Association Rules Extraction using Multi-objective Feature of Genetic Algorithm

Mohit K. Gupta and Geeta Sikka

Abstract—Association Rule Mining is one of the most well - liked techniques of data mining strategies whose primary aim is to extract associations among sets of items or products in transactional databases. However, mining association rules typically ends up in a really large amount of found rules, leaving the database analyst with the task to go through all the association rules and find out the interesting ones. Currently Apriori Algorithm plays an important role in deriving frequent itemsets and then extracting association rules out of it. However Apriori Algorithm uses Conjunctive nature of association rules, and single minimum support threshold to get the interesting rules. But these factors don't seem to be alone sufficient to extract interesting association rules effectively.

Hence in this paper, we proposed a completely unique approach for optimizing association rules using Multi-objective feature of Genetic Algorithm with multiple quality measures i.e. support, confidence, comprehensibility and interestingness. A global search might be achieved using Genetic Algorithm (GA) in the discovery of association rules as GA relies on the greedy approach. The experiments, performed on numerous datasets, show a wonderful performance of the proposed algorithm and it will effectively reduce the quantity of association rules.

***Index Terms*—Apriori Algorithm, Association Rule, Comprehensibility Measure, Genetic Algorithm**

I. INTRODUCTION

DATA Mining is a very active and apace growing research area in the field of computer science. Knowledge Discovery in Databases (KDD) has been a vigorous and attractive research challenge both in the areas of computing and Data Mining. Its aim is to discover interesting and useful data from an oversized variety of data stored in the transactional databases. Association Rule Mining is one of the most well-known methods for such knowledge discovery. It can effectively extract interesting relations among attributes from transactional databases to help out in decision making. A widely accepted definition is introduced by Fayyad et al. [1] in which knowledge discovery is defined as the non-trivial process of discovering valid, novel, useful and interesting patterns in database. This definition focuses on KDD as a complex process having variety of steps. Data Mining is one such step during this process where intelligent techniques are applied so as to extract interesting data patterns [2].

Manuscript received July 8, 2013; revised July 15, 2013.

Mohit K. Gupta is an M.Tech Scholar with the Dr. B R. Ambedkar National Institute of Technology Jalandhar, 144011, Punjab (India). (Contact No. 9781277505; email id: mohitguptaakg@gmail.com).

Dr. Geeta Sikka is currently working as an Associate professor in Dr. B.R. Ambedkar National Institute of Technology Jalandhar, 144011, Punjab (India). (Contact No. 9888582299; email id: sikkag@gmail.com).

In this paper we have thought about Association Rule Mining and tried to improve this technique by applying Genetic Algorithm on the rules generated by Association Rule Mining algorithms such as Apriori Algorithm.

A brief introduction about Association Rule Mining and GA is given in the following sub-sections. Proposed methodology is described in Section II, which will elaborate the implementation details of Association Rule Mining using GA. In section III, we will discuss about the experimental results and its analysis; and conclusion and scope for future work is given in the last section IV.

A. Association Rule Mining

Since its introduction [3], the area of Association Rule Mining has got a huge deal of attention. Association rules are intended to identify strong rules discovered in transactional databases using different measures of interestingness and for discovering regularities and correlation between products in large-scale transaction data recorded by point of sale (POS) systems in supermarkets.

An Association Rule is an implication of the form $X \rightarrow Y$, where X is called antecedent, Y is called consequent Both X and Y are frequent item-sets in a transactional database and $X \cap Y = \emptyset$. The rule $X \rightarrow Y$ can be interpreted as "if itemset X occurs in a transaction, then itemset B will also be there in the same transaction". For example, suppose in a database 45% of all transactions contain both beer and snack and 85% of all transactions contain beer. An Association Rule Mining system might derive the rule beer → snack with 45% support and 85% confidence. Rule support and confidence are two important quality measures of rule interestingness.

