Association Rule Mining Using Firefly Algorithm

Poonam Sehrawat
Dept. of Computer Science
Banasthali University, Banasthali

Manju
Department of Computer Engg.
CDL Govt. Polytechnic Education Society
Nathusari Chopta, Sirsa

Harish Rohil
Department of Computer Science& Applications
Chaudhary Devi Lal University, Sirsa

Abstract: Data mining is the process of extracting previously unknown patterns from large amount of data. Association rule mining is one of very important data mining models. Swarm intelligence is a new subfield of artificial intelligence which studies the collective behavior of groups of simple agents. In this paper, a new efficient approach is proposed for exploring high-quality association rules. The proposed approach is based on firefly algorithm, which is an optimization algorithm used in swarm optimization. The proposed approach mines interesting and understandable association rules in single run without using the minimum support and the minimum confidence thresholds. The proposed approach was implemented using Microsoft Visual Studio 4.0 to prove its efficiency in terms of computation time.

Keywords: Data mining, Association rule mining, Firefly algorithm.

1. INTRODUCTION

Data mining has attracted a great deal of attention in the information industry, scientific analysis, business application, medical research and in society due to huge amounts of data. Data mining is the process of extracting previously useful, meaningful and unknown knowledge from the large database. Data mining is also called knowledge discovery in database (KDD). KDD is the process of turning the low-level data into high-level knowledge [12].

Data mining can be classified into several techniques, including association rules, clustering and classification, time series analysis and sequence discovery. Among these techniques, association rule mining is the most widely significant method for extracting useful and hidden information from large amount of database. The task of association rule mining in large database is to find out the frequent patterns or itemsets and also find out the interesting relationship between frequent itemsets. The following is a formal statement of the problem [1]: Consider I= {i1, i2,……, im}called set of items. D is a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. An association rule is an implication of the form $X \Rightarrow Y$ in which X represents the antecedent part of rule and Y represents consequent part of the rule, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$. The itemset which satisfies the minimum support is called frequent itemset. However, the mining of association rule depends upon two parameters:

(1) Support: The support of rule $X \Rightarrow Y$ is the fraction of transactions in D, containing both X and Y.

$$\text{Support } (X \Rightarrow Y) = \frac{(X \cup Y)}{|D|} \quad (1)$$

where |D| is the total number of records in the database.

(2) Confidence: The confidence of rule $X \Rightarrow Y$ is the fraction of transactions in D containing X that also contain Y and indicates the strength of rule.

$$\text{Confidence } (X \Rightarrow Y) = \frac{\text{Sup}(X \cup Y)}{\text{Sup}(A)} \quad (2)$$

Rules that satisfy both a minimum support threshold (min_sup) and minimum confidence threshold (min_conf) are called strong. These values can be set by users or domain experts.

In recent years, some evolutionary algorithms have been used with multi-objective functions for extracting interesting rules. Evolutionary algorithms such as genetic algorithm, ant colony, and simulated annealing and particle swarm optimization have been used for mining association rules. This paper proposed firefly algorithm for discovering the best rules without considering the minimal support and confidence and extract valuable rules in once executed, rather than previous methods that extract association rules in two stages.

Lampyridae is a family of insects that is capable to produce natural light (bioluminescense) to attract a mate or a prey. They are commonly called fireflies or lightning bugs. In the species of Lampyris noctiluca, the fireflies are also known as glow-worms. In this species, it is always the female who glows, and only the male has wings. In other species, Luciola lusitanica, both male and female firefly may emit light and both have wings. If a firefly is hungry or looks for a mate its light glows brighter in order to make the attraction of insects or mates more effective. [3][4][5]. Operation performed during Firefly algorithm (FA) is shown in Figure 1.

