Genetic Based Method for Mining Association Rules

Dr. Bushra Khireibut Jassim
Computer Center, Collage of Economic and Administration,Baghdad University/Baghdad
Email:d.bushra14@yahoo.com

Received on: 29/4/2012& Accepted on: 4/10/2012

ABSTRACT

In this paper genetic based method proposed for mining association rule, the benefit of this method it mining association rule in one step and it does not require the user-specified threshold of minimum support and minimum confidence deciding suitable threshold values of support and confidence is critical to the quality of association rule technology. Specific mechanisms for crossover operators have been designed to extract interesting rules from a transaction database.

The method proposed in this paper is successfully applied to real-world database. The results demonstrate that the proposed algorithm is a practical method for mining association rules.

Keyword: genetic algorithm, association rule, data mining.

طريقة معتمدة على الخوارزميات الجينة لاستكشاف قواعد الارتباط

الخلاصة

في هذا البحث تم استخدام طريقة معتمدة على الخوارزميات الجينية لاستكشاف قواعد الارتباط في خطوة واحدة .الطريقة المقترحة لا تتطلب من المستخدم التحديد المسبق لمعامل الدعم ولا معامل الثقة حيث يتم اكتشاف قيمها بواسطة الخوارزمية من البيانات .علما بان تحديد القيم الاولية من قبل المستخدم لهذان المعاملان امر في غاية الصعوبة.تم التعديل على العمليات الجينية التعابر والطفرة للتتناسب مع مشكلة البحث .الطريقة المقترحة طبقت بنجاح على مشاكل حقيقية واظهرت ان الطريقة المقترحة هي طريقة عملية لاكتشاف قواعد الارتباط.

INTRODUCTION

The search for association rules in data mining has the aim to identify the phenomena that are recurrent in a data set. The solution of this problem finds application in many fields, such as analysis of basket data of supermarkets, failures in telecommunication networks, medical test results, lexical features of texts, and so on. The extraction of association rules from very large databases has been solved by researchers in many different ways and the proposed solutions are embedded in, as many powerful algorithms.

An association rule X ⇒ Y is a pair of two sets of items (called itemsets), X and Y, which are often found together in a given collection of data. For instance, the association rule {milk, coffee} ⇒{bread, sugar} extracted from the market basket

https://doi.org/10.30684/etj.31.3B.5
2412-0758/University of Technology-Iraq, Baghdad, Iraq

domain, has the intuitive meaning that a customer purchasing milk and coffee together is likely to also purchase bread and sugar.

The validity of an association rule has been based on two measures: the support, the percentage of transactions of the database containing both X and Y; the confidence, the percentage of the transactions in which Y occurs relative only to those transactions in which also X occurs. For instance, with reference to the above example, a value of 2% of support and a value of 15% of confidence would mean that in 2% of all the transactions, customers buy together milk, coffee, bread and sugar, and that the 15% of the transactions in which customers have bought together milk and coffee contain also bread and sugar. In the original application domain of market basket analysis, support meant that only the association rules with decision making value were extracted.

Association rules are intuitive, and constitute a powerful tool that have been applied recently also to new typologies of problems, such as collaborative recommender systems for e-commerce, intrusion detection and in a distributed information system with the purpose to reduce the amount of objects traversed by the queries.

In order to deal with the association rule problem we have developed a Genetic Algorithm. The major advantage of using GAs in the discovery of prediction rules is that they perform global search and its complexity is less compared to other algorithms as the genetic algorithm is based on the greedy approach. The main aim of this paper is to find all the possible optimized rules from given data set using genetic algorithm. GA has been successfully applied in many search, optimization, and machine `learning problems 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 [3]. The aim is to find all the solutions of best association rules,a new challenge has been added in this work when data sets choose from two different domain and try to find the relationships between these two domain, so the generated rule must has antecedent from the first domain and the consequent from the second domain this need especial method to represent the search space of the problem and some modification to the genetic algorithm operators such as crossover.

Table[1] The association rules from different domain.

The association rules	The antecedent from the first domain (the symptoms)	The consequent from the second domain (the diseases)
s2,s3,s5→ d2	s2,s3,s5	d2
s1,s4→ d3,d4	s1,s4	d3,d4

However, mining algorithms are mostly based on the assumption that users can specify the minimum support appropriate to their databases, and thus referred to as the Apriori-like algorithms have pointed out that setting the minimum support is quite subtle, which can hinder the widespread applications of these algorithms[2]. So a good benefit from the proposed method, it does not require the user-specified threshold of minimum support or minimum confidence, it was calculated by the algorithm directly from the data so that the algorithm is dependent on data not on user, as known its very difficult to predefined the two measures the minimum supported and the minimum confidence.

