Application of data mining algorithm with genetic algorithm

Maytham M. Hammood

College of Computer Science and Mathematics , University of Tikrit

(Recieved 11 / 5 / 2008, Accepted 5 / 6 / 2008)

Abstract

The *Attribute-Oriented Induction* (AOI) technique is one of the data mining techniques used to analyzing database. It is designed to address unique AOI problems that cannot be solved by other data analysis tools.

An AOI is used to analysis the data to generalize the attributes according to two thresholds which determine the depth of rolling-up or rolling-down. The end result is some of attributes that succeeds in generalizing by passing the thresholds test, each one of them raises one value that represents the phenomenon in that database which is mined. But AOI ask the user to determine analysis dimensions, it is non trivial for users to determine which dimensions should be included in the analysis of class characteristics. Data relations often contain 50 to 100 attributes, and a user may have little knowledge regarding which attributes should be selected, so these problems have huge search space compatible with genetic algorithm search to determine the relevant attributes, since genetic algorithm need to fitness function so we use association rule for that.

1. Introduction

A huge amount of information needed to turning to get useful information and knowledge which it is gained by using in applications like business management, production control, and market analysis. So data mining is used to extract this knowledge from database which it is represent a large search space and AOI which it is one of data mining techniques as a search strategy mixed with another successful search strategy we me meaning by that genetic algorithm to get efficient search strategy which is has been applied successfully in a wide range of applications ant get a perfect results.

2. GA AND DATA MINING PROCESSES

In computer science database precedes data mining for many years and that leads to unsuitability between the database structure and data mining, also the data mining works heuristically, database may store for many years and that makes the database very huge which makes the performance of computations impossible, these computations are very important for data mining such as aggregate functions (Count ( ), Sum ( ), average ( ), and max ( )). AOI is one of the data mining techniques and depends on sum ( ) function [2] we accelerate the computation by using genetic algorithm in our research.

2.1 Genetic Algorithms Process

GAs start by creating an initial population of genotypes, which are the initial search points in the solution space. The generation of an initial population is usually carried out by assigning randomly selected alleles to all genes in a chromosome. The generation of a population of genotypes continues until there exists the desired number of genotype in the population. The initial population is usually generated randomly[7].

A central instrument in a genetic algorithm is the fitness function. Since genetic algorithm is aimed for optimization of such a function, this function is

One of the keys for success. Consequently, a fitness function should represent all issues that play a role in the optimization of a specific problem. In our project the support and confidence factors of associated rules are used as a fitness function [8].

2.1.1 Crossover

It is the process where information from two parents is combined to form children. The crossover takes two chromosomes and swaps all genes residing after a randomly selected crossover point to produce new chromosomes [7]. This operator does not add new genetic information to the population of chromosomes but manipulates the genetic information already present in the mating pool. By combining the genetic materials in the parents, the hope is to obtain new more fit children. It works as follows:

1- Select two parents from the mating pool (the largest two chromosomes)

2- Find a position k between two genes randomly according to uniform distribution in the range (1, m-1), where m is the length of the chromosome.

3- Swap the genes after k between the two parents.

Crossover can produce children that differ substantially from their parents and thus, introduce important new search points in the solution space which can be explored in order to find better solutions.

2.1.2 Mutation

The basic idea of mutation is to add new genetic information to chromosomes. It is especially important when the chromosomes in the population, after a number of generations are very similar and the GA may be get stuck in a local maximum. A way to introduce new information is by changing the allele of some genes, which is exactly what the mutation operator does [7].

2.1.3 Reproduction

After manipulating the genetic information already present in the mating pool by fitness function, the reproduction operator adds new genetic information to the population of chromosomes by combining strong parents with strong children, the hope is to obtain new more fit children. The reproduction procedure must imitate the natural selection, where the fittest survive. The reproduction procedure must select genotypes according to their fitness. Genotype with high fitness must be selected with a higher probability than those with lower fitness.

2.2 Data Mining Techniques:

This is a part from selecting the appropriate combination of data mining algorithms, it is quick and automated. During the data mining steps will vary with the kind of application that is under development. For example, in the case of database segmentation, one or two runs of the algorithm may be sufficient to clear this step and move into analysis of the results [4].