A confidence of 85% means that 85% of the customers who purchased beer also bought snack. Typically, association rules are considered interesting if they satisfy both the minimum support criteria and minimum confidence criteria [2]. These criteria are set by users or by experts. Those rules having support and confidence greater than or equal to the user specified criteria are extracted by association rule discovery task.

Extracting Association Rules is not full of merits; it also has some limitations, first the number of generated rules grows exponentially with the number of items or products, it means that Association rule mining has algorithmic complexity. But this complexity can be overcome by some latest algorithms which can efficiently reduce the search space. Secondly, the problem of extracting interesting rule from set of rules. Hence in this paper we tried to overcome these issues. For first problem we apply Genetic Algorithm for reducing the number of rules since GA finds better solution as it perform global search and it cope better with attribute interaction than the greedy rule induction

ISBN: 978-988-19253-1-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

techniques used in Data Mining; And for second problem we can discover useful association rules from set of rules through development of useful quality measures on the set of rules [4], [5].

Therefore in this paper we will apply proposed Genetic Algorithm on the rules that were generated by Apriori Algorithm in order to achieve above mentioned objectives.

B. Genetic Algorithms

Genetic Algorithms (GAs) are a family of optimization methods based on biological mechanisms [6], such as, Mendel's laws and Darwin's principle of natural selection. It imitates the mechanics of natural species evolution with biological science principles, like natural selection, crossover and mutation. A GA searches for good solutions to a problem by maintaining a population of candidate solutions and making subsequent generations by choosing the current best solutions and using operators like Crossover and Mutation to create new candidate solutions. Thus, better and better solutions are "evolved" over time. Commonly, the algorithm terminates when either a maximum number of generations has been made, or a satisfactory fitness level has been reached for the population [6], [7], [8]. The advantage of GA becomes clearer once the search space of a task is enormous [9].

The GAs are important when discovering association rules because the rules that GA found are usually more general due to its global search nature to find the set of items frequency and they are less complex than other induction algorithms usually used in data mining, where these algorithms usually performs a kind of local search[10]. As a result of their global search, GAs tend to cope better with attribute interactions than inductions algorithms [9], [11]. Wilson soto *et al.* [12] designed a method which uses a unique kind of crossover known as subset size–oriented common feature (SSOCF) and permits the sets of useful information continuance in order to be inherited, regardless the number of generations individuals have. Peter P. wakabi-waiswa *et al.* [13] given a new approach known as MOGAMAR to generate high quality association rules with five quality metrics i.e. confidence, support, interestingness, lift, J-measure. A completely unique association rules approach base on GA and fuzzy set strategy for web mining is presented in [14]. It is based on a hybrid technique that combines the strengths of rough set theory and GA. GAs for Multi-objective Rule Mining is proposed in [15]. In their work they used alternative useful other i.e. comrehensibility and interestingness. In addition to the predictive accuracy, M. Anandhvalli *et al.* [16] presented a method find all the potential optimized rules from given dataset using GA. In their work they designed a system that can predict the rules which contain negative rule in the generated rules along with more than one attribute in consequent body.

II. PROPOSED METHODOLOGY

In this work, a Multi-objective Genetic Algorithm approach is used for the automated extracting of interesting association rules from large datasets. In this section we will discuss the representation of rules (encoding), genetic operators, and fitness function used in this proposed work as given below:

A. Representation of Rule and Encoding Scheme

To apply GA, initially an accepted encoding needs to be chosen to represent candidate solution to the given problem. Representation of rules plays a significant role in GAs; mainly there are two approaches of how rules are encoded in the population of individuals. One such technique is Michigan approach [9], in which each rule is encoded into an individual. Second technique is referred to as Pittsburg approach [9], where a set of rules are encoded into an individual. In this paper, we opted Michigan's approach [9] i.e. each individual is encoded into a single rule. The structure of an individual is made up of genes and is represented as:

Suppose there are *n* predicting attributes in the data being mined. An individual (sequence of genes) corresponds to a single association rule is divided into two parts: antecedent body consisting of a conjunction of conditions on the values of the predicting attributes, and consequent body consisting of conjunction of conditions on the values of predicting attributes.