RELATED WORK

There is some work that used GAs to discover association rule some of them

1.X. Yan et al. (2009) [8] used genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. In this approach, an elaborate encoding method is developed, and the relative confidence is used as the fitness function. two points crossover are used, such that any segment of chromosome may be chosen.

2.Anandhavalli M., Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K.(2009) [1] applied the genetic algorithm over the rules fetched from Apriori association rule mining. The frequent itemsets are generated using the Apriori association rule mining algorithm. The genetic algorithm has been applied on the generated frequent itemsets to generate the rules containing positive attributes.

3.Rupali Haldulakar , Jitendra Agrawal (2011) [5] they design a a method for generation of the rule in which a general Apriori algorithm is used to generate the rules after that they used the optimization techniques. Genetic algorithm to optimize the rules.

In this paper genetic algorithm used to generate the association rules directly without the need to generate the frequent item set , without the need to predefined the support and the confidence of the rule . The consequent of the generated rules may have more than one attribute ,the research suggested new method for crossover, and the method not need encoding schema so it work directly on the data.

GENETIC ALGORITHMS FOR ASSOCIATION RULE MINING

Most of the methods that used GAs in mining association rule are used in discovering frequent item set or in optimizing the discovered association rule by apriori algorithm[1][5] . In this research GAs have been used to discover the association rule directly in one step where most of the previous method generate association rule in the following two steps:

(1) Generating all itemsets for which supports are greater than, or equal to, the user-specified minimum support, that is, generating all large itemsets.

(2) Generating all the rules which satisfy the minimum confidence constraint in a naive way as follows. For each large itemset X, and any $B \subset X$, let $A = X - B$.

If the confidence of a rule $A \rightarrow B$ is greater than, or equal to, the minimum confidence (or $\sup(X) / \sup(A) \geq minconf$), then $A \rightarrow B$ can be extracted as a valid rule.

To examine the Proposed GAs based method to discover the association rules a real world problem has been studied in this research the problem is discover the relationships between some symptoms and some diseases .

In this problem the domain of the antecedent of rules from one domain and the consequent from another domain this lead us to do some modification to GAs in order to applied it to this type of problem.

The developed GAs based method discover association rules in the following steps:

1. initial population

The initial population is generated randomly, each individual in the population which represent a candidate association rule that consist of the two part antecedent and consequent was generated in the following steps :

1. The length of an antecedent of the rule chosen randomly .
2. The items of the first part of the rule is chosen randomly from the symptoms.
3. The length of the consequent of the rule also chosen randomly.
4. The items in the second part of the rule chosen randomly from the diseases .

Table[2] the candidate association rules.

The candidate rule	The antecedent	The antecedent length	The onsequent	The consequent length
s1,s4,s5→ d1	s1,s4,s5	3	d1	1
s1,s3→ d1,d4	s1,s3	2	d1,d4	2

Fitness Assignment

The two measures: the support and the confidence, have been used as a fitness function to the discovered rule in spite of there are another method have been suggested to measure the validity of association rule but this method is the most applied one and used for comparison with other methods used in discover association rule.

Table [3] some candidate association rules with their confidence.

Rule no.	The candidate rule	The confidence
1.	s1→s4 s5 d1	80%
2.	→s1 s3 d1,d4	75%
3.	s1 s3 s4→s5 d2	72%
4.	s1→s2 s6 d3	55%
5.	s1 s3→s4 d3 d6	28%
6.	s5→s7 d1 d4 d5	7%
7.	s1 s4 s6→ d4 d6 d7	0%
8.	s4 s7→s8 d4 d5 d6	0%

As the confidence and support calculated from the data and it doesn't predefined this very useful feature of this method and this give a complete knowledge about all the items that are associated and the confidence of the association, the method discover the rule with high confidence(rule 1 in above Table) or with low confidence(rule 6) .

A novel good feature of the algorithm it discover the items that are not associated where the confidence of these rules is zero as examples the rules 7 and 8 in table 3.

3. Crossover and mutation

To implement the crossover operation to the case study of this research where the items in the rule antecedent is differ from the items in the rule consequent the crossover must modified .

After the crossover operation the problem that well arise is how to determine the antecedent of the rule , the length of it , the consequent of the rule and its length , that determination is important in calculate the confidence and the support of the next generation .

