Eng. & Tech. Journal, Vol. 28, No. 2, 2010

Proposed Enhancement algorithm for Company Employers Management using Genetic Algorithm in Data Mining

Dalia Nabeel Kamal*

Received on: 17/2/2009
Accepted on: 6/8/2009

Abstract

Data mining is a process of automatically discovering useful information in large data repositories that uses a variety of data analysis tools to discover patterns and relationships that can be hidden among vast amount of data. From these patterns and relationships, businesses and organizations can make valid predictions about future trends in all areas of business. Association rule mining is a typical approach used in data mining domain for uncovering interesting trends, patterns and rules in large datasets.

This research concentrates on one particular aspect to improve the efficiency of the association rules technique in data mining and implement the proposed algorithm on employers management system. The resulted association which introduced by applying rule technique, will be treated by genetic algorithm to find a new rules that might be more efficient and powerful for proposed data base by propose cross point ,threshold for fitness to deal consistently with the formula of the association rules, and gives good results.

Keywords: association rules, frequent item sets, closed frequent item sets and maximal frequent item sets.

مقترح تحسين خوارزمية ادارة موظفي الشركات باستخدام الخوارزميات الجينية في تنقيب البيانات

الخلاصة

تنقيب البيانات هي عملية استخدام ادوات التحليل لمجموعة من البيانات المختلفة من اجل الحصول على انماط وعلاقات غير ظاهرة بين كمية هائلة من البيانات . ومن خلال هذه الانماط والعلاقات تستطيع شركات الاعمال والمؤسسات من بناء توقعات جديدة في جميع مجالات العمل المختلفة. قاعدة الارتباط او التداخل في تنقيب البيانات تعتبر الطريقة الملائمة لتنقيب البيانات في مجال استخدام طرق واساليب جديدة,وانماط وفواعد بين كمية هائلة جدامن البيانات.
هذا البحث يطرح فكرة جديدة تهدف الى زيادة كفاءة قاعدة الارتباط او التداخل في تنقية تنقيب البيانات وذلك من خلال استخدام الخوارزميات الجينية وامكانياتها في الحصول على مقترح يعمل على زيادة وقوة قاعدة البيانات.
تم تطبيق قاعدة الارتباط او التداخل على كمية هائلة من البياناتوتنفيذه على نظام ادارة الموظفين في اي شركة او مؤسسة واعتماد النتائج التي تم الحصول عليها كمدخلات الى الخوارزمية الجينية للوصول الى مقترح يعمل على زيادة كفاءة وقوة قاعدة البيانات وذلك بسبب المميزات التي تمتلكها الخوارزمية الجينية من حيث عملية التقاطع بين النتائج وتحديد مقياس

* Computer Science Department, University of Technology / Baghdad

https://doi.org/10.30684/etj.28.2.4
2412-0758/University of Technology-Iraq, Baghdad, Iraq

كفاءة كحد حرج للخوارزمية ومطابقة القواعد الناتجة من الخوارزمية الجينية مع الشروط لبناء القواعد المترابطة والتي كانت مشجعة.

1. Introduction

Data mining is an information analysis tool that involve the automated discovery in order to find novel and useful patterns that might otherwise remain unknown they also provide capabilities to predicate the outcome of a future observation. Data mining cannot be considered all information discovery tasks [1].The enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that help in decision-making. *Data Mining*, also known as *Knowledge Discovery in Databases* (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. Data mining is not specific to one type of media or data; data mining should be applicable to any kind of information repository.

However, algorithms and approaches may differ when applied to different types of data. Indeed, the challenges presented by different types of data vary significantly. The kinds of patterns that can be discovered depend upon the data mining tasks employed.

1-1Data mining types

There are two types of data mining tasks: *descriptive data mining* tasks that describe the general properties of the existing data, and *predictive data mining* tasks that attempt to do predictions based on inference on available data. Data characterization is a summarization of general features of objects in a target class, and produces what is called *characteristic rules* [2].

Data mining (DM) is the nontrivial extraction of implicit, previously unknown, interesting, and potentially useful information (usually in the form of knowledge patterns or models) from data. The most reticulated DM problems are reduced to traditional statistical and machine learning methods: classification, prediction, association rule extraction, and sequence detection [3].