For any rule the set of attributes forming antecedent body and the set of attributes forming consequent body would be disjoint, i.e. (set of attributes present in the antecedent body) ∩ (set of attributes present in the consequent body)= Ø.

In mathematical form: If a rule is composed of the form $X \rightarrow Y$ then $X \cap Y = \emptyset$.

The genes are come according to their position i.e. the first gene represents the first attribute similarly the second gene represents the second attribute and so on. If an attribute is absent in the rule then the corresponding value in gene is "#".The structure of individual is shown in Fig.1.

Antecedent Body	**Consequent Body**
n Predicting Attributes	n Predicting Attributes

Fig. 1. The Structure of Individual.

For example, consider the following Balloon Dataset given in Table I.

TABLE I
DESCRIPTION OF THE BALLOON DATASET

Attributes	**Values**	**Allies**
Colour	Yellow, Purple	'1','2'
Size	Large, Small	'1','2'
Act	Stretch, Dip	'1','2'
Age	Adult, Child	'1','2'
Class Inflated	Yes, No	'1','2'

According to above described Balloon Dataset, an Association Rule:
If Age= Adult ∧ Act= Dip → inflated = True, would be encoded as:

Color	Size	Age	Act	Inf.	Color	Size	Age	Act	Inf.
#	#	1	2	1	#	#	#	#	#

B. Genetic Operators

Genetic Operators are some of the most essential components of GAs. Standard GA applies Genetic Operators such as Selection, Crossover and Mutation on an initially random population in order to compute an entire generation of new strings. GA runs to come up with better

ISBN: 978-988-19253-1-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

solutions for the next generations. The probability of an individual reproducing is proportional to the goodness of the solution it represents. Therefore quality of the solutions in the next generations increases. The process is terminated when either an appropriate solution is found or the criteria for maximum number of generations has been reached.

The function of genetic operators is as follows:

i) Selection: The selection operator chooses a chromosome in the current population according to the fitness function and copies it without changes into the new population. The selection of member from the population can be done with number of selection methods. In this paper we used Roulette Wheel Sampling Procedure.

Roulette Wheel Sampling is a process of choosing members from the population of chromosomes in a way that is proportional to their fitness. It does not give assurance of that the fittest member goes through to the next generation; however it has a very good chance of doing so.

In Roulette Wheel method following steps is applied:

- The population is sorted by descending fitness values.
- Accumulated normalized fitness values are calculated. The accumulated fitness of the last individual should be 1 (otherwise something went wrong in the normalization step).
- A random number *R* between 0 and 1 is chosen and evaluated; and
- Corresponding to this value and the fitness normalized value, the candidate is selected.

ii) Crossover: The crossover operator used to produce two new chromosomes from two selected chromosomes by swapping segments of genes according to a certain probability. Crossover is a genetic operator used to vary the programming of a chromosome or chromosomes from one generation to the next. There are many types of crossover exist for organisms which use different data structures to store themselves. In this paper, we used one-point crossover with crossover probability of 95%.

In one point crossover, a single crossover point on both parents' organism strings is selected. All data beyond that point in either organism string is swapped between the two parent organisms. The resulting organisms are the children.

For example two chromosomes are given in Fig.2.

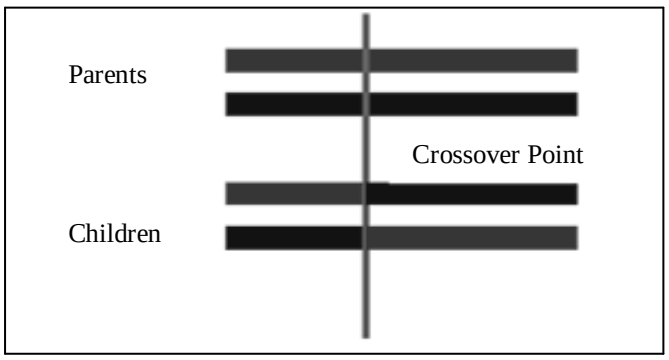

Fig.2. Example of One-Point Crossover

iii) Mutation: It works on the bits of individuals. It adds the information in a random manner that introduces diversity in the population. Mutation alters one or more gene values in a chromosome from its initial state. This is the chance that a bit within a chromosome will be mutated from 0 to 1 and 1 to 0. This can result in extremely new gene values being added to the gene pool. Hence GA can come to better solution by using mutation operator.