In the following subsections we illustrate the two important techniques of data mining which they used in our research:

2.2.1 Association Rule

The typical example of association rules is the basket data analysis. In a given database D, all the records consist of two attributes: transaction ID (TID) and the item which customer bought it in the transaction.

TID	ITEM
100	1
100	3
100	4
200	2
200	3
200	4
300	1
300	2
300	3
300	5
400	2
400	5

Table 2.1 Example of the Basket Data set

Usually the item attribute in each record contains only one item, so in the database, there will be more than one row for a transaction ID since each transaction will involve more than one item. Table (2.1) shows one example of the basket data set with four transactions. The formal definition of association rules is the following:

Let $I = \{I_1, \ldots I_n\}$ 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 $T \subseteq I$, and each transaction is associated with a unique identifier called TID.

An item set X is a set of items in I. An item set X is called a K – item set if it contains K items from I.

A transaction T satisfies an item set X if $X \subseteq T$.

An association rule is an implication of the form X Y.

where $X \subseteq I$,

$Y \subseteq I$ and $X \cap Y = \varnothing$.

The support of an item set X in D is :

$$\text{Support} = \frac{\text{The number of Transactions that contains X\&Y}}{\text{The total Number of Transactions}}$$

The confidence of an item set Y associated to X.

$$\text{Confidence} = \frac{\text{The number of Transactions that contains X\&Y}}{\text{The number of Transactions that contains}}$$

These two factors are used in this project as fitness function for GA process to find the optimal solution that is represented by the attributes associated to user query attributes [5].

2.2.2 Attribute – Oriented Induction (AOI)

AOI is a set – oriented database mining method which generalizes the task – relevant subset of data attribute – by – attribute, compresses it into a generalized relation, and extracts from it the general features of data[6].

2.2.2.1 Data Generalization

Given the large amount of data stored in database, it is useful to be able to describe concepts in concise and succinct terms at generalized levels of abstraction. Allowing data sets to be generalized at multiple levels of abstraction facilitates examination of the general behavior of the data. The essential operation of attribute–oriented induction (AOI) data generalization is 1-attribute removal and 2-attribute generalization.

2.2.2.2 Attribute Removal

An attribute–value pair represents a conjunct in generalized and thus generalizes the rule. If, as in case 1, there is a large set of distinct values for an attribute but there is no generalization operator for it, the attribute should be removed because it cannot be generalized, and preserving it would imply keeping a large number of disjuncts which contradicts the goal of generating rules. On the other hand consider case 2 where the higher level concepts of the attribute are expressed in terms of other attributes. For example of case 1 the Name and Phone attributes have large number of distinct values but there is no hierarchy concept, and of case 2 the street attribute, whose higher level concepts are represented by the city attribute.

2.2.2.3 Attribute Generalization

If there is a large set of distinct values for an attribute in the initial working relation, and there exists a set of generalization operators on the attribute, then a generalization operator should be selected and applied to the attribute.

Use of generalization operator to generalize an attribute value within a tuple or rule, in the working relation will make the rule cover more of the original data tuples thus generalization the concept it represents. This corresponds to the generalization rule known as climbing generalization trees. For example the GPA (grade – point – average) attribute will be generalized from (50, 65, 70, etc.) to (accept, middle, good, very good, and excellent).

2.2.2.4 Thresholds

It is concept to solve the problem: how many distinct values for an attribute are considered to be generalized attribute? To control of attribute generalization we use two thresholds, a user will determine the values of them but he must be careful to choose the values of thresholds.

A- Threshold1

If the number of distinct values in an attribute is greater than the attribute threshold, further attribute removal or attribute generalization should be performed. Data mining systems typically have a default attribute

threshold value (typically ranging from 2 to 8). And should allow experts and users to modify the threshold values as well. If a user feels that the generalization reaches too high a level for a particular attribute, he can increase the threshold. This corresponds to drilling down along the attribute. Also, to further generalize a relation. He can reduce the threshold of a particular attribute, which corresponds to rolling up along the attribute [2].