- **Data discrimination**
- produces what are called discriminate rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class.

Associationanalysis

- Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules.
- **Classification analysis** is the organization of data in given classes. Also known as *supervised classification*, the classification uses

given class labels to order the objects in the data collection. Classification approaches normally use a *training set* where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. Clustering: is the organization of data in class Similar to classification, but in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes [4, 5].

1-2. Association rules

Association rule mining is an important problem in the rapidly growing field called data mining and knowledge discovery in databases (KDD). There are many application areas for association rule mining techniques, which include catalog design, store layout, customer segmentation, and telecommunication alarm diagnosis and so on. The task of mining all frequent associations in very large datasets is quite challenging. The search space is exponential in the number of attributes and with millions of records of dataset. However, most current approaches are iterative in nature, requiring multiple database scans, which is clearly an expensive solution [6]. The efficient discovery of such rules has been a major focus in the data mining research community. Many algorithms and approaches have been proposed to deal with the discovery of different types of association rules discovered from a variety of databases. However, typically, the databases relied upon are alphanumerical and often transaction-based. The problem of discovering association rules is to find relationships between the existence of an object (or characteristic) and the existence of other objects (or characteristics) in a large repetitive collection.Such a repetitive collection can be a set of transactions for example, also known as the market basket. Typically, association rules are found from sets of transactions, each transaction being a different assortment of items, like in a shopping store ({milk, bread, etc}). Association rules would give the probability that some items appear with others based on the processed transactions, for example milk →bread [50%], meaning that there is a probability 0.5 that bread is bought when milk is bought[2,7].

1-3 Representation For problem in Data mining

Essentially, the problem consists of finding items that frequently appear together, known as frequent or large item-sets.

The problem is stated in the following example, Let $I = \{i1, i2 \ldots im\}$ be a set of literals, called items.

Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$.

- A unique identifier *TID used to refer to (the Colum that represent one employment for each raw)* is given to each transaction.
- A transaction T is said to contain X, a set of items in I, if $X \acute{I} T$.
- An *association rule* is an implication of the form "$X \text{ Þ } Y$", where $X \acute{I} I$, $Y \acute{I} I$, and $X \text{ Ç } Y = \varnothing$. The rule $X \text{ Þ } Y$ has a *support s* in the transaction set D is s% of the transactions in D contain $X \text{ È } Y$. In other words, the support of the rule is the probability that X and Y hold together among all the possible presented cases.

- It is said that the rule $X \Rightarrow Y$ holds in the transaction set D with *confidence* c if c% of transactions in D that contain X also contain Y.
- In other words, the confidence of the rule is the conditional probability that the consequent Y is true under the condition of the antecedent X.
- The problem of discovering all association rules from a set of transactions D consists of generating the rules that have a *support* and *confidence* greater that given thresholds. These rules are called *strong rules [7]*.

2. Genetic algorithms [8, 9]

The genetic algorithm is a search algorithm based on the mechanics of natural selection and natural genetics.

This means that survival and reproduction of an individual is promoted by the elimination of useless or harmful traits and by rewarding useful. *Genetic Algorithm* begins with a randomly selected population chromosomes represented by strings The GA uses the current population of strings to create a new population. Such that the strings in the new generation are on better than those in current population (the selection depends on their fitness value), as shown in figure (1). The selection process determines which string in the current will be used to create the next generation. The crossover process determines the actual form of the string in the next generation. Here two of the selected parents are paired. A fixed small mutation probability is set at the start of the algorithm.

2-1 Basic element of GA

Most GAs methods are based on the following elements, of chromosomes, selection according to fitness, produce new offspring, and random mutation of new offspring.