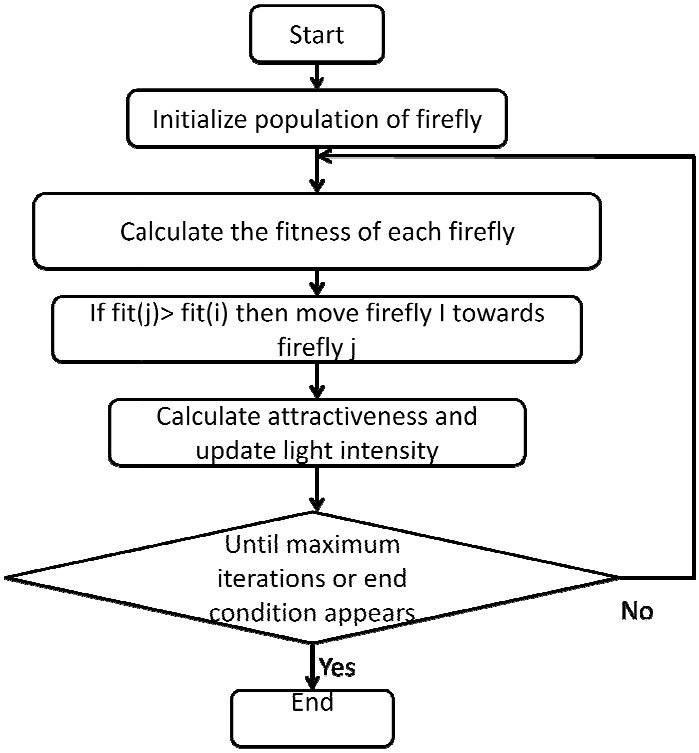

Figure 1: DFD of Firefly Algorithm

2. RELATED WORK

Association rule mining, one of the important and well researched techniques of data mining, used to find out relationships among data items in the database and some hidden relationships exist between purchased items in transactional databases, was first proposed by Agrawal in 1993. Support and Confidence are two important interesting measures. There are many algorithms for association rule mining in literature .Some of these is follows as:

An algorithm for finding all association rules known as the AIS algorithm [1] and another algorithm is SETM algorithm [2]. The Aprior algorithm (Agrawal, Imielinski & swami, 1993; Agrawal & Srikant, 1994; Agrawal & Shafer, 1996) is best-known basic algorithms of association rule mining. The Apriori and AprioriTid algorithms generate the candidate itemsets to be counted in a pass by using only the itemsets found large in the previous pass - without considering the transactions in the database. Apriori uses the downward closure property of itemset support to prune the itemset lattice – the property that all subsets of a frequent itemset must themselves be frequent. Thus only the frequent k-itemsets are used to construct candidate (k + 1)-itemsets [6]. The AprioriTid algorithm has the additional property that the database is not used at all for counting the support of candidate itemsets after the first

pass. Rather, an encoding of the candidate itemsets used in the previous pass is employed for this purpose. In later passes, the size of this encoding can become much smaller than the database, thus saving much reading effort [6].
Since the processing of the Apriori algorithm takes more time, its computational efficiency is a very important issue. In order to improve the efficiency of Apriori, many researchers have proposed modified association rule-related algorithms such that Partition algorithm [8], Sampling algorithm [9], Pincer-Search [9], Éclat [10], etc.
Yan et al proposed association rules based on the genetic algorithm without considering minimum support. In this method, relative confidence is used as the fitness function and only the most interesting rules are returned according to the interestingness measure defined by the fitness function. The fitness function of this method is easily trapped into local optimum, and hence many rules are generated [13]. This problem is improved by other researchers by using multi-objective fitness function [14].
Ghosh et al proposed multi-objective rule mining using genetic algorithm. In this method, use three measures like support count, comprehensibility and interestingness, used for evaluating a rule can be thought of as different objectives of association rule mining problem and also used genetic algorithm to extract some useful and interesting rules from any market-basket type database [15].
Kuo et al proposed association rule mining using particle swarm optimization. In this method, particle swarm optimization algorithm first searches for the optimum fitness value of each particle and then finds corresponding support and confidence as minimal threshold values after the data are transformed into binary values. These minimal threshold values are used to extract valuable rules. But, there is no studies exist to extract the association rules by using the firefly algorithm. Using this approach tried to find out the high-frequency association rules [16].
Firefly Algorithm (FA) is among the most powerful algorithms for optimization. The Firefly Algorithm was developed by Yang is inspired by biochemical and social aspects of real fireflies and it was based on the idealized behaviour of the flashing characteristics of fireflies. Most fireflies produce short and rhythmic flashes are to attract mating partners (communication), and to attract potential prey. In addition, flashing may also serve as a protective warning mechanism. The following three rules are idealized for flashing characteristics of FA [17]:
All fireflies are unisex so that one firefly is attracted to other fireflies regardless of their sex.
Attractiveness is proportional to their brightness, thus for any two flashing fireflies, the less bright one will move towards the brighter one. The attractiveness is proportional to the brightness and they both decrease as their distance increases. If no one is brighter than a particular firefly, it moves randomly.
The brightness or light intensity of a firefly is determined by the landscape of the objective function to be optimized
Feature selection methods fail to find optimal data reductions or require more time to achieve better results. Hema et al. proposed a new feature selection algorithm (FA_RSAR) that incorporates the basic behavior of firefly algorithm with RST to improve the performance [18].