B- Threshold2

If the number of (distinct) tuples in the generalization are greater than the threshold, further generalization should be performed (that is applied in our project).Otherwise, no further generalization should be performed, such a threshold may also be preset in the data mining system (usually within a range of 10 to 30), or set by an expert or user, and should be adjustable. For example, if a user feels that generalized relation is too small, he can reduce the threshold, which implies rolling up[2].

These two techniques can be applied in sequence: first apply the threshold1 to generalize each attribute and then apply threshold2 to further reduce to the size of the generalized relation.

2.2.2.5 Attribute-Oriented Induction Algorithm

Step1. The initial working is derived to consist of the data relevant to the mining task defined in the user provided data mining request.

Step2. The working relation is scanned once to collect statistics on the number of distinct values per attribute.

Step3. Refers to attribute removal as described earlier.

Step4. Perform attribute generalization, this can be implemented by replacing each value of attribute by its ancestor in the concept hierarchy.

Step5. Identical tuples in the working relation are merged in order to create the

3. DESIGN PROPOSED RESEARCH

The project is used to control and view the data mining operations. One of the most important data mining techniques is AOI. Attribute oriented induction is a set – oriented database mining method which generalizes the task-relevant sub set of data attribute-by-attribute, compresses it into generalized relation, and extracts from it the general feature of data. AOI method has been developed as an interesting technique for mining knowledge from data. Database operation, extracts generalized rules from an interesting set of data and discovers high-level data regularities [6].

The problem of attribute selection in data mining is an important real-world problem that involves multiple objectives to be simultaneously optimized, in order to solve this problem; this work proposes a multi objective GA for attribute selection based on the association rule factors (confidence and support factors). In our research we discovered a mutual search strategies, the major one is AOI technique which request to input for its program the attributes of data base that we want to extract the knowledge from it, but since the most applications have huge attributes that lead exponential running and that mean allot of time to solve the exponential problem we asked the users to select the important attributes to apply the AOI technique on them and lead us to another problem that represent human mistakes, so our research solve it by using GA which it is in turned need to fitness function to able it to run suitable with specific problem, in our research represent the relevant attributes, therefore we use another data mining technique to aid GA in its work and this technique is association rule.

3.1Database

he research database is about the students of computer science of University of Tikrit taken randomly (75 students**).** In our project we applied many steps to database to pave for our project and we used eighteenth attributes, these attributes are list below:

1- stname 2- gender
3- branch 4- level
5- levelyear 6- casestudy
7- graduatyear 8- acceptyear
9- studytyp 10- grid
11- street 12- city
13- country 14- cycle
15- sumertrain 16- acceptmonth
17- acceptday 18- phone