C. Fitness Function

Fitness function is a particular type of objective function that is used to summarize as how close a given design solution is to achieving the set aims. It is very important to define a good fitness function that rewards the right kinds of individuals. The fitness function is always problem dependent. Multi-objective processing can be fostered for extracting the interesting association rules. Based on that, in this present work, four significant measures of the rules such that support, confidence, simplicity and interestingness are considered. These metrics are converted into an objective fitness function with user-defined weights. Using these four measures, some previously unknown, easily understandable and compact rules can be generated. So, Association Rule Mining problems can be thought of as a Multi-objective problem instead of as a single objective one [15].

The support $\sigma(X)$, of an item-set X is defined as the proportion of transaction in the dataset which contain the itemset. The support can be formulated as:

$$S = \frac{\sigma(X \cup Y)}{\sigma(N)} \quad (1)$$

Where $\sigma(N)$ is the total number of transactions and $\sigma(X \cup Y)$ is the number of transactions containing both X and Y. Support is typically used to eliminate non-interesting rules.

A measure to predict the association rule precision is the confidence or predictive accuracy. It measures the conditional probability of the consequent given the antecedent and formulated as:

$$C = \frac{\sigma(X \cup Y)}{\sigma(X)} \quad (2)$$

Where $\sigma(X)$ is the number of transactions containing X. A higher confidence suggests a strong association between X and Y. Although confidence favours the rules overfitting the data [17].

The generated rule may have a large number of attributes involved in the rule, thus making it difficult to comprehend. If the discovered rules are not simple and comprehensible to the user, the user will never use them. So the Comprehensibility (comp.) measure is needed to make the discovered rules easy to understand. The comprehensibility tries to quantify the understandability of the rule. Comprehensibility of an association rule can be defined by the following expression:

$$\text{Comp} = \frac{\log(1 + |Y|)}{\log(1 + |X \cup Y|)} \quad (3)$$

Where $|Y|$ and $|X \cup Y|$ are the number of attributes involved in the consequent body and the total rule respectively. If the number of conditions in the antecedent body is less, the rule is considered as more simple.

ISBN: 978-988-19253-1-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Interestingness of a rule, denoted by Interestingness X→Y, is used to quantify how much the rule is surprising for the users. As the most important point of rule mining is to find some hidden information, it should discover those rules that have comparatively less occurrence in the database. The following expression can be used to define the interestingness [15].

$$\text{Interestingness } X \rightarrow Y = \frac{\text{Sup}(X \cup Y)}{\text{Sup}(X)} \times \frac{\text{Sup}(X \cup Y)}{\text{Sup}(Y)} \left(1 - \frac{\text{Sup}(X \cup Y)}{\sigma(N)}\right) \quad (4)$$

Where $\sigma(N)$ indicates the total number of transactions in the database.

As described above, Association Rule Mining is considered as Multi-objective problem rather than Single Objective one. So, the fitness function is defined as:

$$F = ((W_1 \times \text{Sup}) + (W_2 \times \text{Con.}) + (W_3 \times \text{Comp}) + (W_4 \times \text{Interest.})) / ((W_1 + W_2 + W_3 + W_4)) \quad (5)$$

Where W_1 , W_2 , W_3 and W_4 are user-defined weights. Since finding the frequent itemsets for any given transactional database is of huge computational complexity, the problem of extracting association rules can be reduced to the problem of finding frequent itemsets. On this basis, in this work the weight values of W_1 = 4, W_2 =3, W_3 =2 and W_4=1 were taken according to the relative importance of the quality measures support, confidence, comprehensibility and interestingness. It is noted that fitness values should be in the range of [0…1].