- Chromosomes: The chromosomes in GAs represent the space of candidate solutions. Possible chromosomes encodings are binary, permutation, value, and tree encodings.
- Fitness function: GAs requires a fitness function which allocates a score to each chromosome in the current population Thus; it can calculate how well the solutions are coded and how well they solve the problem.
- Selection: The selection process is based on fitness. Chromosomes that are evaluated with higher values (fitter) will most likely be selected to reproduce, whereas, those with low values will be discarded. The fittest chromosomes may be selected several times, however, the number of chromosomes selected to reproduce is equal to the population size, keeping the size constant for every generations. This phase has an element of randomness just like the survival of organisms in nature. The most used selection methods are roulette-wheel, rank selection, steady-state selection, and some others. Moreover, to increase the performance of GAs, the selection methods are enhanced by elitism. Elitism is a method which first copies a few of the top scored chromosomes to the new population and then continues generating the rest of the population. Thus, it prevents loosing the few best found solutions.

- Crossover: Crossover is the process of combining the bits of one chromosome with those of another. This is to create an offspring for the next generation that inherits traits of both parents. Crossover randomly chooses a locus and exchanges the subsequences before and after that locus between two chromosomes to create two offspring. For example, consider the following parents and a crossover point at position 3:

Parent 1 100 | 0111

Parent 2 111| 1000

Offspring1 1000 1000

Offspring 2 111 0 111

Mutation: mutation is performed after crossover to prevent falling all solutions in the population into a local optimum of solved problem.Mutation changes the new offspring by flipping bits from 1 to 0 or from 0 to 1. Mutation can occur at each bit position in the string with some probability, usually very small (e.g. 0.001). For example, consider the following chromosome with mutation point at position 2:

Not mutated chromosome: 1000111

Mutated: 1100111

The 0 at position 2 flips to 1 after mutation.

3. The Proposed Genetic Algorithm for using GA in Data mining

Step 1. Initialization

The GA randomly draws values to begin the search. , which only draws weights for one solution, the GA will draw weights for a population of solutions. The population size for this study is set to 20 solutions. Once the population of solutions is drawn, the training begins with this first generation.

Step 2. Evaluation

Each of the 20 randomly drawn solutions in the population is evaluated based on a pre-selected objective function, which is not necessarily differentiable. Since our objective is to find a global solution that can identify relevant variables, an objective function is needed that will not only measure the error between estimates and real outputs but will also include additional penalties for the number of non-zero connections in the model.

Step 3. Reproduction

In the first generation are more likely to be drawn for the second generation of 20 solutions. This is known as reproduction, which parallels the process of natural selection or 'survival of the fittest'. The solutions those are most favorable in optimizing the objective function will reproduce and thrive in future generations, while poorer solutions die out.

Step 4. Crossover.

These 20 solutions, which only include solutions that existed in the prior generation, are now randomly paired into 10 sets of solutions. For each paired set of solutions, a random integer value in the range [1, *n*], with *n* being equal to the number of weights in a solution, is drawn to decide where the crossover operation is to take place. Once *n* is determined, all the weights above that value are switched between the pair resulting in two new solutions. For example, if a solution contains 10 weights, a random integer value is drawn from 1 to 10 for the first pair

of solutions. For this example, assumes the value selected is six. Every weight above the sixth weight is now switched between the paired solutions, resulting in two new solutions for each pair. Once this is done for each pair of solutions, crossover is complete.

Step 5. Mutation.

To sample the entire parameter space, and not be limited only to those initially random drawn values from the first generation, mutation must occur. Each solution in this new generation now has a small probability that any of its weights may be replaced with a value uniformly selected from the parameter space.

4. The implementation of the proposed algorithm

Step one:

The first step in the proposed system is to build a virtual data base for any company or firm employer as in the following Table (1) that used as data to implement the proposed algorithm:

The conversion of the above database to be suitable for the mining process is done by giving each attribute an alphabet will appear if it give the real value and will not appear if it give not the value such that example are shown in table (2):

1. If the (Age) is more than (30) then (A) will appear, if the age is equal or less than (30) then (A) will not appear.
2. If the (Sex) is (Male) then (B) will appear, if the (Sex) is (Female) then (B) will not appear.
3. If the (nationality) is (Iraqi) then (C) will appear, if the (nationality) is (not Iraqi) then (C) will not appear.
4. This is done for each attribute (column) related to each employ (row).
5. Then each employ will take a TID number pointed for the converted set of alphabet that represent it.

