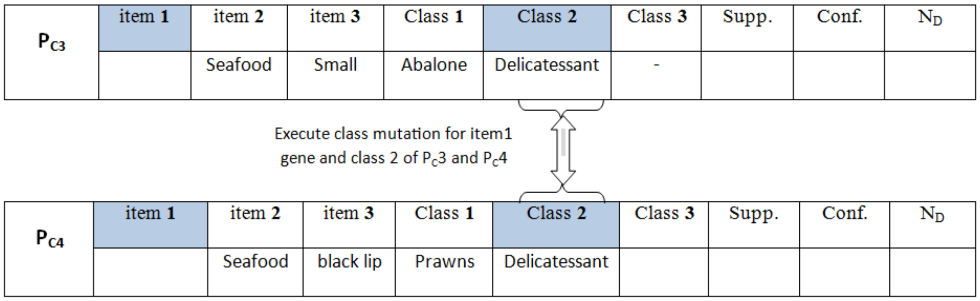
Fig. 4. Chromosomes after one-point crossover process using real life data.

Also, we used an elitist reproduction strategy, where the best individual of each generation was passed unaltered to the next generation. The main propose of Mutation process is generate association rules with different number of classes. This means that gene items moves between items and classes sides as illustrated in Fig. 5 below.

We developed two mutation operators tailored for our genome representation, namely item mutation and class mutation. Each of these operators acts once on a different field of item gene or class gene. We used mutation rates of 100% for both kind of mutations.

- **Item mutation** modifies the itemsets currently being used in a condition of the rule, by moving random class gene item of another chromosome into blank item of items side of current chromosome.
- **Class mutation** modifies the class(es) item, by moving random itemset gene of another chromosome into blank item of classes side of current chromosome.

Now, if we assume the child chromosomes that generated before in crossover process as shown in Fig. 5 as next generation and we can perform one point class mutation process on them randomly. The results are two chromosomes returned to dealing with it as new parents chromosome as shown in Fig. 5.



P_{C3}	item 1	item 2	item 3	Class 1	Class 2	Class 3	Supp.	Conf.	N_D
		Seafood	Small	Abalone	Delicatessant	-			

Execute class mutation for item1 gene and class 2 of P_c3 and P_c4

P_{C4}	item 1	item 2	item 3	Class 1	Class 2	Class 3	Supp.	Conf.	N_D
		Seafood	black lip	Prawns	Delicatessant				

Fig. 5. Chromosomes after class mutation process using real life data

2.2.6 Novelty as a fitness measure process

The proposed approach evaluates the degree of novelty for just association rules that have minimum confidence value or greater than it, this means the minimum confidence is a pre condition of novelty measure. The evaluation process performs during the mining process by fitness function. Novelty measure is based on threshold value that specified by user. The novelty evaluation should be applied separately on both items and classes sides of each chromosome, and the final decision of novelty depends on value of both sides. The main steps of this process presented as the following.

1. Find in CHRi about association rule that has highest confidence values.
2. Find in CHRi$_{\mathbf{PDNAR}}$ about association rule that has highest corresponding at items and classes genes to new rule in CHRi$_{\mathbf{CDNAR}}$.
3. Compute the grade of novelty for association new rule using the equation (6) [7, 9]:

$$\Psi(AR1\,,AR2)=\frac{\{|AR1|+|AR2|-2*k\}}{|AR1|+|AR2|} \tag{6}$$

Where,

- AR_1, AR_2 = are two gene sets(chromosome) with cardinalities $|AR1|$ and $|AR2|$ respectively.
- K = the pairs of compatible genes between $AR1$ and $AR2$.

IF ($\Psi(AR1, AR2)$>= THRESHOLD), THEN CHRi IS NOVEL, ADD CHRi TO CHRi$_{PDNAR}$.