D. Algorithmic Structure

In this section, we are presenting structure of the proposed algorithm. The GA is applied over the rules fetched from Apriori algorithm. The procedure of the proposed algorithm for generating optimized association rule through GA is as follows:

1. Start
2. Import a dataset from UCI Machine Learning Repository that fits into memory.
3. Apply Apriori Algorithm to find the frequent item-sets. Suppose A is set of the frequent item-set generated by Apriori Algorithm.
4. Set Z=Φ where Z is the output set, that contains all generated association rules.
5. Set the termination condition of Genetic Algorithm.
6. Represent each item set of A in above described encoding scheme.
7. Apply Genetic Algorithm on selected members to generate association rules.
8. Evaluate the fitness function for each rule X→ Y.
9. If fitness function satisfies the selection criteria then
10. Set Z= Z∪{X→ Y}.
11. If the desired number of generations is not completed, go to step 3.
12. Stop.

The flow chart of the proposed algorithm is shown in Fig.3.

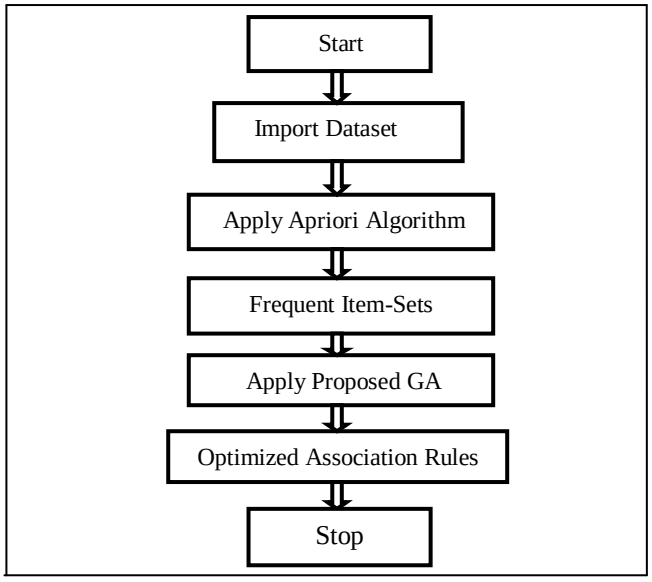

Fig.3. Block Diagram of Proposed Algorithm

III. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed algorithm is implemented using MATLAB 2012 MathWorks, Inc. software tool with 3GB RAM and 2.67 GHz processor. The performance of the proposed approach is tested on four datasets collected from UCI Machine Learning Repository [18], which is a collection of widely used real-world datasets for Data Mining and KDD community. For each dataset the proposed GA had 100 individuals in the population and was executed for 200 generations. The proposed algorithm was terminated when the maximum number of generations has reached. The performance of proposed algorithm is evaluated and compared with the well-known Apriori Algorithm and previous technique proposed by M. Ramesh et al. [19]. The default parameters of the Apriori Algorithm and proposed GA are used to make the comparison completely fair. The results for four datasets are an average over 10 executions. The summary of used datasets is given in Table II.

TABLE II
SUMMARY OF DATASETS

Dataset	Instances	Attributes
Adult	48842	14
Chess	3196	36
Wine	178	13
Zoo	101	17

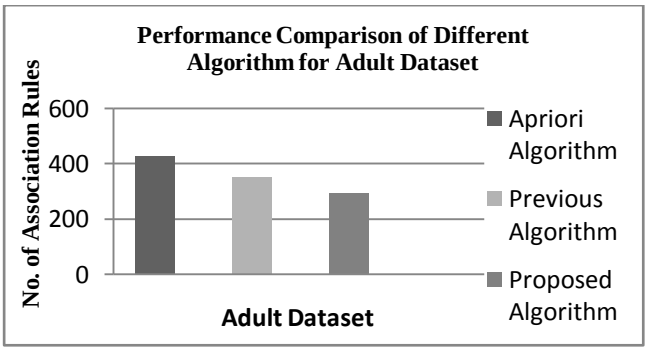

Fig.4. Performance Comparison of Different Techniques for Adult Dataset. It shows number of association rules generated using Apriori Algorithm, previous algorithm and proposed algorithm.