And the following are explained some applications according to the figure 3.1

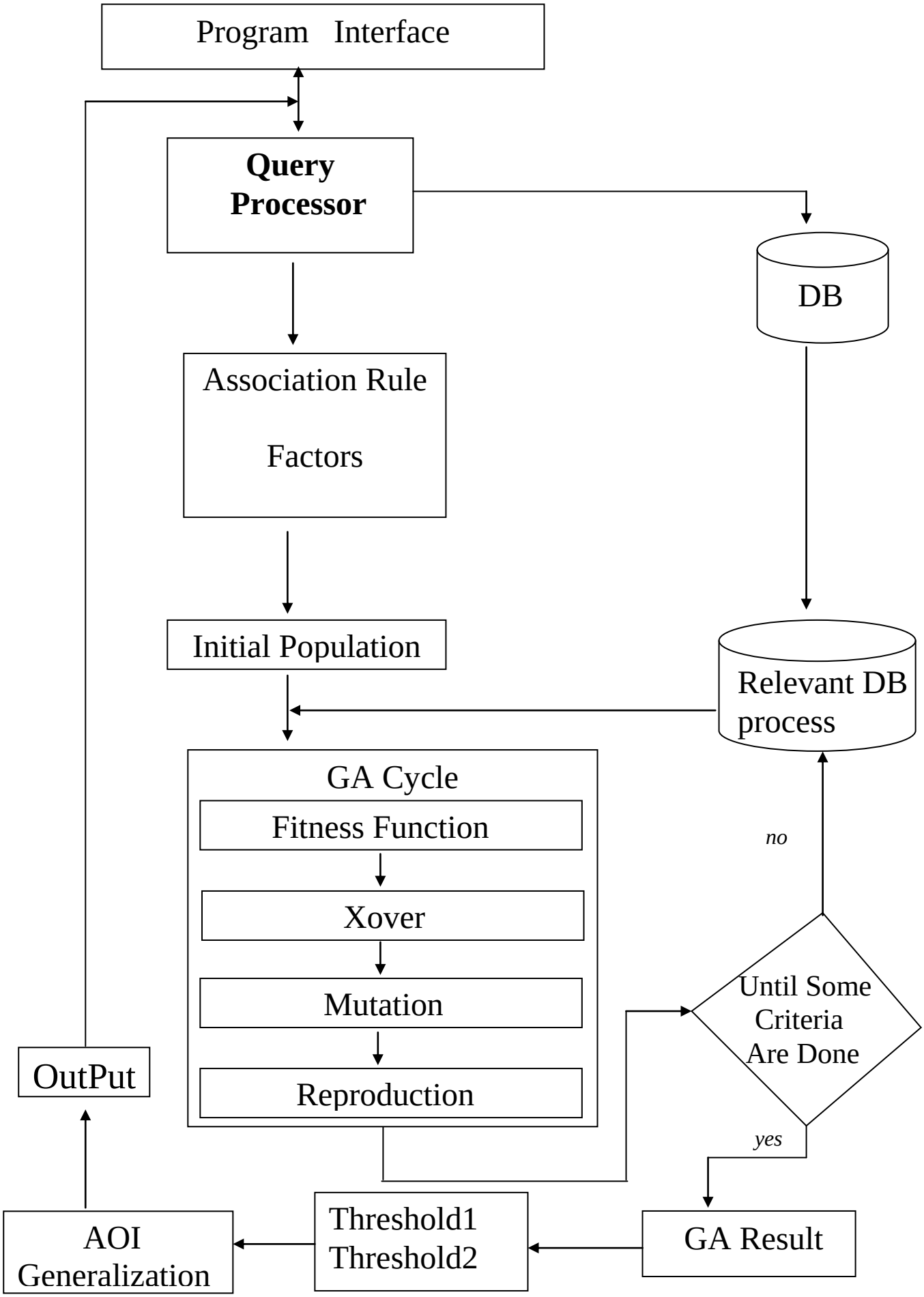

Figure 3.1 System Diagram

Relevant Database Process

It is the pure database, which represents this part of clean database. It just has the records that contain the values of attributes that are determined by filtering process. The task relevant database produced by the filtering process contains a portion of database which is relevant to the values of specific attributes determined by filtering process. Task relevant database is very important in data mining technique, so in our research we reduce executing time of research algorithm by scanning just relevant database part for data mining query instead of scanning the entire database.

3.3 Query Process

As we see in the figure 3.2, this interface contains:
1-information: Instructions text.
2-exit: Exit button.
3-text: The text where is the user writing his data mining query.
4-data mining attributes: A list of attributes names and the possible values of each attribute.
5-go: Go button is triggered of NL compiler. It must be used after the user writes his query.

3.4 Association Rule Factors

As we see in the figure 3.2 which appears after interface processing finishes its execution there are two texts which allow the user to enter confidence and support factors and the interface system tells the user the range of these two factors (0.125 to 1) and also tells him if he ignores them, the system will take (0.5, o.5) as a default numbers for confidence and support factors respectively for using them as a base of association rules which are in turn used as a fitness function of GA process which is triggered by GA button. You see this in the figure 3.2