Step two:

Applying the association rule technique on the converted database that for extracting the association rules that will be explained in the following implementation, figure (2) (Note: before applying the technique converted the last database to notepad file), because notepad file will accept huge data.

Association rule reach to the following results as shown in table (3):

Step three:

The genetic algorithm used the resulted of that obtained from (step 2) association rules see the following stepping:

1. Initial populations consist of all the resulted association rules that obtained in (step 1).
2. End the loop of genetic algorithm for (X) generation where (X) is half number of association rules.
3. Take each two association rules randomly. (Such A $\longrightarrow$ BC and B $\longrightarrow$ A).
4. Crossover these two association rules in the middle of them ($\longleftrightarrow$).
5. The resulted crossover association rules will submit to the fitness measure which will be two thresholds.

Note: (First: no letter appears in both sides second the confidence of the association rules pass the minimum confidence as determined above) such in table (4):

6. From my point view there is no need for mutation.End.

6. Conclusion

From this paper concealed to the following results:

1. Using Genetic algorithm on association rules in data mining create un expected association rules these new association rules give a new knowledge and pattern support the official work.

2. Using the association rules basis as fitness for Genetic algorithm make the obtained result much more powerful and confidential.
3. Applying the proposed work on official work give a new results un expected, since these results differ from applying traditional association rules technique.

7. Future works

The work done in this paper can be extended to several interesting directions, among these are:

1- Use neural network (NN) with the proposed method in this paper in order reach to learning NN for reach to the optimal association rules for data mining.
2- Build secure system based on capabilities for genetic algorithm with data mining.
3-Modofy on the proposed algorithm that produced in this paper by insert security requirements by add protection-condition on fitness function that make the work more robust in front of the intruder.

8. References

[1] Michel Steinbach, Pang-Nang Tan,Vipin Kumar," **Introduction** to Data Mining ",Addison Weasly,2006.

[2] M. S. Chen, J. Han, and P. S. Yu. **"Data mining: An overview from a database perspective"**. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.

[3] R˘azvan Andonie and Boris Kovalerchuk," Neural Networks for Data Mining: Constrains and Open Problems", Computer Science Department,Central Washington University, Ellensburg, USA,2007

[4] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. **"Advances in Knowledge Discovery and Data Mining"**. AAAI/MIT Press, 1996.

[5] J. Han and M. Kamber. **"Data Mining: Concepts and Techniques"**. Morgan Kaufmann, 2000.

[6] S. Dehuri, A.K. Jagadev, A. Ghosh, R. Mall," Multi-objective genetic algorithm for association rule mining using a homogeneous dedicated cluster of workstations",American Journal of Applied Sciences, Nov, 2006

[7] Dunham Margraet H., **"Data mining: Introductionary and Advanced Topics**", Southern Methodist University, 2003.

[8] Bethany Delman, "Genetic Algorithms in Cryptography" ,A Ms.C Thesis, Kate Gleason College of Engineering,2004.

[9] Garg P.,Shastri A.," An improved cryptanalytic attack on Knapsack cipher using genetic algorithm", International journal of Information Technology ,November 2006.

[10] Randall S. Sexton* and Naheel A. Sikander." **Data Mining Using a Genetic Algorithm-Trained Neural Network",** Computer Information Systems, Southwest Missouri State University, USA.2001.

Table (1) the data base that used in proposed algorithm

ID	Age	Sex	Nationality	Religion	Date of birth	address	Graduation	Marriage state	No. of child	Property
1	24	M	Iraqi	Muslim	1984	Baghdad	High school	single	0	No
2	30	M	Iraqi	Not Muslim	1978	najf	B. Sc	marred	2	Yes
3	40	F	Not Iraqi	Muslim	1968	England	PhD	single	0	Yes
4	44	F	Iraqi	Not Muslim	1964	duhok	PhD	marred	1	Yes
5	50	M	Iraqi	Muslim	1958	anbar	High school	marred	1	No
6	22	F	Iraqi	Muslim	1986	tekret	B. Sc	single	0	No
7	28	F	Iraqi	Not Muslim	1980	mousl	M.Sc	single	0	No
8	33	F	Iraqi	Muslim	1975	dyala	B. Sc	single	0	No
9	60	F	Not iraqi	Muslim	1948	Basra	M.Sc	marred	3	Yes
10	55	M	Iraqi	Muslim	1943	Baghdad	High school	marred	4	No

