Improving Classification Knowledge Using An Integrated Knowledge Discovery Approach

Ahmed Rafea[1] , Said Mabrouk[2], Ahmed Yousef[2]

[1] Computer Science Dept., American University in Cairo
113 Sharia Kasr El-Aini, Cairo, Egypt
rafea@aucegypt.edu
[2] Comp. Science. Dept., ISSR, Cairo University
5, Tharwat Street, Giza , Egypt
said51@hotmail.com, ayas@hotmail.com

Abstract

Attribute-oriented induction approach (AOA) has been developed for knowledge discovery in large relational database. Several kinds of knowledge, such as characteristic rules and discrimination or classification rules can be discovered. These rules may contain unnecessary conditions and/or unnecessary condition-values. A Tuple-oriented approach (TOA) examines one tuple at a time since there are large number of possible combination in such testing, this approach is quite inefficient when performing learning from large databases. This paper introduces an Integrated Discovery System (IDS) based on combining techniques from both the AOA and the TOA. It captures the advantages and overcomes the difficulties associated with each of these approaches when used separately. Therefore IDS discovers more efficient classification rules, compared with the rules discovered by the pure AO.

Key words: Data Mining, Knowledge Discovery, Attribute Oriented Approach, Tuple Oriented Approach

1 INTRODUCTION

Knowledge Discovery in Databases (KDD) is an active research area with the promise for a high payoff in many business and scientific applications. The grand challenge of Knowledge discovery in databases is to automatically process large quantities of raw data, to identify the most significant knowledge and to present this knowledge in an appropriate form for achieving the user's goal [4,7].

The objective of this paper is to present a proposed system that integrates a variety of knowledge discovery algorithms namely: attribute-oriented algorithms [2,10], MIN-ATTR algorithm [5] for minimum attributes values and MIN-RULE algorithm [5] for minimum rules values. This integration allows us to exploit the strengths of diverse discovery techniques. The proposed integrated discovery system (IDS), has been developed and applied to discover classification knowledge from a database on Egyptian Scientists Living- Abroad The system was able to discover more efficient classification knowledge compared with the knowledge discovered by the pure attribute-oriented approach.

Section 2 presents a background on previous related algorithms. Section 3, presents the architecture of the integrated model and illustrates the different discovery modules. Section 4, illustrates a case study. We conclude with section 5.

2 RELATED WORK

This section reviews previous algorithms of Attribute-Oriented and Tuple Oriented approaches.

2.1 Attribute Oriented Approach

Attribute-oriented induction approach [2,10] has been developed for knowledge discovery in relational database. Several kinds of knowledge, such as characteristic rules and discrimination or classification rules can be learned.

The discovered rules are not efficient since they may contain unnecessary conditions and / or unnecessary condition-values.

The algorithm for discovering characteristic rules applies an attribute oriented concept tree ascending technique which substitutes the lower level concept of each attribute in a tuple by its corresponding higher level concept [2,10]. Applying this algorithm on databases, one will need to get the learning task that is consisting of the target concept and relevant attributes, the concept hierarchies of the attributes, and the form in which the learning results are to be represented.

In case of discovering discrimination rules, the collected relevant data set is partitioned into two classes, one representing the target class while the other representing the contrasting class, if there are overlapping tuples in both target and contrasting class, these tuples should be marked [10].

2.2 Tuple Oriented Approach

Machine Learning methods, such as learning from examples [3,6] are also used to learn from databases. These methods are based on Tuple-Oriented approach, where one tuple is examined at a time; efficient rules can be learned. Since there are large number of possible combination in such testing, this approach is quite inefficient when performing learning from large databases.