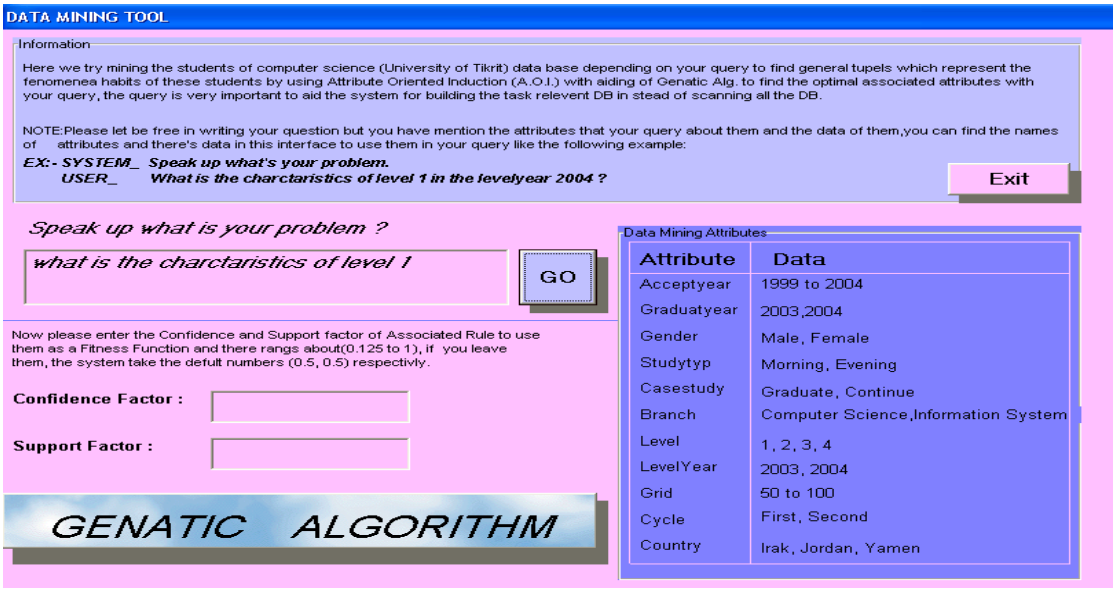

Figure 3.2 Association and GA interface

3.5 Initial Population

It is random population constructed by ten chromosomes, each one has ten alleles which in turn are represented by integer numbers, these numbers represent the serial number of each attribute in the meta data, the last number of each chromosome is zero, this number represents the initial number of fitness function. This number increases after each GA cycle and is referred to as the number of alleles in that chromosome associated to user data mining query. By used the proposal system for testing the database we found the largest set of associated attributes for DM user query is nine items so the length of chromosome is ten alleles