ISBN: 978-988-19253-1-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

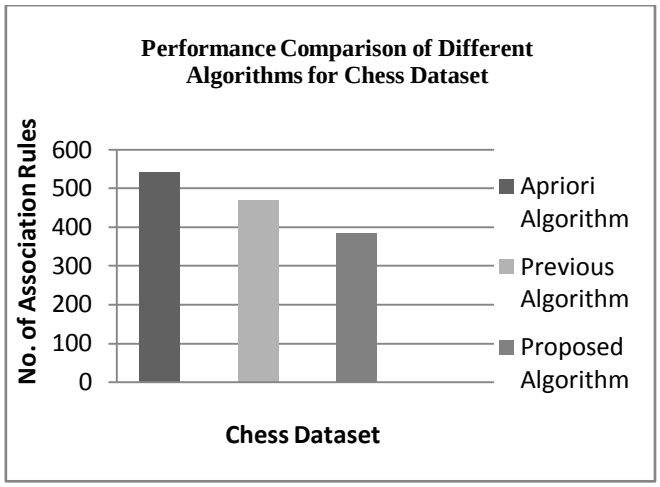

Fig.5. Performance Comparison of Different Techniques for Chess Dataset. It shows number of association rules generated using Apriori Algorithm, previous algorithm and proposed algorithm.

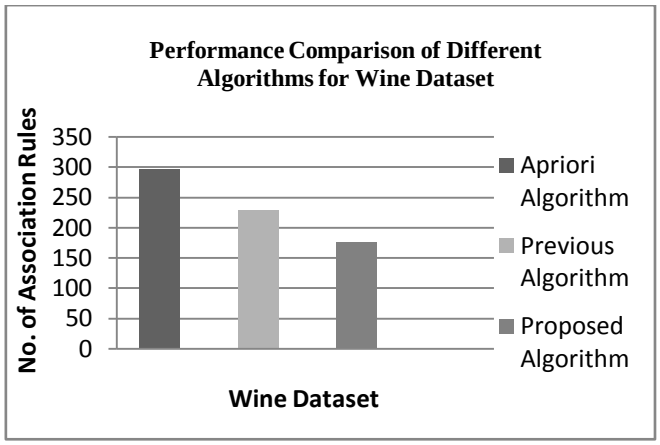

Fig.4. Performance Comparison of Different Techniques for Wine Dataset It shows number of association rules generated using Apriori Algorithm, previous algorithm and proposed algorithm.

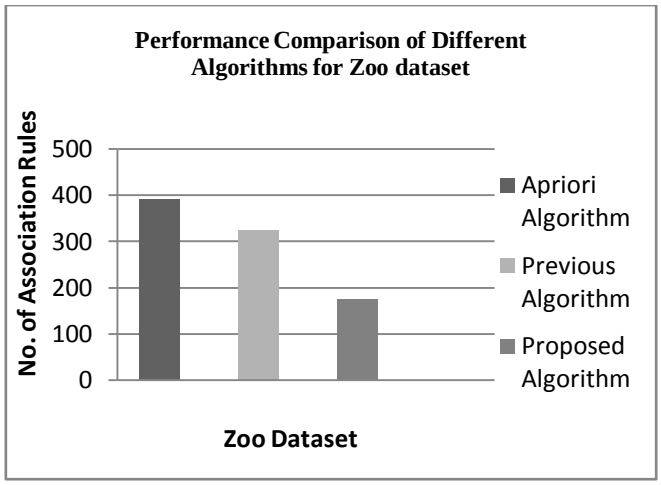

Fig.7. Performance Comparison of Different Techniques for Zoo Dataset. It shows number of association rules generated using Apriori Algorithm, previous algorithm and proposed algorithm.

Fig.4-7 shows the performance comparison of proposed algorithm with Apriori Algorithm and previous algorithm discussed in [19].