2.2.7 The proposed algorithm

```
Begin
  Find the frequent items in dataset transactions
  Find the frequent itemsets F in T
  Compute support value for each frequent itemset
  Initialize the population P(n)
  While the termination condition is not met do
begin
    Generate association rule
    Compute support and confidence values for each generated association rule CHRi
            // Fitness function
     IF confidence(CHRi) >= minconf do
        For each CHRi ($item_i$→$Class_i$) in CHRi
        begin
        Find $CHRi_{CDNAR}$(itemj→ Class)j from $CHRi_{PDNAR}$, such that $\Psi$ ($item_i$ , $class_j$) is minimum.
          Compute $\Psi$ ($CHRi_{CDNAR}$(i) , $CHRi_{PDNAR}$(i))  //novelty measure
        IF ($\Psi$(CHRi**CDNAR**(i) , CHRi**PDNAR**(i))>= threshold), THEN
        $CHRi_{CDNAR}$(i) is novel, add $CHRi_{CDNAR}$(i) to   $CHRi_{PDNAR}$
   end
Select the best chromosomes from population p(n) based on fitness value.
Execute crossover and mutation processes for produce the offspring.
Replace, based on fitness, candidates of p(n), with these offspring.
  end
END
```

3 Results and Discussion

This section presents detailed of experimental setup, results and discussion.

3.1 Experimental Setup

The proposed approach is implemented and tested using public dataset. Since, there are no other approaches available, which pushed incorporating novelty during mining and generating association rules processes, we could not perform any comparison against our approach.

We considered these dataset as evolving with time, and partitioned them into 2 increments: D1, and D2 mined at times T_1, and T_2 respectively. We took each of these partitions to be equal for purpose of generating association rules.

The following experiment were conducted to show the effectiveness of the proposed approach using shopping basket dataset and based on the following parameters:

- **n** (number of transactions in given dataset) = 140.
- $\mathbf{F_L}$ (frequent itemset level) = 4.
- $\mathbf{F_n}$(number of frequent itemsets) = 10.
- **k** (number of possible items in each role) = 3.
- **m** (number of possible classes of each role) = 3.
- $\mathbf{C_L}$(chromosomes length, dynamic) = [k + m + 3].
- $\mathbf{R_L}$(rule length, dynamic) = $\mathbf{F_L}$.
- $\mathbf{M_C}$ (the maximum cycles, dynamic) = $(2^4 - 2) * 10$..
- $\mathbf{C_R}$ (the crossover rate) = 100%.
- $\mathbf{M_R}$ (the mutation rate) = 100%.
- $\mathbf{T_V}$ (the threshold value) >= 60%.
- **minsup** (minimum support value) = 25%.
- **minconf** (minimum confidence value) = 50%.

3.2 Computational Results and Discussion

In this sub section, we evaluate the performance of the proposed approach by conducting a series of experiments with all the well known items and frequent itemsets.

Actually, No comparisons have been done with other previous closely related works from literature because the experimental parameters are totally different from those the previous proposed algorithms in which not considered the novelty factor for improving the discovered association rules as mentioned above in literature review section.

For simplicity, we state running the proposed approach by 4th level of frequent itemset. Also we assume that the proposed approach will look for a solution in 10 frequent itemsets that have minimum support value and 140 subsets of a frequent itemset that generated as association rules. Fourteen association rules have been generated from each frequent itemset.

The GADNAR approach discovered 10 novel association rules only during 140 generation. Results shows that novel association rules have been generated from frequent itemsets number 1, 3, 6, 7, 8, 9 an 10. On the other hands, the proposed approach didn’t discovered any novel association rules from frequent itemsets number 2, 4 and 5 as shown in Table 1.

Note that the set of discovered novel association rules includes a default association rule which is rule number(1), a rule with no conditions which is automatically applied when no other association rule has its conditions satisfied by the example to be classified. The novelty values of discovered novel association rules is around between 60%-100% in both left (items) and right (classes) sides of rules.

Table 2 presents the final 10 association rules discovered by the proposed approach. These discovered rules filtered by some criteria measurements which are minimal confidence value and fitness criteria measures.

Table 1. Discovered novel association rules based on frequent itemset

# frequent itemset	Number of discovered novel association rules
F1	3
F2	0
F3	1
F4	0
F5	0
F6	2
F7	1
F8	1
F9	1
F10	1

Table 2. Discovered novelty association rules for shopping basket dataset

#	Association Rule	Confidence	Novelty degree
1	Delicatessant, seafood, prawns → small	33%	[1, 1]
2	Seafood, small, prawns →Delicatessant	100%	[0.67, 1]
3	Delicatessant, small → prawns, seafood	100%	[0.60, 1]
4	Chillimedum→Delicatessant, seafood, prawns	100%	[1, 0.60]
5	Delicatessant, chicken, roast → quarter	50%	[0.67, 1]
6	Delicatessant, quarter → roast, chicken	100%	[0.60, 1]
7	Roast →Delicatessant, chicken, haif	50%	[1, 0.67]
8	Fillets, chicken, breast →Delicatessant	100%	[0.67, 1]
9	Delicatessant, BBQ, wings → chicken	100%	[0.67, 1]
10	Sweet, chicken, wings →Delicatessant	100%	[0.67, 1]

As observed from the results in Table 2, its seems that the discovered novel association rules differs of others in term of number of items, classes and degree of novelty.