3. Proposed Association Rule Mining Using Firefly Algorithm (ARMFA)

In this work, firefly algorithm which is one of the new evolutionary algorithms is applied to explore association rules from transactional databases. This approach is called ARMFA. The following of this section are some important parts of the algorithm which are explained: encoding, fitness function, and the last part of the section explain the pseudo-code of ARMFA.

3.1. Encoding of Fireflies

In this paper, each firefly represents a rule and each rule contains of a series of decision variables which represent the status of every item in the rule. According to Figure 2 in the proposed algorithm, every firefly has n decision variables in lieu of n items in any dataset.

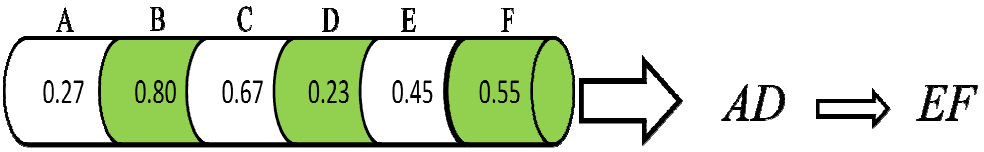

Figure 2: Encoding of Fireflies

This means that the ith variable which is known as ESi indicates the status of ith item and can take values between 0 and 1. In this way, if 0.00≤ ESi ≤0.33, the ith attribute is in the antecedent of the rule and if 0.33≤ESi≤0.66, this attribute is in the consequence of the rule and if 0.66≤ESi≤1, it means the lack of ith attribute in the rule.

3.2. Fitness Function

The fitness function provided in this study is in Eqs. (3)

$$fit(i)=\square 1[\frac{Sup(A\cup C)}{Sup(A)}].[\frac{Sup(A\cup C)}{Sup(C)}].[1-\frac{Sup(A\cup C)}{|D|}]+\square 2\frac{NumberField(i)}{MaxField} \quad (3)$$

Since the mining association rule is a task that extracts some hidden information, it must discover those rules that have a comparatively less occurrence in the entire database which are more interesting for the users; discovering such rules is more difficult. For classification rules, it can be defined by information gain theoretic measures. But it is not efficient for evaluating the association rules. Therefore, interestingness measure [15] is used in the fitness function according to the first parameter. In this parameter |D| is the total number of records in the database. The first parameter has three parts:

i). $[\frac{Sup(A\cup C)}{Sup(A)}]$ shows the probability of creating the rule depending on the antecedent part;

ii). $[\frac{Sup(A\cup C)}{Sup(C)})]$ shows the probability of creating the rule depending on the consequent part and iii). $[\frac{Sup(A\cup C)}{|D|}]$ gives the probability of generating the rule depending on the whole data-set. So complement of this probability will be the probability of not generating the rule. Thus, a rule having a very high support count will be measured as less interesting, because such rules easily predictable by user.

In fact most of these are interesting rules in which the rate of acquired information is approximately the same in both antecedent and consequent parts of the rule. In above parameter the support count of the rule antecedent and the support count of the rule consequent are considered.