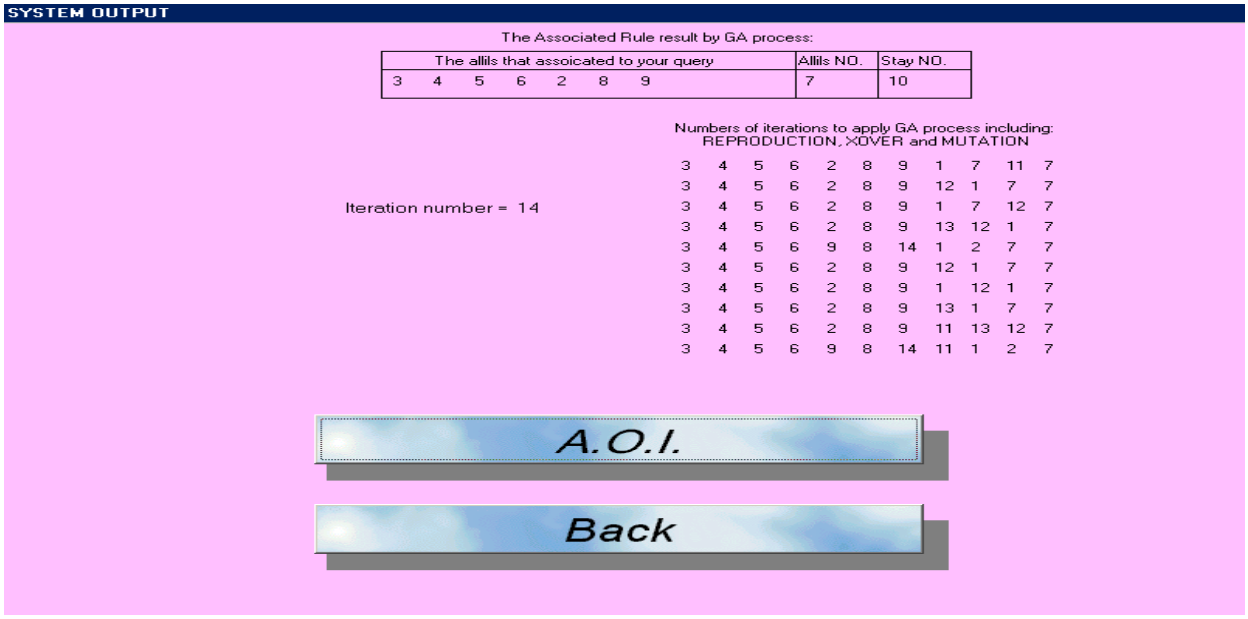

Figure 3.3 GA operator

3.6 Genetic Algorithm Cycle

It contains the processes of GA which are iterated fifteenth times or reach chromosome which has ten associated alleles to task relevant database, as we see in figure 3.3, one of the iterations is submitted to GA cycle.

3.6.1 Fitness Function

It is depends on filtering and recursive techniques by taking the first allele in the chromosome and bringing the attribute which that allele represents and checks every possible value for this attribute by filtering the task relevant instant to that value to obtain new task relevant, which is tested by confidence and support factors.

3.6.2 Reproduction

The reproduction must select genotypes according to their fitness. Genotypes with high fitness must be selected with a higher probability than those with lower fitness. We apply reproduction in our research by constructing temporary population pool for the new population (children) which results from fitness function then compare it with the old population (parents) and put the strongest chromosomes in the original pool to produce new population which contains ten chromosomes representing the strongest parents and children.

3.6.3 Crossover

Apply mating processes between each two chromosomes, assign the cut point for crossover process of each chromosome depending on the number of fitness function to specify the two position of cutting and replace the left part with the right part of the other one and return after the end of the right part to the original alleles' chromosome as we see in the figure 3.4

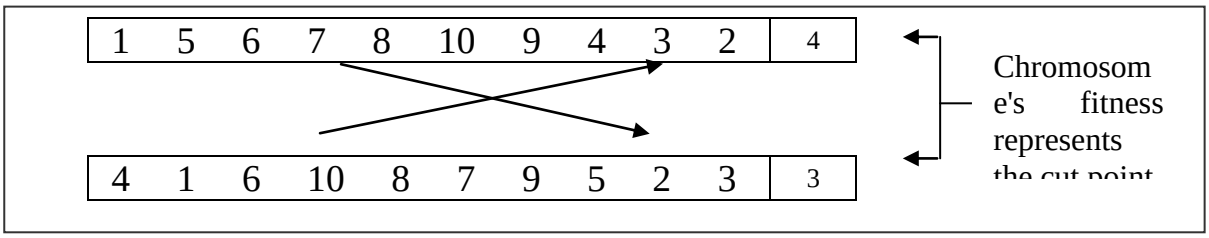

Figure 3.4 Crossover Operation

Then check for redundant alleles in the children chromosomes and then replace them with new alleles generated by mutation process. If the number of alleles on the right side of cut point is smaller than the number of alleles which would replace it (the alleles of the left side of cut point for the other parents) then we replace just the alleles of small number and stop replacing process, and vice versa if the left side alleles are smaller than right side ones.

3.6.4 Mutation

Choose randomly integer number of attributes (alleles) and replace it with redundant alleles to avoid the local maximum. This step is embedded into crossover step. In our research we use the number (15 iterations) of GA iterations as criteria for stopping the GA cycle, additional if the association attributes number are equal to ten also GA cycle must be stopped. The last field in the GA result is very useful to determine the number of iteration, in the proposal DB.

It represents the output of GA and at the same time the input of AOI with the threshold1 and threshold2.