For each rule in Table 2 the fourth column shows two values, the first one indicate to the degree of novelty for the itemsets of the association rule, and the second one indicate to the degree of novelty for the class(s) value of the association rule, namely the degree of novelty of the association rule computed by equation 6.

As shown by Table 2 and Fig. 6 below, the novelty average of classes side with value 92.7% is better than novelty average of items side with value 75.47%. This means that the class side of association rule represent the knowledge and result of mining to support decision makers.

In general, it is interesting to evaluate the novelty of the set of discovered association rules as a whole, by applying the confidence and novelty measures. The percentage of discovered repeated, conformed, generalized, and specialized knowledge and hence not interesting association rules was 86% (out of 140 discovered rules, 130 rules were not satisfied to confidence and novelty measures).

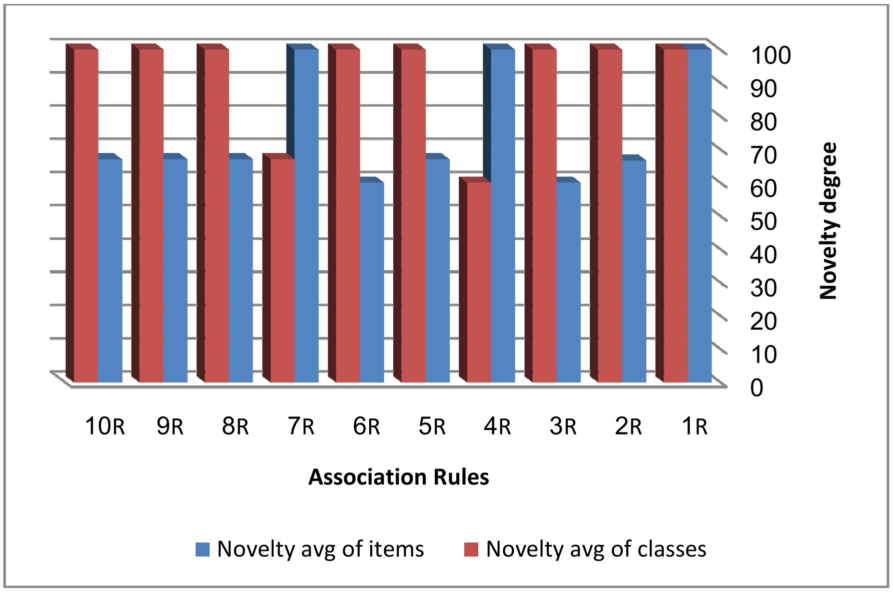


Fig. 6. Discovered novelty association rules for shopping basket dataset

4 Conclusion

In this paper, new strategy has been proposed for association rule set reduction based on confidence and novelty measures. Pushed the confidence and novelty criterion into a genetic algorithm in order to generate association rules from huge datasets and keep just novel and hence interesting knowledge.

Novelty and confidence indexes of a newly discovered association rules are the quantification of its deviation with respect to the known rule set. User subjectivity should be captured by specification of minimum confidence and threshold of fitness function for association rule categorization.

The proposed approach has been experimented and evaluated using real-life datasets and results have been presented. The generated association rules were categorized as conforming, generalized/specialized, unexpected and novel rules. Then, these generated association rules have been filtered to get novel and hence interesting knowledge which finds by 14% approximately.

5 Future Work

This work can further be extended to cover and evaluate by some factors such as time and complexity factors. In the other hands, it should be include the high order level of frequent itemsets and generalize to be applicable on wide range of data types. Also, it should consist of more experiments with other datasets, as well as more elaborated experiments to optimize several parameters of the algorithm, and the proposed approach will be implemented by computer program. In addition, this work can further be extended to cover more technological applications of GA in order to improve discovered association rules.

Acknowledgements

Thanks to Dr. Fahd Al-Wesabi for collecting the data.

Competing Interests

Author has declared that no competing interests exist.

References

[1] Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha P. Sarkar. Mining frequent itemsets using genetic algorithm. Nternational Journal of Artificial Intelligence & Applications (IJAIA). 2010;14.