Two tuple-oriented algorithms are applied to decision tables with binary attributes Min-attribute and Min-rules algorithms [5]. First algorithm is used to eliminate the redundant attributes in the set of learned rules and find the

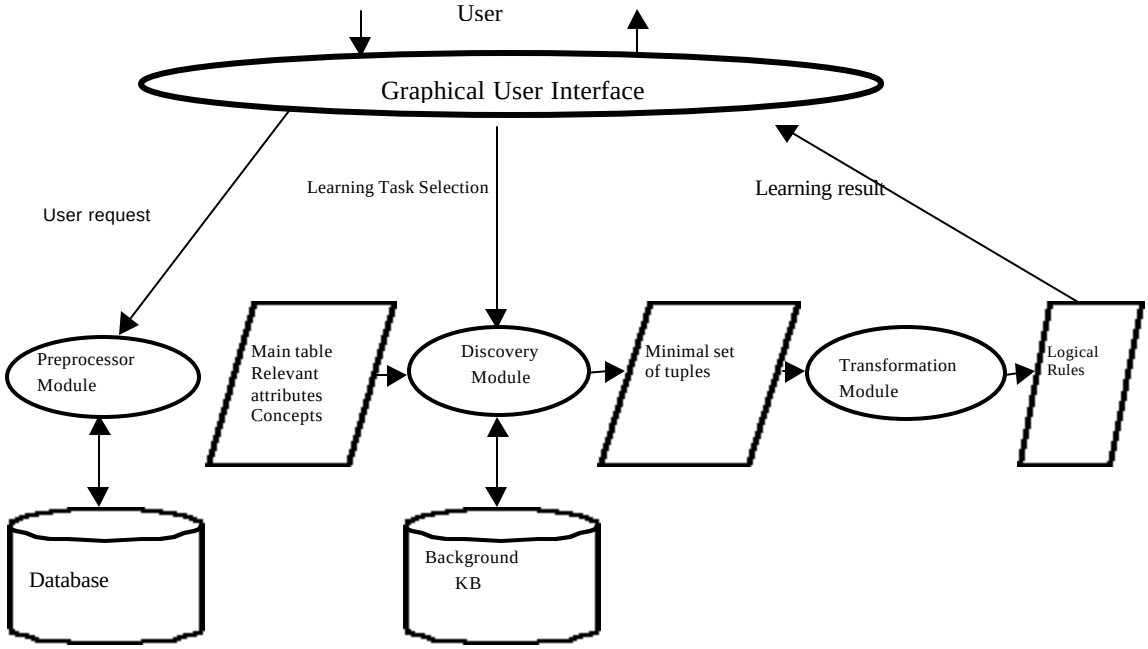

Fig. 1. Integrated Discovery System Architecture

minimal set of attributes in the rules. The second algorithm removes the redundant attribute-values and minimizes the number of rules describing the target concepts. These algorithms are not suitable, as they are, to be used with databases because they are only dealing with binary attributes.

3. THE PROPOSES SYSTEM

The proposed knowledge discovery system uses three main techniques that constitute the core of this work. These techniques are data generalization, attribute reduction, and rules reduction. In data generalization, attribute induction is applied to the initial database relation, using tree ascension and attribute removal generalization rules to obtain a prim relation. The generalized prim relation only contains a small number of tuples and it is feasible to apply tuple-oriented techniques to eliminate the irrelevant or unimportant attributes and choose the best minimal attribute set. In the data reduction phase, our method finds a minimal subset of interesting attributes that have all the essential information of the generalized relation, thus a subset of the attributes can be used instead of the whole attribute set of the generalized relation. Finally the tuples in the reduced relation are transformed into logical knowledge rules.

Figure 1, illustrates the basic architecture of the proposed Integrated Discovery System (IDS). It consists of four modules: User interface, Preprocessor, Discovery, and Transformation modules. It also contains the background KB and the Databases.

3.1 Graphical User Interface (GUI)

It facilitates communication between the system and the user, by accepting user request, target and contrasting concepts, relevant attributes, and learning tasks. It then provides this information to the preprocessor and the discovery modules. It also browses the learning result for the user.

3.2 Preprocessor Module

This module is responsible for creating initial tables, selecting relevant attributes, loading database, and describing learning concept.