3.7 Thresholds

There is AOI requirements text which contains the instructions for using threshold1, which represents the number of distinct data values for generalization process and threshold2 represents the number of records, these are the generalized tuple cover, and considered success. The default ranges of the two thresholds are (3 to 8) and (4 to 30) respectively (by try and error.

3.8 Attribute-Oriented Induction by Using Genetic Algorithm

AOI program works like fitness function program notation by depending on filtering and recursive techniques and that is done by taking the first allele of the GA result chromosome and bring the attribute which that allele is represented and checks the distinct values of this attribute with threshold1, if the number of values is equal to or greater than threshold1 then apply removal attributes process, else check every possible value for this attribute and this is done by filtering the task relevant instant to that value to obtain new task relevant which is tested by threshold2, and this is done by comparing the records number of new task relevant with threshold2, if equal to or greater then it keep this tuple for listing it, then repeat the process for another attribute and so on

.

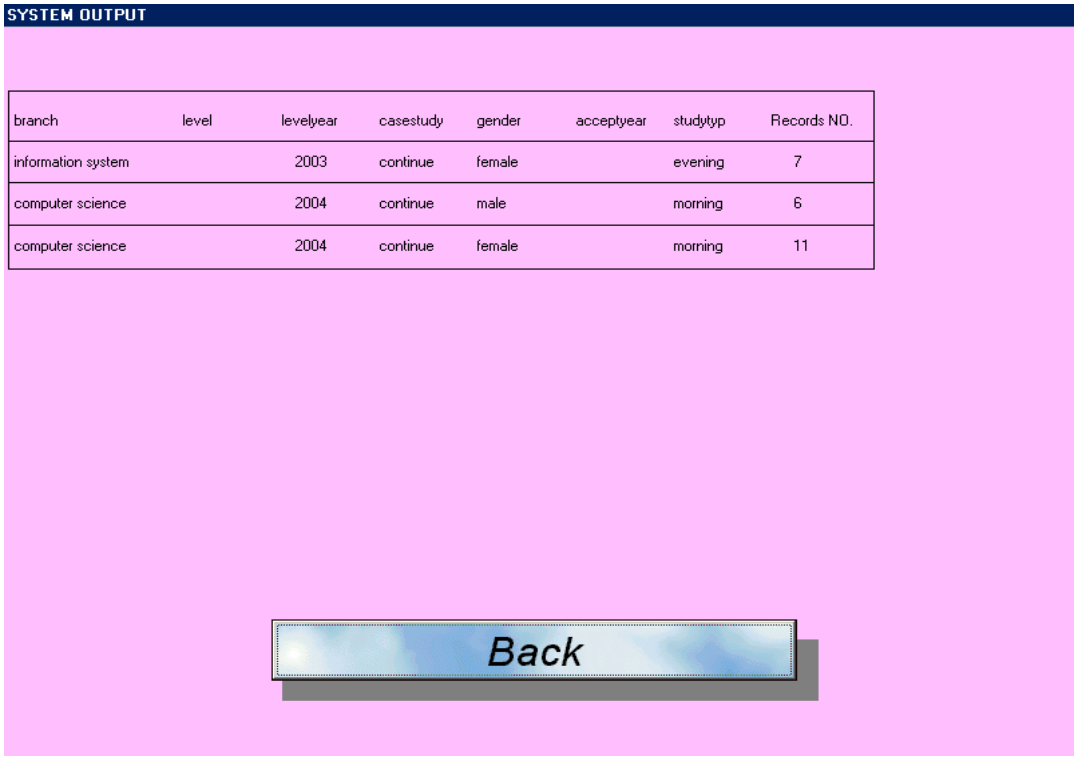

Figure 3.5 Final System Output

3.9 Output of Proposed System

It is represents the final output of the system as we see in the figure 3.5, the table represents the attribute name and its specific value in that general tuple which is contains the pure attributes resulting from AOI process after applying the generalization and the last column of tuple (Record No.) represents the specific records numbers which that generalized tuple cover, the table in the Figure

3.9 tell us There are 20 students that have good degrees; all of them female and belong to computer science branch.