The second parameter is used for number of attributes in rule and it gives the shorter rules with a smaller number of attributes. Numberfield(i) returns the number of attributes that exist in firefly i. This term leads to shorter rules. In result, comprehensibility of rules that are important in data mining is increased. Larger rules are more likely to contain redundant information.

It should be noted that α1 and α2 will be specified by the percent of user interests and one might increase or decrease the effects of parameters of the fitness function.

3.3. ARMFA Approach

In this approach, each firefly considers as a rule and calculates fitness value of each firefly. Using this approach tried to find out the high-frequency association rules. The proposed association rule mining algorithm is shown in Figure 3.

The proposed algorithm contains two parts:

The first part provides procedures related to encoding and calculating the fitness values of the firefly swarm.

In the second part of the algorithm, which is the main contribution of this study, the firefly algorithm is employed to mine the association rules.

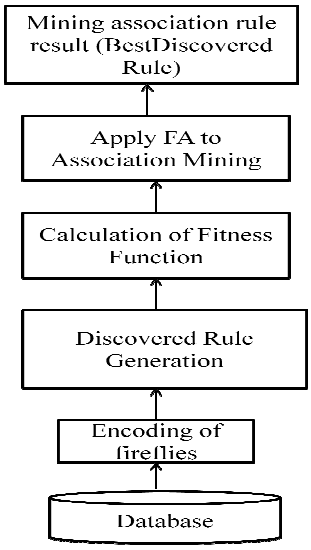

Figure 3: The Proposed Association Rule Mining Algorithm

The pseudo code of above approach is shown in Figure 4. In this, first specify t as an iteration number. The algorithm is run as number of iterations or number of desired rules. At the beginning of the algorithm, the DiscoveredRules set is empty. At the first iteration of algorithm, each firefly is initialized randomly as a rule. In each of iterations, until reaching the termination conditions, calculate fitness of each fireflies. If light intensity of jth firefly is lagre than ith firefy then calculate distance between them. The individuals of population are sorted in descending order according to their distance and then select minimum distance between ith and jth firefly and then move i towards j and increase the intensity of jth firefly

Input: Database of Transaction (D)

Output: Best discovered association rules

1. t=0 // here t is iteration number
2. DiscoveredRules = ∅
3. Generate initial population of fireflies x_i (i=1,2,...n)
4. While(t < no of iterations)
5. Compute objective function of each firefly
6. for i=1: n (all n fireflies)
7. for j=1: i (all n fireflies)
8. if(fit(j) > fit(i))
9. Calculate distance $r_{ij} = \| x_i - x_j \|$
10. end if
11. sort distance in descending order
12. min ← firefly[mindistanceindex]
13. end for j
14. Movement phase:
15. j = mindistanceindex
16. $x_i = x_i + f(x_i)\times(x_j-x_i) + \alpha(rand -1/2)$
17. $I_j = I_j + x_i$ //increase the intensity of firefly j
18. end for i
19. Best = firefly[Bestindex]
20. Until not terminate
21. DiscoveredRules = Best ∪ DiscoveredRules
22. t++ // End of algorithm

Figure 4: Pseudo-code for the Proposed ARMFA

In each iteration, the best discovered rule is added to Best vector and combines with DiscoveredRules. The rule is valid if it has at least one attribute in the antecedent of the rule and one in the consequence of the rule. This process is continued until termination conditions not occur.

4. Experiment

A simulator with GUI was designed and developed with Microsoft Visual Basic 4.0 using C# and run on a 2.10 GHz machine with a relatively RAM of 2GB. This simulator is accepting input in the form of text file. This experiment evaluates the efficiency and usefulness of the proposed algorithm. For conducting the examination, the authors have used the dataset; format of dataset is taken from the Iris dataset. Iris dataset is online available at http://www.ics.uci.edu/~mlearn.The specifications of dataset are given in Table 1. Assessment of proposed approach was made on the basis of execution time refers most literally to the time during which a computer program is executing.