TABLE III
AVERAGE PERFORMANCE OF TWO ALGORITHMS

Dataset	Algorithms	Sup	Conf
Adult	The Proposed Algorithm	.343	1.000
	The Previous Algorithm	.318	1.000
Chess	The Proposed Algorithm	.283	.920
	The Previous Algorithm	.253	.908
Wine	The Proposed Algorithm	.095	.740
	The Previous Algorithm	.153	.680
Zoo	The Proposed Algorithm	.142	.603
	The Previous Algorithm	.127	.540

Table III shows the comparison of two algorithms based on average performance. The values of support and confidence in the above table refer to the total average of support and confidence for the discovered rules respectively using the proposed algorithm and previous algorithm for the dataset we chosen.

Fig.8 and Fig.9 depicts the comparative performance of two algorithms based on the average support and confidence value respectively.

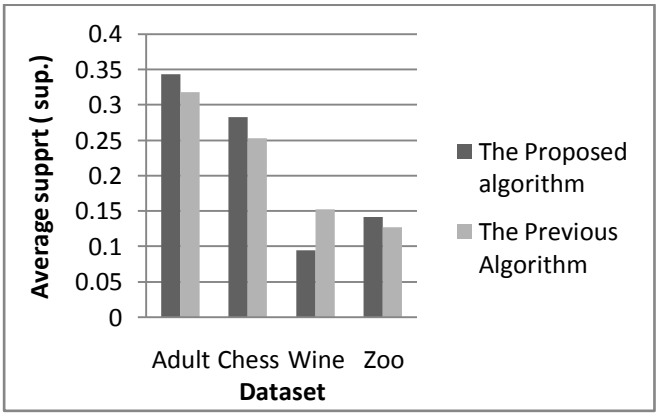

Fig. 8. Average Support of Extracted Rules by Proposed Algorithm and Previous Algorithm

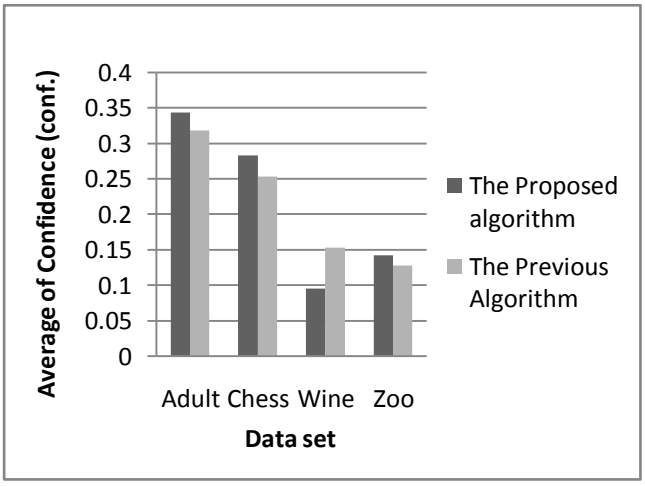

Fig.9. Average Confidence of Extracted Rules by Proposed Algorithm and Previous Algorithm

The above results show that the proposed algorithm performs better than the previous algorithm. Except in wine dataset, the average of support in the previous algorithm is better than the proposed algorithm. Since it is a Multi-

ISBN: 978-988-19253-1-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

objective problem, we can't prioritize one objective over another.

IV. CONCLUSION AND SCOPE FOR FUTURE WORK

Although a variety of works has already been published in this area, however in this research paper we have tried to use Multi-objective feature of GA for discovering the association rules. When proposed algorithm is applied on different datasets, we get results containing desired rules with maximum accuracy and interestingness. The resulted accuracy of the generated rules is 100%. It has been observed that proposed algorithm can attain considerable performance improvement in terms of the interesting association rules extraction. The numbers of rules generated by proposed algorithm are significantly less as compared to Apriori Algorithm and previous technique [19]. Hence we can say proposed algorithm optimize the association rule efficiently and effectively.