Conclusions

Our research is the first one combine between two different heuristic search(GA and Data Mining), since it is used GA output as an A.O.I. input and also used Association Rule which it is another Data Mining techniques as a GA fitness function and by this ultimate advantage used of multi heuristic search we conclude a large reduce of exponential computation, addition to canceled the human mistake by make the GA work his job which it is relative attributes selection from large numbers embedded in database.

Since the AOI result of the tool appears to be promising, we are setting up research plan to evaluate this tool thoroughly. Also GA operations of our research have two fitness functions, which it is represent by association rule factors: confidence factor as the first fitness function and support factor as the second fitness function. To summarize the results of research, we summarize as below:

1- GA aids the AOI algorithm by removing for these attributes which fail to associate with user data mining query.

2- The aim of our research is reduce manual work and that is done by using expert interface and GA which enables the user to write in NL without entering the relevant attributes and their hierarchy.

3- Using GA leads to more meaningful information by discovered the precisely number of associated attributes to the task relevant database.

4 Thresholds are important step for the AOI output to make it more informative by select the suitable thresholds and that done by reduce the thresholds (rolling up) or increase them (drilling down) .

References

[1] Jamal A. A., MSc Research, Using Data Mining for Decision Support, Iraq Commissions for Computers and Informatics/ Informatics Institute for Postgraduate Studies, 2002.

[2] Jiawei H., Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, 2001.

[3] David E. G., Genetic Algorithms, Addision Wesley, 1999.

[4] Peter C., Discovering Data Mining from Concept to Implementation, Prentice Hall PTR, 1998.

[5] Alaa H.H., Mining of Maximal Itemsets by Using Machine Learning, PDF, Al-Rafedain University College/Computer Science Department, 2003.

[6] Jiawei H., Exploration of the Power of Attribute – Oriented Induction in Data Mining, PDF, Simon Fraser University, 1995.

[7] Imad F. T., Ph.D. Research, Solution of Some Hard Problems Using Genetic Algorithms (GAs), University of Pune, 2000

[8] S. Choenni, On the Suitability of Genetic – Based Algorithms for Data Mining, Springer verag (serie LNCS), 1998.

تطبيق خوارزمية تعدين البيانات مع الخوارزميات الجينية

ميثم مصطفى حمود

كلية علوم الحاسبات والرياضيات / جامعة تكريت

الملخص:

أن تقنية أستقراء الصفة الموجهة تعتبر من التقنيات الحديثة الخاصة بتعدين قواعد البيانات و هي مصممة لحل المشاكل المستعصية على بقية طرق التعدين وهي تعتمد على مبدأ التعميم للصفات الخاصة بقواعد البيانات أستناداً الى أستخدام نوعيين من العتبات اللذان يحددان مدى الصعود للحصول على صفات أكثر عمومية أو النزول للحصول على صفات أكثر تخصص. والنتجة النهائية لعملية التعميم هو الحصول على صفات نجحة بأجتيازفحص التعميم و المتمثل بالنجاح بفحص العتبتين. أن كل قيمة من قيم هذه الصفات الناجحة تمثل ظاهرة تم أستخلاصها من قاعدة البيانات التي تمت عليها عملية التعدين. لكن المشكلة تكمن في أن تقنية أستقراء الصفة الموجهة تطلب من المستخدم بأن يختار صفات معينة ذات علاقة بالمشكلة التي يطرحها والتي هي عملية ليست بسيط حين تكون صفات قواعد البيانات تتراوح بين ٥٠_١٠٠ صفة مع أحتمالية بأن يكون المستخدم قليل المعرفة بتلك الصفات، أن هذه المشكلة تعتبر ذات فضاء بحث كبير و هذا يتناسب مع عمل الخوارزمية الجينية لذا تم أستخدامها في هذا البحث لتحديد تلك الصفات بدل من طلب ذلك من المستخدم. أن قوة الخوارزمية الجينية تكمن بتحديد دالة الهدف فأذا تم تحديدها تبقى بقية العمليات الخاصة بالخوارزمية الجينية هي عملية روتينية وقد تم تحديدها بهذا البحث بنجاح عن طريق أستخدام تقنية الدوال التجميعية كدالة هدف.