Weighted coefficients α1 and α2 that have been used in fitness function were selected as 1 and 0.2, respectively. These coefficients are specified by the percent of user interests. Hence, they do not need to re-initialize for each database. This experiment evaluates the efficiency and usefulness of the proposed algorithm. For conducting the examination, the authors have the modified Iris dataset.

Table 1: Specifications of Dataset

	Dataset
No. record	50
No. attribute	19

This algorithm extract only those rules that have a comparatively less occurrence in the entire database which are more interesting for the users; discovering such rules is more difficult. There are 42 different DiscoveredRules that came in existence after running above algorithm ARMFA on simulator. After each iteration best rules are generated that have high fitness or intensity value are considered as a best rule from all rules. ARMFA algorithm was run for 10 iterations and 5 Best Discovered Rules were generated as shown in Table 4. Best discovered rules was changed in each run because each time algorithm uses different random value.

Table 2: Best Discovered Rules of ARMFA Approach

Iteration Number	Best Discovered Rules
1	OF ⇒RB
2	OF ⇒RB
3	OF ⇒RB
4	MHC ⇒R
5	SOG ⇒B
6	KH ⇒RB
7	SOG ⇒B
8	OH ⇒RB
9	OH ⇒RB
10	OH ⇒RB

Comparison with the Existing Approaches

The proposed association rule mining using firefly algorithm was compared with a similar existing approach proposed by Kuo et al. [16]. Kuo et al. proposed application of particle swarm optimization to association rule mining. The proposed approach is better than the existing approach due to following reasons:
In existing approach, first find out the minimal support and confidence threshold values by applying particle swarm optimization algorithm and then set these values as minimum support and confidence to extract association rules. This particular makes algorithms dependent on datasets and it must execute several time. But in proposed approach, there is no need of minimum support and confidence to extract association rules.
The specification of existing approach is different from proposed approach. But it is obvious, when dataset of existing approach was run on ARMFA simulator takes less time. At population size of 5 and 10 run time of existing approach is less than the proposed approach. After that run time is exponentially increasing with population size but in proposed approach run time is almost similar for all population size.

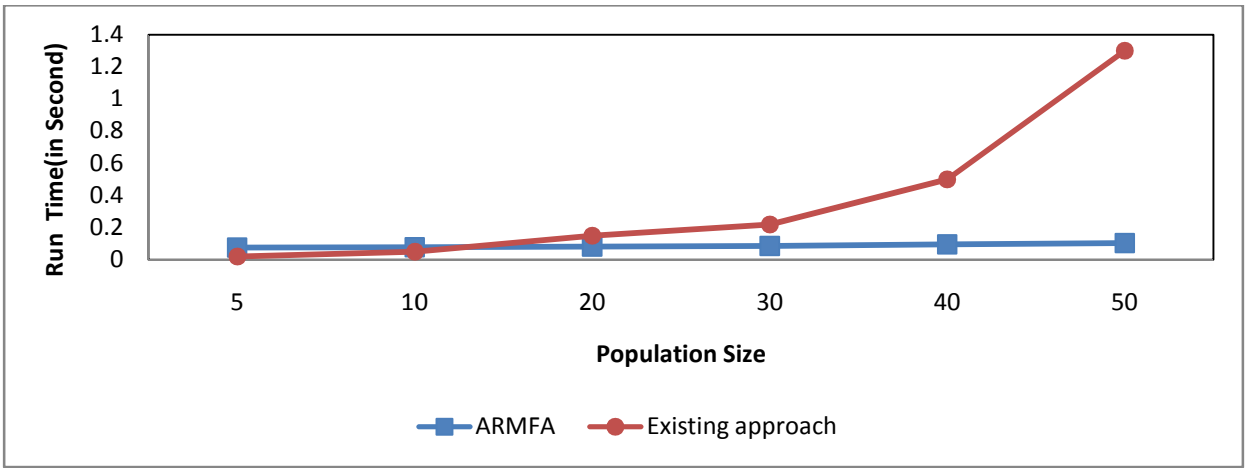

Figure 5: Relationships between Population Size and Computation Time for ARMFA and the Existing Approach Proposed by Kuo et al.