Suppose we have the following two chromosome

s1 s2 s4: →d1 d2

s1 s5: → d5

And suppose is the following two point crossover as following
s2 s3 |s4 : d1| d2
s1 |s5 | :d5
So the offsprings well be as following
s1 s2 s5 :d2
s1 s4 :d1 d5

To calculate the length of the antecedent part of the rule after crossover operation the following formula is suggested .

$$nal1=al-it1+it2 \qquad (1)$$

Where nal1 is the length of the antecedent part of the rule in the new generation ,al is the length of the antecedent part of the rule in the old generation ,it1 number of items cuts from the first chromosome, ,it2 number of items cuts from the second chromosome.

$$nal2=al1-it2+it1 \qquad (2)$$

Where nal2 is the length of the antecedent part of the second rule in the new generation ,al1 is the length of the antecedent part of the second rule in the old generation ,it2 number of items cuts from the second chromosome, ,it1 number of items cuts from the first chromosome.

In order to know the number of items cut from first chromosome the first point of crossover is examined if it in the consequent the it1 is zero, if it's in the antecedent part the second point is examined if it's also in the antecedent then.

$$\text{It2=location of the second point- the location of the first point} \qquad (3)$$

If the second point of crossover in the consequent then

It2=al1-location of the first point of crossover (4)

The same formulas used to calculate the length and items of the antecedent of the second rule and the consequent of the first and second rule. In mutation operation also the mutation point has been examined if it in the antecedent part, the item in the individual is replaced with item from the symptom chosen randomly ,if it in the consequent part, the item in the individual is replaced with item from the dieses and it also chosen randomly.

EXPERIMENTS

Experiments were conducted using a real-world datasets, the dataset of 300 patient that savored from some diseases and have some common symptoms .

The following transaction are examples of the data set.

Table [4] examples of the data set.

The symptoms	the disease the patient saver from it
S1 S2 S3	D1 D3
S1 S2 S3	D1,D2,D3,D4
S1 S2 S3 S4	D1,D4
S1 S2	D1
S3 S5	D4
S1 S2 S6 S7	D1 D2 D3 D6
S1 S2	D2
S4 S6	D1 D2

CONCLUSION AND FUTURE WORK

We have dealt with a challenging of association rule mining problem of finding interesting association rules. The results reported in this paper are very promising since the discovered rules are of a high comprehensibility, high predictive accuracy and of a high interestingness values.

This paper produce a very promising method that can be expanded to solve a lot of problem such as classification problems and can used to find classification rule for different classes .

For future work first the method suggested in this research will be applied to study the relationship of the absences of students in their lessons and the results of them in these loosens this indicate the effect of the teachers in the results of their students. Second the method suggested in this work will be compared with other methods used in generate association rules.

REFERENCES

[1].Anandhavalli M., Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K.(2009). Optimized association rule mining using genetic algorithm .Advances in Information Mining, ISSN: 0975–3265, Volume 1, Issue 2, 2009, pp-01-04

[2].Han, J., Wang, J., Lu, Y., & Tzvetkov, P. (2002). Mining top-k frequent closed patterns without minimum support. In Proceedings of the 2002 IEEE international conference on data mining (ICDM 2002) (pp. 211– 218).

[3].Jos´e Mar´ıa Luna, Jos´e Ra´ul Romero, Sebasti´an Ventura (2011). Mining and Representing Rare Association Rules through the Use of Genetic Programming 2011 Third World Congress on Nature and Biologically Inspired Computing.

[4].Peter P. Wakabi-Waiswa and Venansius Baryamureeba. (2008)Extraction of Interesting Association Rules Using Genetic Algorithms. International Journal of Computing and ICT Research, Vol. 2, No. 1, pp. 26 – 33. http:www.ijcir.org/volume2-number1/article4.pdf.

[5].Rupali Haldulakar , Jitendra Agrawal (2011) Optimization of Association Rule Mining through Genetic Algorithm , International Journal on Computer Science and Engineering (IJCSE) Vol. 3 No. 3 Mar 2011.

[6].Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar and Partha Pratim Sarkar.(2010) Mining Frequent Itemsets Using Genetic Algorithm. *International* Journal of Artificial Intelligence & Applications (IJAIA), Vol. 1(4):133–143, October 2010.

[7].Wilson Soto_ and Amparo Olaya–Benavides (2011) A Genetic Algorithm for Discovery of Association Rules

[8]. Yan et al. X. (2009). Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support Expert Systems with Applications 36 (2009) 3066–3076