As for future work, we are currently working on GA in parallel for optimization of Association Rule Mining through which we can further improve its complexity.

ACKNOWLEDGEMENT

Author gratefully acknowledges the authorities of Dr. B. R. Ambedkar National Institute of Technology Jalandhar for providing facilities and encouragement to carry out this work.

REFERENCES

[1] U. Fayyad and R. Uthurusamy, "Data Mining and Knowledge Discovery in Databases", *Communications of the ACM*, vol. 39, no. 11, 1996, pp.24–34.

[2] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann, 2006.

[3] R. Agrawal, T. Imielinski and T. Swami, "Mining Association Rules Between Sets Of Items In Large Databases", In *Proceedings of ACM SIGMOD International Conference on Management of Data (SIGMOD' 93)*, 1993, pp. 207–216.

[4] C. Silverstein, S. Brin, R. Motwani and J.D. Ullan, "Scalable Techniques For Mining Causal Structures", In *the Proc. of ACM SIGMOD Int'l Conf. on Management of Data, Seattle, Washington, USA*, 1998.

[5] S. Brin, R. Motwani, and C. Silverstein, "Beyond Market Baskets: Generalising Association Rules to Correlations", In *the Proc. of the ACM SIGMOD Int'l Conference on Management of Data (ACM SIGMOD '97)*.

[6] D. E. Goldberg, "Genetic Algorithms in Search, Optimization and Machine Learning", *Addison-Wesley*, 1989.

[7] S. N. Sivanandamand and S. N. Deepa, "Introduction to Genetic Algorithms", *Springer-Verlag Berlin Heidelberg*, 2008.

[8] K. K. Bharadwaj, N. M. Hewahi and M. A. Brando, "Adaptive Hierarchical Censored Production Rule-Based System: A Genetic Algorithm Approach", *Advances in Artificial Intelligence, SBIA '96, Lecture Notes in Artificial Intelligence*, no. 1159, Berlin, Germany, Springer-Verlag, 1996, pp. 81-90.

[9] Alex A. Freitas, "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery" Postgraduate Program in Computer Science, Pontificia Universidade Catolica do Parana Rua Imaculada Conceicao, 1155. Curitiba - PR. 80215-901. Brazil.

[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, 2009, pp. 3066–3076.

[11] S.Dehuri and R Mall, "Predictive and Compehensible Rule Discovery using a Multi-objective Genetic Algorithm", *Knowledge Based Systems*, vol. 19, 2006, pp. 413-421.

[12] W. Soto and A. Olaya-Benavides, "A Genetic Algorithm for Discovery of Association Rules." In *Computer Science Society (SCCC)*, 2011, pp. 289-293.

[13] 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*, 2009, pp. 73-78.

[14] C. Chai and B. Li, "A Novel Association Rules Method Based on Genetic Algorithm and Fuzzy Set Strategy for Web Mining", *Journal of Computers*, vol. 5, no. 9, 2010, pp. 1448-1455.

[15] A. Ghosh and B. Nath, "Muti-Objective Rule Mining using Genetic Algorithms", *Information Sciences*, vol. 163, 2004, pp. 123–133.

[16] M. Anandhavalli and S. Kumar Sudhanshu, A. Kumar and M.K. Ghose, "Optimized Association Rule Mining Using Genetic Algorithm", *Advances in Information Mining*, vol. 1, issue 2, 2009, pp. 01-04.

[17] Y. Cheung and A. Fu, "Mining Frequent Itemsets without Support Threshold: with and without Item Constraints", *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, 2004, pp. 1052-1069.

[18] UCI Repository of Machine Learning Databases, Department of Information and Computer Science University of California, 1994, Available: [http://www.ics.uci.edu /~mlearn/MLRepositry.html].

[19] M. Ramesh Kumar and Dr. K. Iyakutti, "Application Of Genetic Algorithms For The Prioritization Of Association Rules", *IJCA Special Issue on Artificial Intelligence Techniques- Novel Approaches & Practical Applications (AIT)*, 2011, pp. 1-3.