Table (2) the data base after conversion

TID	item sets
1	BCDE
2	BCE
3	ACDE
4	A CD

Table (3) Association rules

New Association rules	Fitness
A ⟶ BC	(100%)
B ⟶ A	(100%)
A ⟶ **A**	**Omitted**
B ⟶ **B**C	**Omitted**
B ⟶ C	(70%) **Omitted**

Table (5) new Association rules

Association rules	Fitness
A ⟶ B	**(100%)**
A ⟶ C	**(100%)**
A ⟶ D	**(100%)**
AB ⟶ C	**(100%)**

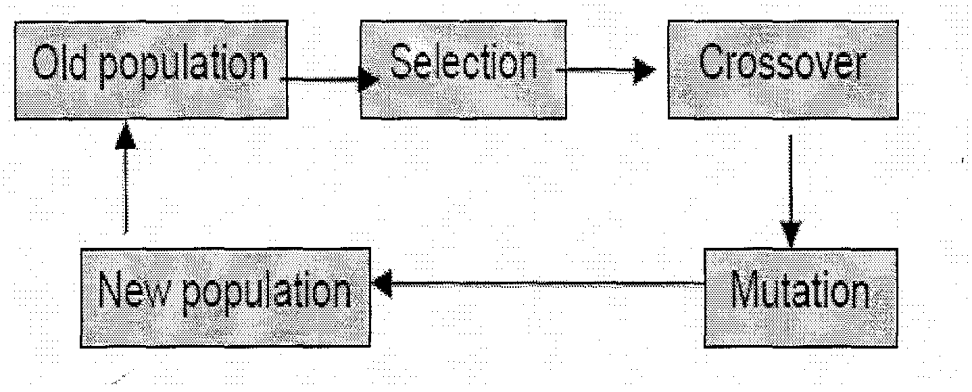

Figure (1) The basic genetic algorithm concept

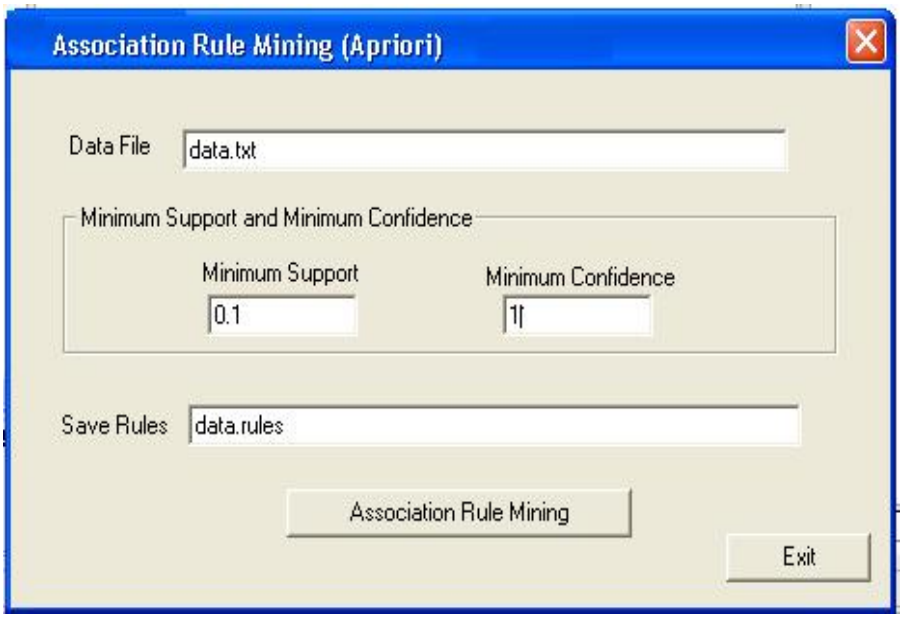

Figure (2) implementation of Data mining