3.3 Discovery Modules

This module is responsible for discovering non-enhanced / enhanced rules through four sub-modules as shown in Figure 2. These sub-modules are the characterizer, the discriminator, the min-attribute, and the min-rule.

The Characterizer. This algorithm discovers a set of characteristic rules from the relevant set of data in a database. A characteristic rule summarizes the general characteristics of a set of user-specified data. The core of this module is the attribute-oriented algorithm described in [2} and [10].

The Discriminantor. This algorithm discovers a set of discrimination rules from the relevant set(s) of data in a database. A discrimination rule distinguishes the general features of one set of data, called the target class, from some other set(s) of data, called the contrasting class(es).

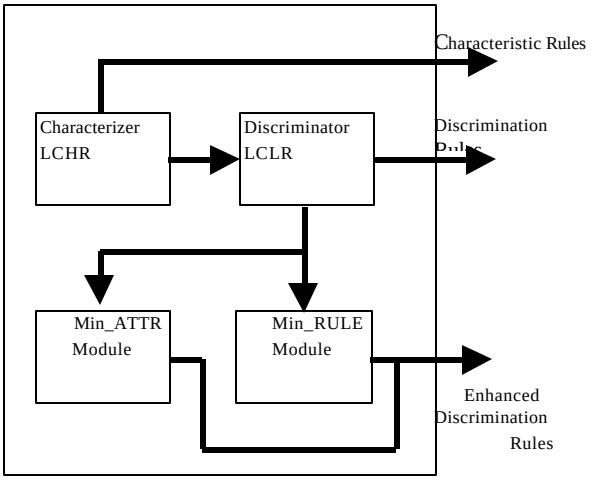

Fig. 2. Discovery Modules

Attribute-Oriented algorithm described in [10] is used in this module.

The Min-Attribute. This algorithm eliminates unnecessary attributes by performing dependency analysis (attribute reduction). The algorithm described in [5] is modified to cover all kind of attributes not only binary value attributes. The algorithm is as follows:

```
For each attribute Aj,1 <= j <= n ,do
{
{For each tuple Ti, 1 <= i <= n do
For each tuple Tk ,i+1 <= k <= n do
If any two tuples Ti and Tk have
identical values for Subset of
attributes A –{ Aj }, but  have
different class values
          Then Attribute Aj is needed;
Break
          Else continue
}
If Aj is needed
then continue
Else remove this attribute
}
```

The Min-Rule. This algorithm eliminates redundant attribute values and minimizes the number of rules (value reduction). The algorithm described in [5] is modified to cover all kind of attributes not only binary value attributes. The algorithm is as follows:

```
For each column attribute do
  For each tuple Ti, 1   i   n do
      For each tuple Tk ,i+1   k   n
         If the two tuples Ti and Tk
           have identical values for
           all columns attributes –
           {current column attribute
           value}, but have different
           class values
           Then Value of this column
             attribute is needed;
         Else the value of the tuple
           corresponding to the class
           under discovery column
           attribute value is replaced
           by do not care
```

3.4 Transformation Module

This module is responsible for transforming the generalized minimal set of tuples into the logical rules.

3.5 Database and Background Knowledge

Relational database files are used as the source of the raw data to learn from. Database operations are used to retrieve and select the initial data. Attributes in the database are numeric, nominal and structure. Each attribute is represented in the knowledge base in terms of attribute-hierarchy, which is used during the generalization process.

3.6 Rules-Discovery Flow

The flow of rules discovery can be briefly described in the following steps:

The preprocessor module receives the user request through the graphical user interface. This request contains the database table name, relevant attributes, and the learning concept attributes.

The preprocessor module invokes the database component to extract the main table, relevant attributes, and the learning concept attributes.