Figure 5 clearly indicates that the proposed ARMFA algorithm outperforms the application of particle swarm optimization in terms of relationship between population size and computation time. In proposed approach there is little variation in run time with respect to population size so association rule mining using firefly algorithm is efficient in terms of execution times with respect to population size.

5. Conclusion

The work reported in this paper presents an efficient approach for exploring high quality association rule. The proposed approach is based on firefly algorithm. In this proposed approach, encoding of fireflies is applied to extract rules from database. For extracting rule, fitness value of individual rule is computed instead of minimum support and minimum confidence thresholds. This has an advantage that the database is scanned once only which enhances efficiency of the approach in terms of CPU time and memory consumption. The proposed approach was implemented using Microsoft Visual Studio 4.0 and the result was compared with existing similar approach proposed by Kuo et al. The proposed ARMFA algorithm betters the application of particle swarm optimization in terms of relationship between population size and computation time.

REFERENCES

[1] R. Agrawal, T. Imielinski, A. Swami, “Mining association rules between sets of items in large databases”, Proceedings of the ACM SIGMOD Conference on Management of Data, 1993, 207-216.
[2] M. Houtsma, and A. Swami, “Set-oriented mining of association rules”, Research Report RJ 9567, IBM Almaden Reseach Center, 1993.
[3] H. Fraga, “Firefly luminescence: a historical perspective and recent developments”, Journal of Photochemical & Photobiological Sciences, 2008, Vol. 7, 146–158.
[4] O. Shimomura, “Bioluminescense: Chemical Principles and Methods”, World Scientific Publishing, 2006.

[5] R.S. Parpinelli, and H.S. Lopes, "New inspirations in swarm intelligence: a survey", International Journal of Bio-Inspired Computation, 2011, Vol. 3, No. 1, 1-16.
[6] R. Agrawal, and R. Srikant, "Fast algorithms for mining association rules", Proceedings of the 20th VLDB Conference, 1994, 487-499.
[7] R. Agrawal, and J.C. Shafer, "Parallel mining of association rules", IEEE Transactions on Knowledge and Data Engineering, 1996, Vol. 8, No. 6, 962-969.
[8] A. Savasere, E. Omiecinski, and S. Navathe, "An efficient algorithm for mining association rules in large database", Proceedings of the 21st VLDB Conference, 1995, 432–444.
[9] H. Toivonen, "Sampling large databases for association rules", Proceedings of the 22nd VLDB Conference, 1996, 134–145.
[10] D.I. Lin, and Z.M. Kedem, "Pincer search: a new algorithm for discovering the maximum frequent set", 1997, 1-17.
[11] M.J. Zaki, S. Parthasarathy, and M. Ogihara, "Parallel Algorithms for Discovery of Association Rules", Data Mining and Knowledge Discovery, 1997, 1, 343–373.
[12] J. Han, and M. Kamber, "Data Mining: Concepts and Techniques", 2nd ed., San Francisco, Morgan Kauffmann Publishers, 2001.
[13] X. Yan, CH. Zhang, and SH. Zhang, "Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support", Elsevier Expert Systems with Applications, 2008, Vol. 36, No. 2, 3066-3076.
[14] H. Qodmanan, M. Nasiri, and B. Minaei-Bidgoli, "Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence", Elsevier Expert Systems with Applications, 2010, Vol. 38, No. 1, 288-298.
[15] A. Ghosh, and B. Nath, "Multi Objective Association Rule Mining Using Genetic Algorithm", Elsevier Information Sciences, 2004, Vol. 163, No. 1, 123–133.
[16] R.J. Kuo, C.M. Chao, and Y.T. Chiu, "Application of particle swarm optimization to association rule mining", Applied Soft Computing, 2011, Vol. 11,326–336.
[17] X.S. Yang, "Firefly algorithms for multimodal optimization", SAGA 2009, Lecture Notes in Computer Science, 5792, 2009, 169-178.
[18] H. Banati, and M. Bajaj, "Fire Fly Based Feature Selection Approach", International Journal of Computer Science Issues, 2011, Vol. 8, Issue 4, No 2, 473-480.