[2] Sushmita Mitra, Senior Member, IEEE, Sankar K. Pal, Fellow, IEEE, and Pabitra Mitra. Data mining in soft computing framework: A Survey, IEEE Transactions on Neural Networks. 2002;13:1.

[3] Arvind J, Gaurav D. Identifying best association rules and their optimization using genetic algorithm. International Journal of Emerging Science and Engineering (IJESE). 2013;1:7.

[4] Yafi E, Al-Hegami AS, Alam MA, Biswas R. Incremental mining of shocking association patterns. In proceedings of world academy of science. Engineering and Technology. 2009;37. Dubai, UAE.

[5] Zhou Y, Xia S, Gong D, Youwen L. Knowledge discovery of interesting classification rules based on adaptive genetic algorithm. ISKE; 2007.

[6] Bhatnagar V, Al-Hegami AS, Kumar N. A hybrid approach for quantification of novelty in rule discovery. In Proceedings of International Conference on Artificial Learning and Data Mining (ALDM'05), Turkey. 2005;39-42.

[7] Bhatnagar V, Al-Hegami AS, Kumar N. Novelty as a measure of interestingness in knowledge discovery. In International Journal of Information Technology. 2005;2:1.

[8] Alatas B, Arslan A. Mining of interesting prediction rules with uniform two-level genetic algorithm. International Journal of Computational Intelligence, Fall; 2005.

[9] Al-Hegami S, Bhatnagar V, Kumar N. Novelty framework for knowledge discovery in databases. In Proceedings of the 6th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2004), Zaragoza, Spain. 2004;48-55.

[10] Basu S, Mooney RJ, Pasupuleti KV, Ghosh J. Using lexical knowledge to evaluate the novelty of rules mined from text. In Proceedings of the NAACL workshop and other Lexical Resources: Applications, Extensions and Customizations; 2001.

[11] Marsland S. On-line novelty detection through self-organization, with application to robotics. Ph.D. Thesis, Department of Computer Science, University of Manchester; 2001.

[12] Japkowicz N, Myers C, Gluck M. A novelty detection approach to classification. In Proceedings of the 14th International Joint Conference on Artificial Intelligence; 1995.

[13] Roberts S, Tarassenko L. A probabilistic resource allocation network for novelty detection. In Neural Computation. 1994;6(2).

[14] Ypma R. Duin. Novelty detection using self-organizing maps. In Progress in Connectionist-Based Information Systems. 1997;2.

[15] Choenni S. Design and implementation of a genetic-based algorithm for data mining. In Proc. of the 26th Int'l Conf. on Very Large Data Bases, Cairo, Egypt. 2000;33-42.

[16] Fadl Mutaher Ba-Alwi. Knowledge acquisition tool for classification rules using genetic algorithm approach. International Journal of Computer Applications. 2012;60:1.

[17] Usharani J, Dr. K. Iyakutti. Mining association rules for web crawling using genetic algorithm, International Journal of Engineering and Computer Science. 2013;2:8.

[18] Fahd Alwesabi, Ahmed Alhejami. Intelligent discovery of novel classification rules based on genetic algorithms. Journal of Intelligent Computing, London, UK. 2011;2:1.

[19] Kaya M, Alhajj R. Genetic algorithm based framework for mining fuzzy association rules [J]. Fuzzy Sets and Systems. 2005;152(3):587-601.

[20] Yan X, Zhang C, Zhang S. Genetic algorithm based strategy for identifying association rules without specifying actual minimum support [J]. Expert Systems with Applications. 2009;36(2):3066-3076.

[21] Kaya M. Multiobjective genetic algorithm based approaches for mining optimized fuzzy as sociation rules[J]. Soft Computing, 2006:10(7);578-586.

[22] SallebAouissi A, Vrain C, Nortet C. QuantMiner: A Genetic Algorithm for Mining Quantit ative Association Rules[C], IJCAI; 2007.

[23] Han J, Kamber M. Data mining: Concepts and techniques. Morgan Kaufmann, Academic, San Fransisco, NewYork; 2001.

[24] Charanjeer kaur. Association rule mining using apriori algorithm: A Survey Internationl Journal of advanced Research in Computer Engineering & Technology (IJARCET). 2013;2:6.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
www.sciencedomain.org/review-history.php?iid=669&id=6&aid=6201