The discovery module receives the output of the preprocessor module and the learning task from the user, and it uses the background knowledge to generalize the relevant attributes and creates the rules. These rules could be as follows :

- Characteristic rules, if the characterizer module is applied on the target data.
- Discrimination rules without enhancement, if the discriminator module is applied on target and contrasting data.
- Enhanced discrimination rules, when applying the Min-attribute or Min-rule module on discrimination rules.

d. Finally all rules are transformed into logical formulae by the transformation module and then the learning results are browsed for the user.

4.CASE STUDY

The aim of this proposed case study is to test the integrated discovery system (IDS) modules by using sample data of Scientists Living-Abroad database. It is used by the different organization to extract more valuable information about the experts, whenever an advanced technology issue is urgently needed [1]. The integrated discovery system (IDS) is implemented on IBM-PC using Delphi-V3.1 with Open DataBase

Connectivity (ODBC) library, and Microsoft Access database.

Section 4.1 describes the database and background knowledge of the Scientists Living Abroad. The other four sections present four experiments for testing the different discovery modules.

4.1. Database & Background Knowledge Description

Each attribute is represented in knowledge base by a tree, part of the tree of the Specialty attribute is shown in Figure 3.

The Scientists database consists of 7 entities namely city, qualification,, position-name, position-place, specialty, experience, and basic table The Basic Table consists of columns corresponding to the first six entities

{ Engineering, Science, ..., Arts } ñ Any (Specialty)
{ Art, Egyptology, ..., Sociology } ñ Arts
{ Computing, Chemistry, ..., Physics } ñ Science
{ Electrical, Mechanical, ..., Construction } ñ Engineering
{ E-computer, Communication , ..., Power } ñ Electrical
{ Vehicle, Air/c, ..., Ships } ñ Mechanical
{ Math, S-computer, ...} ñ Computing
{ Biology, Environmental, ...} ñ Chemistry
{ Physics, nuclear,...} ñ Physics

Fig. 3 "Specialty" Concept Hierarchy

4.2 Experiment 1: Test The Characterizer Module

Objective. The objective of this experiment was to learn Characteristic rules from Scientists-Living-Abroad database for the compound concept (Country = USA & Experience = Nuclear).i.e., the rules that characterize the scientists living-abroad whose experience in nuclear domain and live in USA.

Result. The Characterizer algorithm succeeded in generating 4 rules, represented by 24 tuples of target class from 50 tuples of database. These rules are shown in figure4 where:

Rule1 represents concept (usa & nuclear) with weight 20.83%
Rule2 represents concept (usa & nuclear) with weight 20.83%
Rule3 represents concept (usa & nuclear) with weight 25.00 %
Rule4 represents concept (usa & nuclear) with weight 33.33%

The weight of the rule represents the percentage of the tuples belonging to the target class that are covered by the rule.

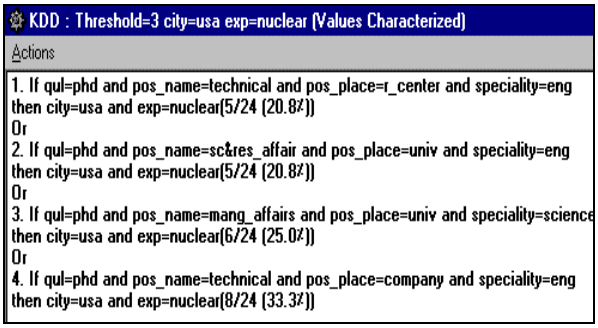

Fig.4 Characteristic Rule

4.3 Experiment2: Test the Discriminator Module

Objective. The objective of this experiment is to learn Discrimination rules from Scientists-Living-Abroad database with learning compound concept (Country = USA & Experience = Nuclear) i.e., the rules that discriminate the nuclear scientists living in USA from other scientists.

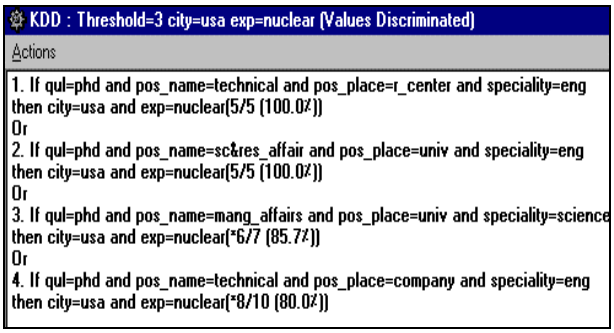

Fig. 5 Discrimination Rules

Result. The Discriminator algorithm succeeded in generating 4 rules, represented by 24 tuples of target class and 26 tuples of contrasting class. These rules are shown in Figure 5 where:

Rule1 represents concept (usa & nuclear) with weight 85.7 %
Rule2 represents concept (usa & nuclear) with weight 85.7 %
Rule 3 represents concept (usa & nuclear) with weight 85.7%
Rule 4 represents concept (usa & nuclear) with weight 80.0%

The weight associated with discriminate rules reflects the degree of confidence in the rule.

4.4 Experiment 4: Test the Min-Rule Module

Objective. The objective of this experiment is to test the effect of the Min-Rule algorithm in improving the set of rule obtained from experiment 2.

Result. The Min-Rule algorithm succeeded in eliminating 6 of nonessential attribute values as shown in figure6.

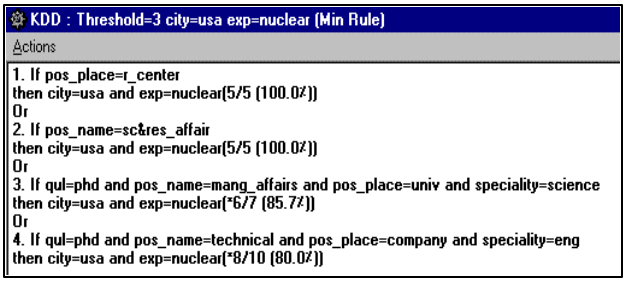

Fig.6. Min-Rules Formulae

The number of condition in the rules has been decreased to an average of 2.5 conditions per rule instead of 4 conditions per rule. As one can see Rule 1 before applying Min-Rule algorithm was:

```
If qul = Ph.d and pos-name = technical and
pos-place = r-center and specialty = eng
Then country = USA and exp = Nuclear 100.0%
```

After applying Min-Rule algorithm, Rule1 becomes

```
If pos-place = r-center Then country = USA
and exp = Nuclear 100.0%
```

It is clearly noticed that the discrimination rule values are optimized by applying the Min- Rule algorithm. The total number of rules has not decreased. However, in other situation the application of the Min-Rule algorithm may lead to reduction of the number of discrimination rules.

4.5. Experiment 4: Test the Min-Attribute Module

Objective. The objective of this experiment is to test the effect of the Min-Attribute algorithm using different learning concept (city = quebec), By learn discrimination rules from Scientists-Living-Abroad database for this and then apply the Min-Attribute algorithm.

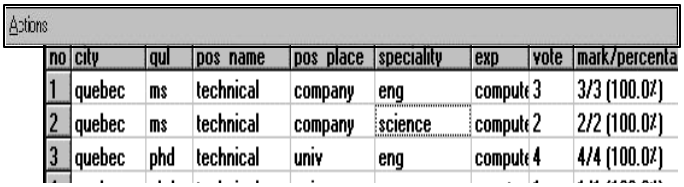

Fig. 7 Discrimination rules before using min-attr algorithm

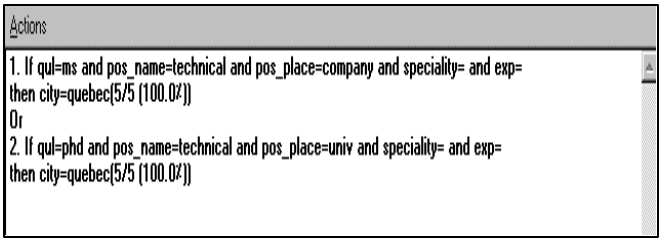

Fig. 8 Discrimination rules after using min-attr algorithm

Result. Figure 7 and Figure 8 show the discovered discrimination rules for the target concept without and with using the min-attribute algorithm. The Min-Attribute algorithm succeeded in eliminating the unnecessary attributes. The number of discrimination rules has been reduced to 2 rules instead of 4 rules. The number of conditions has been reduced to the average of 3 conditions per rule instead of 5 conditions per rule before applying the algorithm.

Rule 1 before applying Min-Attribute

```
If qul = ms and pos-name = technical and
pos-place = company and specialty = eng and
exp = computer Then city = quebec
```

Rule 1 after applying Min-Attribute :

```
If qul = ms and pos-name = technical
and pos-place  = company Then city =
quebec
```

It is clearly observed that applying the Min-Attribute algorithm has optimized the discrimination rule attributes

5. CONCLUSION AND FUTURE WORK

A framework for knowledge discovery in databases integrating techniques from both the attribute-oriented approach and tuple-oriented approach was presented. Attribute-oriented approach was mainly used to generate characteristic and discrimination rules from a database by generalizing attributes using background knowledge. The main objective of using the attribute-oriented approach was to reduce the number of tuples in the database. Two tuple-oriented algorithms were developed and used to eliminate the unnecessary attributes and nonessential attribute values, and hence enhanced discrimination rules were extracted.

In this implementation, extra features were included to learn compound concept, to adjust the threshold value dynamically for each attribute, and to friendly interact with the user.

The developed system was tested using a real database for Scientists Living Abroad. Applying the attribute-oriented and tuple-oriented algorithms on this database enhanced the discovered discrimination rules.

Issues related to the work presented in this paper and that needs further investigation are handling identical rules in the target and contrasting classes, and dealing with numerical attributes. The identical rules in both target and contrasting classes are called overlapping rules, each rule has a percentage value of confidence with respect to the equivalent rule in the other class. This type of rules needs a special handling to distinguish the target class from the contrasting class. Discovering rules with empirical mathematical equations in the right hand side, from database is also a research point that needs further investigation.

6. REFERENCES

1. Yousef, A.: “Knowledge Discovery from database based on inductive-deductive techniques”, Msc thesis, ISSR, Cairo University (1999).

2. Han, J. ,Cai, Y., and Cercone.,N. : "Knowledge Discovery In Databases : An Attribute-Oriented Approach". In Proc. of the 18th VLDB conf. Vancouver, Canada (1992).
3. Michalski R.S, Carbanell, J.G. :Machine Learning: An Artificial Intelligence Approach. Tioga, Palo Alto, ea. (1983).
4. Chen, M.S., Han, J.and Yu, P. S. : Data Mining: An Overview from Database Perspective, IEEE Transactions on Knowledge and Data Engineering (1997).
5. Shan, N., Hamilton, H.J., and Cercame,N. : "Learning Decision Rules in Parallel " Proc. of the 9th Florida Artificial Intelligence Research Symposium. (1996).
6. Shi, Z. (1992). Principle of Machine Learning. International Academic Publishers.
7. Fayyad, U. M., Shapiro, G. , Smyth, P.and Uthurusamy,R. : Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press (1996).
8. Hu, X., Cercone,N., Han,J. : A Rough Set Approach for Knowledge Discovery in Databases, Rough Sets, Fuzzy Sets and Knowledge Discovery, Springer-Verlag Press, W. Ziarko(ed), (1993). 90-99.
9. Hu, X.: Knowledge Discovery in Database : An attribute-oriented Rough Set Approach. Regina, Suskatchewan, Canada (1995).
10. Cai,Y., , Cercone, N. and Han, J. : Attribute-Oriented Induction in Relational Databases, in Shapiro, G. and Frowley, W. (eds), Knowledge Discovery in Databases, AAAI/MIT Press. (1991).