ijcrr
Vol 04 issue 06
Category: Research
Received on:02/01/12
Revised on:20/01/12
Accepted on:12/02/12

A NOVEL APPROACH FOR MINING PECULIAR DATA FROM LARGE DATA SET USING PATTERN MATCHING AND PECULIAR RULE MINING

S.Shahar Banu[1], V.Saravanan[2]

[1]B.S.Abdur Rahman University, Vandalur, Chennai
[2]Dr. NGP Institute of Technology, Coimbatore

E-mail of Corresponding Author: shaharmohamed@gmail.com

ABSTRACT

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. There are many types of data mining techniques. The main and most objective mining method is peculiarity mining. This method mines the peculiar data among the large set of data. In this paper there are certain set of rules which found to find relevant data among large set of data in large set of databases.

Keyword: Data mining, peculiar data,peculiar rules,pattern matching

1. INTRODUCTION

Data mining or knowledge discovery refers to the process of finding interesting information in large repositories of data. The process of data mining is composed of several steps including selecting data to analyze, preparing the data, applying the data mining algorithms, and then interpreting and evaluating the results. The application of data mining techniques was first applied to databases. A better term for this process is KDD (Knowledge Discovery in Databases).Data mining (DM) is a multi staged process of extracting previously unanticipated knowledge from large databases, and applying the results to decision making. Data mining tools detect patterns from the data and infer associations and rules from them. The extracted information may then be applied to prediction or classification models by identifying relations within the data records or between databases. Those patterns and rules can then guide decision making and forecast the effects of those decisions.

In order to discover new, surprising, interesting patterns hidden in data, peculiarity oriented mining and multi database mining are required. The main objective of this work is to fetch the peculiar data. The availability of large quantity of data in the large set of databases from the World Wide Web and business data management systems has made the dynamic separation of data into new categories as a very important task for every business intelligence systems.

We find the association or relation between the dataset in the databases. The main aim is to fetch the peculiar data among the data. This paper describes the attribute level

entity level and record level peculiarity to get the rules. Several software implementations are carried out to demonstrate the peculiarity-based mining. The term data mining also refers to the step in the knowledge discovery process in which special algorithms are employed for identifying interesting patterns in the data. These interesting patterns are then analyzed yielding knowledge.

2. Literature Review

Ribeiro, Kaufman and Kerschberg,[1995] have described a way of multi-database mining by incorporating primary and foreign keys, as well as developing and processing knowledge segments[1]. Wrobel[1997], has extended the concept of foreign keys to include foreign links, since multi-database mining also involves accessing non-key attributes. Aronis *et al.* introduced a system called WoRLD that uses spreading activation to enable inductive learning from multiple tables in multiple databases spread across the network.

Liu, