

DISCOVERY OF NOVEL ASSOCIATION RULES BASED ON GENETIC ALGORITHMS

Fadl Mutaher Ba-Alwi^{1*}

¹ Associate professor, Faculty of Computer and Information Technology, Sana'a University,
Yemen, P.O.Box 1247

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 from 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 weak 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.

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 [17]. The term KDD refers to the overall process of knowledge discovery in databases. Data mining is a particular step

* Tel.: +967-777933489; fax: +967-1-464723.

E-mail address: dr.fadlbaalwi@gmail.com.

in this process, involving the application of specific algorithms for extracting patterns (models) from data [16].

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 [18].

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 [16]. 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 [18].

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. MATERIAL AND METHODS / EXPERIMENTAL DETAILS / METHODOLOGY

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 [1], 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

84 changing data and the user beliefs is attractive in order to make the over all KDD process
85 more effective and efficient.

86 In [14], Presents an algorithm based on attribute information gain which can combine the
87 subjective evaluation method and objective evaluation method together to discover
88 interesting classification rules. The algorithm allows the users themselves to set the weight
89 of each attribute's information gain, and the weights can reflect they preference, with
90 different weights the algorithm can discover different interesting classification rules.

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

94 In [3, 4], The quantification of novelty is performed objectively and user involvement is
95 sought for categorization of rules (as novel, unexpected, generalized, specialized,
96 conformed) based on novelty measure, but the approaches used without mining and
97 extracting new rules.

98 In [5], A new efficient type of genetic algorithm (GA) called uniform two-level GA is proposed
99 as a search strategy to discover truly interesting, high-level prediction rules, Although the
100 task of generalized rule induction requires a lot of computations, which is usually not
101 satisfied with the normal algorithms, it was demonstrated that this method has coped with
102 the problems of GAs such as divergence of genetic search process and remaining stuck on
103 local solution of genetic search, and rapidly found interesting rules.

104 In [6], The authors proposed a framework to quantify the novelty in terms of computing the
105 deviation of currently discovered knowledge with respect to domain knowledge and
106 previously discovered knowledge. The approach presented is intuitive in nature and lays
107 more emphasis on user involvement in quantification process by way of parameter
108 specification. In the present work, the quantification is performed objectively and user
109 involvement is sought in categorization of rules based on novelty index.

110 In [7], The novelty is estimated based on the lexical knowledge in WordNet. The proposed
111 approach defines a measure of semantic distance between two words in WordNet and
112 determined by the length of the shortest path between the two words (w_i, w_j). The novelty is
113 defined as the average of this distance across all pairs of the words (w_i, w_j), where w_i is a
114 word in the antecedent and w_j is a word in the consequent.

115 In [8, 9, 10, 11], Generally, all studies build a model of training set that is selected to contain
116 no examples of the important (i.e. novel) class. Subsequently, the mechanisms detect the
117 deviation from this model by some way.

118 In [12], The use of GAs for rule discovery in the application of data mining has been studied
119 some another studies, These algorithms are based on the Michigan approach in a way that
120 each rule is encoded in a chromosome and the rule set is represented by the entire
121 population.

122 In [19], a new algorithm has been proposed as a knowledge acquisition tool for classification
123 problem and discovering significant IF-THEN rules. The proposed algorithm limited to
124 generates set of classification rules and deletes the weak rules and selects only the
125 significant rules.

126 In [20], a novel genetic based apriori algorithm has been proposed for web crawling. The
127 proposed method yields promising results compared to the ordinary apriori algorithm and the
128 authors present empirical results to substantiate this claim.

129 In [21], a new approach based GA has been proposed for discover classification rules. The
130 proposed approach combined the GA with novelty measure to improve the discovered
131 classification rules. the application of the proposed approach is limited on classification rules.

132 In [22], a new method based clustering has been proposed for mining fuzzy association
133 rules. the proposed method used GA with other techniques for adjusts centroids of the
134 clusters, which are to be handled later as midpoints of triangular membership functions. The
135 proposed method not considered to improve any objective and subjective measures of

136 patterns interestingness such as support, confidence, unexpectedness, novelty, and
137 actionability.

138 In [23], a new genetic algorithm-based strategy has been developed for identifying
139 association rules without specifying actual minimum support. The proposed algorithm does
140 not require the minimum-support threshold but also didn't consider any of objective or
141 subjective measures of patterns interestingness.

142 In [24], a multi-objective based GA approach has been developed for optimizing fuzzy
143 association rule mining in terms of three criteria which are strongness, interestingness and
144 comprehensibility and then discovering these optimized rules. Novelty factor not considered
145 by this approach.

146 In [25], a new algorithm has been proposed for mining quantitative association rules. the
147 main idea in this algorithm is used GA for optimize both the support and the confidence in
148 order to discover good intervals in association rules. the proposed algorithm didn't contribute
149 to improve of objective or subjective measures to discover interestingness association rules.

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

154

155 **2.2 The Proposed Approach**

156 In this paper, genetic algorithm used to incorporate novelty criterion with data mining
157 techniques in order to find the useful knowledge, was novelty measure developed by [5].

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

162

163 **2.2.1 Framework of the Proposed Approach**

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

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

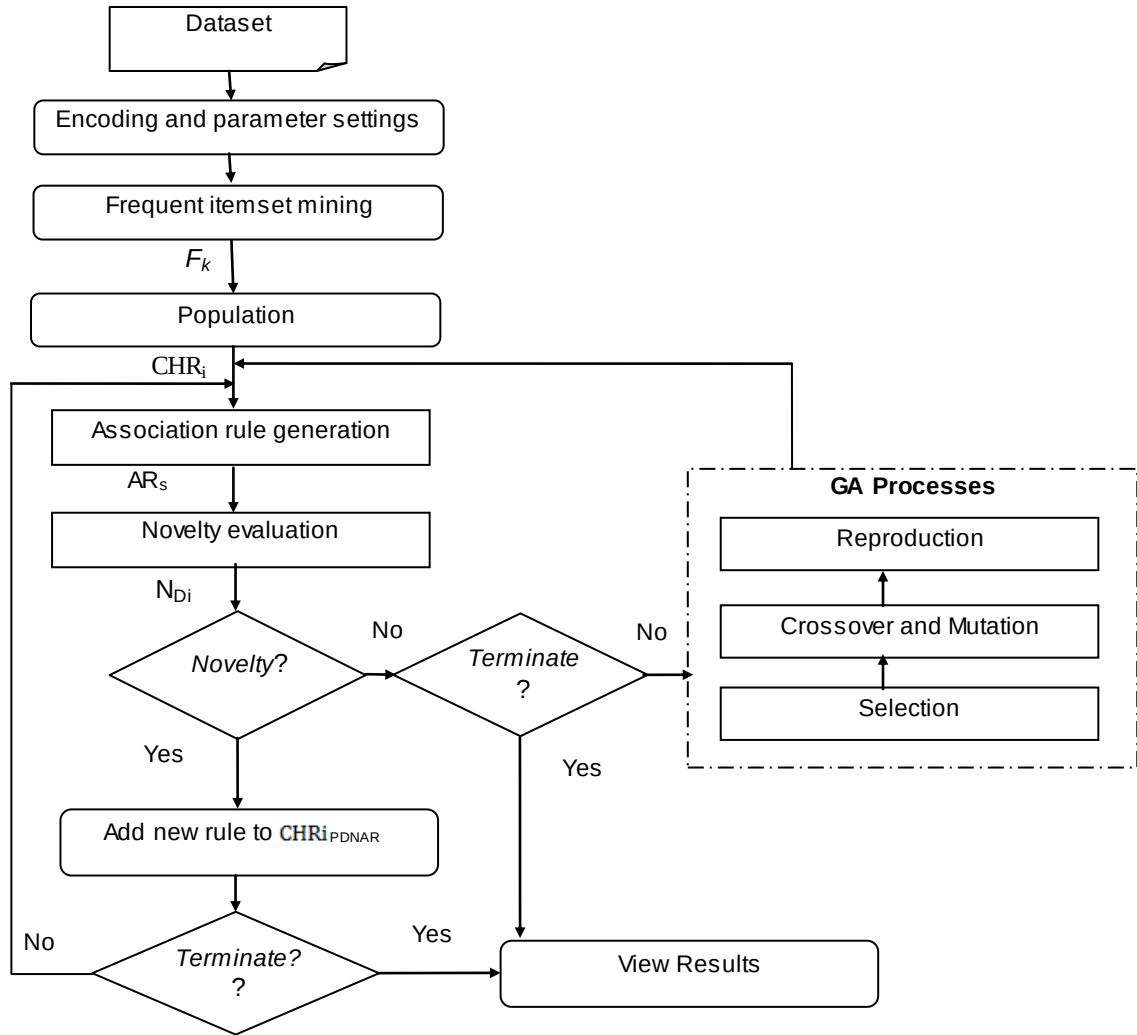


Fig. 1. General framework of the proposed approach.

Figure 1 above involved some terms which are, **ARs** which refers to association rules, **F_k** is **kth** frequency itemsets, **N_{Di}** refers to **ith** novelty degree, and **CHR_{iPDNAR}** 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, **T** = { **Chr₁**, **Chr₂**, ..., **Chr_n**}. **Chr_l** represent a set of items (itemset) within each **Chr_n**, **Chr_l** = {**item₁**, **item₂**, ..., **item_k**}, **k** represent the number of items (attributes) shall be represented in each chromosome (Chr_l). Then, **CHR_i** 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 Figure 2.

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.
- **CHR_i**: refers to set of itemsets in T.
- **supp.**: refers to support value of **CHR_i**.
- **conf.**: refers to confidence value of **CHR_i** to select the optimal class of **CHR_i**.
- **N_D** = refers to fitness value as novelty degree of **CHR_i**.

So, before encoding process the proposed approach should be determine the some important parameters such as **n**, **T**, **k**, **m**, **CHR_i**, **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 **C₁** and find all 1-item frequent itemsets and save them in **L₁**, then all 2-item frequent itemsets, and so on. In each iteration **k**, only consider itemsets that contain some **k-1** frequent itemset.

240 Apriori algorithm is illustrate as the following.

241

242 **Algorithm Apriori(T, k)**

243 Begin

244 1. $q = 1$,

245 2. For each transaction $t \in T$ **do**

246 - Find the frequent itemsets C_q of size 1 and save them in f_q using the Equation
247 (1):

$$f_1 = \left\{ f \in C_q \mid \frac{f.count}{n} \geq \text{minsup}(C_q) \right\} \quad \dots\dots\dots (1)$$

249 3. **For** ($q = 2; F_{q-1} \neq \emptyset; q++$) **do**

250 $C_q = \text{candidate-gen}(F_{q-1})$,

251 **for** each transaction $t \in T$ **do**

252 **for** each candidate $c \in C_q$ **do**

253 **if** c is contained in t **then**

254 $c.count++$,

255 **end**

256 **end**

257 Find the frequent itemsets of C_q and save them in f_q using the Equation (2):

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

259 **end**

260 4. return F .

261 End

262

263

264 Where,

265 - n : number of transactions in dataset T .

266 - C_q : refers to candidates of size q , those itemsets of size q that could be frequent.

267 - F_q : refers to itemsets that are actually frequent.

268

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

273 - **Join** process which aims to generate all possible candidate itemsets C_q of length q .

274 - **Prune** process which aims to remove those candidates in C_q that cannot be
275 frequent.

276

277 The candidate generation function is executes as the following.

278

279 **Function** candidate-gen(F_{q-1})

```

280      Begin
281      1.  $C_q = \emptyset$ ;
282      2. For all  $f_1, f_2 \in F_{q-1}$ 
283          with  $f_1 = \{i_1, \dots, i_{q-2}, i_{q-1}\}$ 
284          and  $f_2 = \{i_1, \dots, i_{q-2}, i'_{q-1}\}$ 
285          and  $i_{q-1} < i'_{q-1}$  do
286       $c = \{i_1, \dots, i_{q-1}, i'_{q-1}\}$ , // join  $f_1$  and  $f_2$ 
287       $C_q = C_q \cup \{c\}$ ,
288      for each (q-1) - subset  $s$  of  $c$  do
289          if ( $s \notin F_{q-1}$ ) then
290              delete  $c$  from  $C_q$ , // prune all candidate itemsets from  $C_q$  where some (q-1)-
291              subset of the candidate itemset is not in the frequent itemset  $F_{q-1}$ 
292          end
293      end
294      3. return  $C_q$ .

```

295 End

296

297 2.2.4 Population and Association Rule Generation Process

298

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

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

304

305 *2.2.4.1 Population Generation*

306

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

310

$$311 \quad P_x = (2^q - 2) * w \quad \dots\dots\dots (3)$$

312

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

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

318

319 Function build_chr_array(w, q)

320 Begin

321 1. Read w and q .

322 2. Find P_x .

323 3. Build a chromosome array based on P_x .

324 4. Call gen_population ($q-1, 1$)

325 End

326
327 **Where**
328 - P_x : is the number of probability populations.
329
330
331 Function gen_population (q, c)
332 Begin
333 ▪ Read q and m,
334 ▪ $k = q - c$,
335 ▪ For each i^{th} frequent itemset in F ,
336 While $j \leq k$ do,
337 - $CHRi[i, k] = F(k)$;
338 - $j++$;
339 end
340
341 While $j \leq k + m$ do,
342 - $CHRi[i, k + m] = F(k + m)$;
343 - $j++$;
344 end
345
346 - Compute support value by Equation (4).
347
348
$$CHRi[i, j + 1] = \text{supp} \left(\frac{\text{item}_{1..k} \cup \text{class}_{1..m} . \text{count}}{n} \right) \quad \dots\dots\dots (4)$$

349
350 - Compute confidence value by Equation (5).
351
352
$$CHRi[i, j + 2] = \text{conf} \left(\frac{\text{item}_{1..k} \cup \text{class}_{1..m} . \text{count}}{\text{item}_{1..k} . \text{count}} \right) \quad \dots\dots\dots (5)$$

353 ▪ End-for
354 End
355
356 **Where**
357 - q: is the maximum number of items in frequent itemsets.
358 - k: is the maximum number of possible items in the right side of rules.
359 - m: is the maximum number of possible classes of each rule.
360
361
362 **2.2.4.2 Association Rules Generation**
363
364 Context of each chromosome will be transform into useful knowledge as conditional rules.
365 This process requires chromosomes context as input, the output is useful association rules
366 which its support and confidence values are equal to or greater than minimum support and
367 confidence. Generating association rules algorithm explained as the following.
368
369 **Algorithm Ass_rule_gen (CHRi)**
370 Begin
371 1. Call build_chr_array(w, q)
372 2. For each pair of population do

```

373         While m < k do
374             For each mth classes of association rule do
375                 o Call gen_population (k,m)
376                 o Compute support value
377                 o Compute confidence value
378                 o  $item_{1,k} \rightarrow class_{1,m}$  is an association rule if
379                     - confidence(CHRi) ≥ minconf do
380                     - add CHRiCDNRE to CHRiPDNAR
381             end-for
382             m++
383         end-while
384     end-for
385 End
386
387 Where
388     - CHRiCDNAR: refers to current discovered useful knowledge.
389     - CHRiPDNAR: refers to previous discovered useful knowledge.

```

392 **2.2.5 Crossover and Mutation Process**

```

393
394     The chromosomes which selected by selection process habituation to dealing with it as
395     new population, then crossover occurs to exchange information between randomly
396     selected parent chromosomes by recombining parts of their genetic materials.
397     We used one-point crossover, with crossover probability = 100%, and one point (gene) is
398     selected at random locations from both items or classes sides of two chromosomes and
399     interchanged.
400     Note: in the proposed approach should be set support and confidence values by zero for
401     each child chromosome before execute the crossover operation between two
402     chromosome.
403     When occur one point-crossover randomly between Pc1 and Pc2, then Assuming that, the
404     two children are illustrated in figures 3 and 4 and formulated as:
405         CC1 = { Gene item 1 , ... , Gene item k, Gene class 1, ... , Gene class m}
406     And
407         CC2 = { Gene item 1 , ... , Gene item k, Gene class 1, ... , Gene class m}
408
409     Where,
410         PC1 = parent chromosome 1
411         PC2 = parent chromosome 2
412         CC1 = Child chromosome 1
413         CC1 = Child chromosome 2
414

```

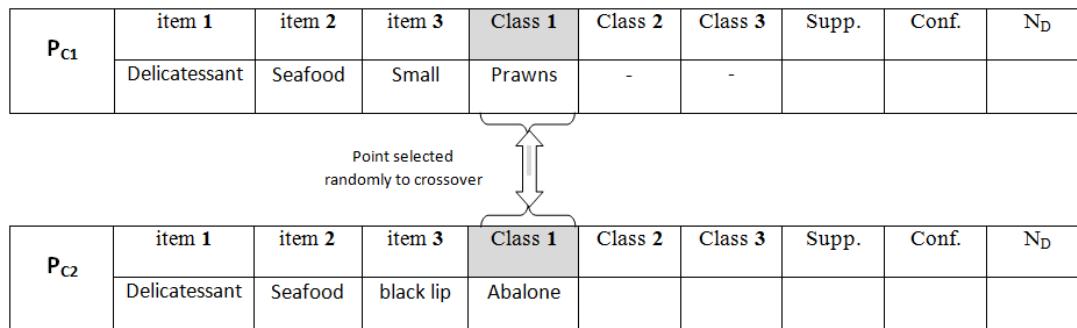


Fig. 3. Chromosomes before one-point crossover process using real life data.

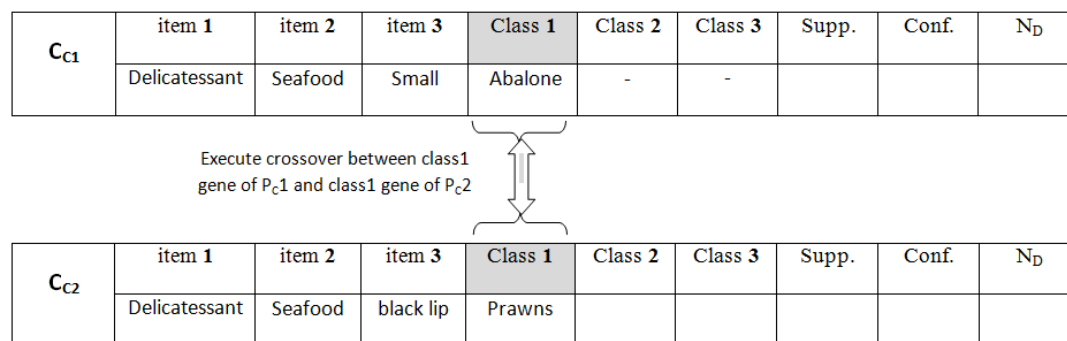


Fig. 4. Chromosomes after one-point crossover process using real life data.

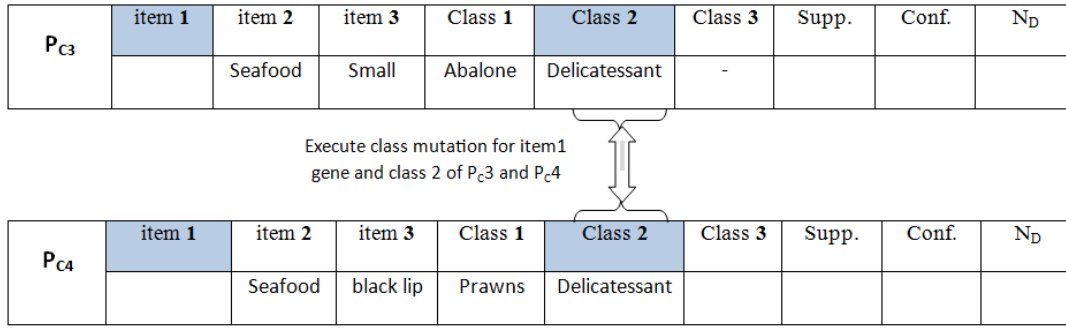
Also, we used an elitist reproduction strategy, where the best individual of each generation was passed unaltered to the next generation. The main propose of Mutation process is generate association rules with different number of classes. This means that gene items moves between items and classes sides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

440



441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

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 CHR_i about association rule that has highest confidence values.
2. Find in CHR_{PDNAR} about association rule that has highest corresponding at items and classes genes to new rule in CHR_{CDNAR} .
3. Compute the grade of novelty for association new rule using the equation (6) [4,6]:

$$\Psi(AR1, AR2) = \frac{(|AR1| + |AR2| - 2 \cdot k)}{|AR1| + |AR2|} \quad \dots(6)$$

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 ($\Psi(AR1, AR2) \geq \text{THRESHOLD}$), THEN CHR_i IS NOVEL, ADD CHR_i TO CHR_{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

* Tel.: +967-777933489; fax: +967-1-464723.

E-mail address: dr.fadlbaalwi@gmail.com.

```

480   Compute support and confidence values for each generated association rule CHRI
481   // Fitness function
482   IF confidence(CHRI) >= minconf do
483       For each CHRI (itemi → Classi) in CHRI
484       begin
485           Find CHRICDNAR(itemj → Classj) from CHRIPDNAR, such that  $\Psi$  (itemi , classj) is
486           minimum.
487           Compute  $\Psi$  (CHRICDNAR(i) , CHRIPDNAR(i)) //novelty measure
488           IF ( $\Psi$ (CHRICDNAR(i) , CHRIPDNAR(i)) >= threshold), THEN
489               CHRICDNAR(i) is novel, add CHRICDNAR(i) to CHRIPDNAR
490           end
491       Select the best chromosomes from population p(n) based on fitness value.
492       Execute crossover and mutation processes for produce the offspring.
493       Replace, based on fitness, candidates of p(n), with these offspring.
494   end
495 END

```

498 3. RESULTS AND DISCUSSION

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

502 3.1 Experimental Setup

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

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

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

- 513 ▪ **n** (number of transactions in given dataset) = 140.
- 514 ▪ **F_L** (frequent itemset level) = 4.
- 515 ▪ **F_n** (number of frequent itemsets) = 10.
- 516 ▪ **k** (number of possible items in each role) = 3.
- 517 ▪ **m** (number of possible classes of each role) = 3.
- 518 ▪ **C_L**(chromosomes length, dynamic) = $[k + m + 3]$.
- 519 ▪ **R_L** (rule length, dynamic) = **F_L**.
- 520 ▪ **M_C** (the maximum cycles, dynamic) = $(2^4 - 2) * 10$.
- 521 ▪ **C_R** (the crossover rate) = 100%.
- 522 ▪ **M_R** (the mutation rate) = 100%.
- 523 ▪ **T_v** (the threshold value) >= 60%.
- 524 ▪ **minsup** (minimum support value) = 25%.
- 525 ▪ **minconf** (minimum confidence value) = 50%.

528 3.2 Computational Results and Discussion

* Tel.: +967-777933489; fax: +967-1-464723.

E-mail address: dr.fadlbaalwi@gmail.com.

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.

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

547

548

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.

557

Table 2. Discovered novelty association rules for shopping basket dataset

558

559

#	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	Chilli medum → 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]

#	Association Rule	Confidence	Novelty degree
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, it 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 Figure 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).

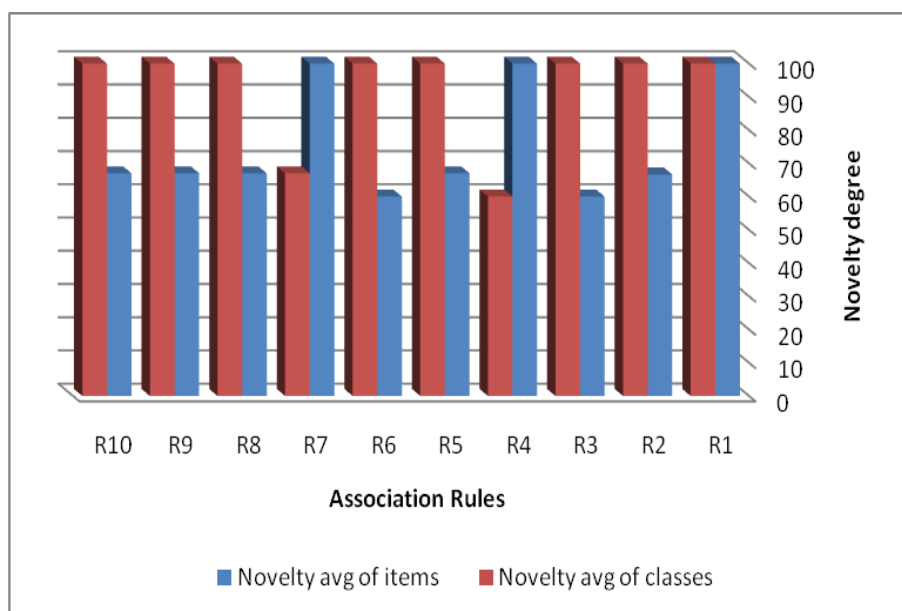


Fig. 6. Discovered novelty association rules for shopping basket dataset

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

A brief acknowledgement section may be given after the conclusion section just before the references. The acknowledgments of people who provided assistance in manuscript preparation, funding for research, etc. should be listed in this section. All sources of funding should be declared as an acknowledgement. Authors should declare the role of funding agency, if any, in the study design, collection, analysis and interpretation of data; in the writing of the manuscript. If the study sponsors had no such involvement, the authors should so state.

COMPETING INTERESTS

The proposed approach presented in this paper realizes an interesting application of the genetic algorithms to discover a novel association rules in a large set of data.

AUTHORS' CONTRIBUTIONS

'Author' developed the approach by using genetic algorithm to incorporate novelty criterion with data mining techniques in order to find the useful knowledge.

REFERENCES

1. E. Yafi, A. S. Al-Hegami, M. A. Alam, R. Biswas. Incremental Mining of Shocking Association Patterns. In Proceedings of World Academy of Science. Engineering and Technology. Volume 37. January 2009. Dubai, UAE.

- 641 2. Min Wang, Qin Zou, Caihui Liu. Multi-dimension Association Rule Mining Based on
642 Adaptive Genetic Algorithm. in IEEE International Conference on Uncertainty Reasoning
643 and Knowledge Engineering, 2011, pp.150-153
- 644 3. V. Bhatnagar, A. S. Al-Hegami, N. Kumar. A hybrid approach for Quantification of
645 Novelty in Rule Discovery. In Proceedings of International Conference on Artificial
646 Learning and Data Mining (ALDM'05), Turkey, Feb. 25-27, 2005, pp 39-42.
- 647 4. V. Bhatnagar, A. S. Al-Hegami, N. Kumar. Novelty as a Measure of Interestingness in
648 Knowledge Discovery. In International Journal of Information Technology, Volume 2,
649 Number 1, 2005.
- 650 5. B. Alatas, and A. Arslan. Mining of Interesting Prediction Rules with Uniform Two-Level
651 Genetic Algorithm. International Journal of Computational Intelligence, Fall 2005.
- 652 6. S. Al-Hegami, V. Bhatnagar, N. Kumar, Novelty Framework for Knowledge Discovery in
653 Databases. In Proceedings of the 6th International Conference on Data Warehousing
654 and Knowledge Discovery (DaWaK 2004), Zaragoza, Spain, 2004, pp 48-55
- 655 7. S. Basu, R. J. Mooney, K. V. Pasupuleti, J. Ghosh, Using Lexical Knowledge to Evaluate
656 the Novelty of Rules Mined from Text. In Proceedings of the NAACL workshop and other
657 Lexical Resources: Applications, Extensions and Customizations, 2001.
- 658 8. S. Marsland. On-Line Novelty Detection Through Self-Organization, with Application to
659 Robotics. Ph.D. Thesis, Department of Computer Science, University of Manchester,
660 2001.
- 661 9. N. Japkowicz, C. Myers, M. Gluck. A Novelty Detection Approach to Classification. In
662 Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995.
- 663 10. S. Roberts, L. Tarassenko, A Probabilistic Resource Allocation Network for Novelty
664 Detection. In Neural Computation, 6(2), 1994
- 665 11. Ypma, R. Duin. Novelty Detection Using Self-Organizing Maps. In Progress in
666 Connectionist-Based Information Systems. Volume 2, 1997.
- 667 12. S. Choenni. Design and Implementation of a Genetic-Based Algorithm for Data Mining.
668 in Proc. of the 26th Int'l Conf. on Very Large Data Bases, Cairo, Egypt, 2000, pp. 33-42.
- 669 13. Han J, Kamber M (2001) Data mining: concepts and techniques, Morgan Kaufmann,
670 Academic, San Fransisco, NewYork
- 671 14. Y. Zhou, S. Xia, D. Gong, L.Youwen. Knowledge Discovery of interesting Classification
672 Rules Based on Adaptive Genetic Algorithm. ISKE, 2007.
- 673 15. Charanjeer kaur, Association Rule Mining Using Apriori Algorithm: A Survey International
674 Journal of advanced Research in Computer Engineering & Technology(IJARCET)
675 Volume2,Issue 6,June 2013
- 676 16. Sushmita Mitra, Senior Member, IEEE, Sankar K. Pal, Fellow, IEEE, and Pabitra Mitra,
677 Data Mining in Soft Computing Framework:A Survey, IEEE transactions on neural
678 networks, VOL. 13, NO. 1, JANUARY 2002.
- 679 17. Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha P. Sarkar, Mining
680 Frequent Itemsets Using Genetic Algorithm, nternational Journal of Artificial Intelligence
681 & Applications (IJAIA), Vol.1, No.4, 2010.
- 682 18. Arvind J., Gaurav D., Identifying Best Association Rules and Their Optimization Using
683 Genetic Algorithm, International Journal of Emerging Science and Engineering (IJESE),
684 Volume-1, Issue-7, 2013.
- 685 19. Fadl Mutaher Ba-Alwi, Knowledge Acquisition Tool for Classification Rules using
686 Genetic Algorithm Approach, International Journal of Computer Applications, Volume
687 60– No.1, 2012.
- 688 20. J. Usharani, Dr. K. Iyakutti, Mining Association Rules for Web Crawling using Genetic
689 Algorithm, International Journal Of Engineering And Computer Science, Volume 2 Issue
690 8, 2013.
- 691 21. Fahd Alwesabi and Ahmed Alhejami, Intelligent Discovery of Novel Classification Rules
692 Based on Genetic Algorithms. Journal of Intelligent Computing, London, UK, Volume 2,
693 Issue 1, March 2011.

- 694 22. Kaya M, Alhajj R. Genetic algorithm based framework for mining fuzzy association rules[
695 J]. Fuzzy Sets and Systems, 2005, 152(3): 587-601.
- 696 23. Yan X, Zhang C, Zhang S. Genetic algorithm based strategy for identifying association
697 rules without specifying actual minimum support[J].
698 Expert Systems with Applications, 2009, 36(2): 3066-3076.
- 699 24. Kaya M. Multiobjective genetic algorithm based approaches for mining optimized fuzzy a
700 ssociation rules[J]. Soft Computing, 2006, 10(7): 578-586.
- 701 25. SallebAouissi A, Vrain C, Nortet C. QuantMiner: A Genetic Algorithm for Mining Quantita
702 tive Association Rules[C], IJCAI, 2007.
- 703 26. Saggar M, Agrawal A K, Lad A. Optimization of association rule mining using improved g
704 enetic algorithms[C],
705 Systems, Man and Cybernetics, 2004 IEEE International Conference on. IEEE, 2004, 4:
706 3725-3729.
- 707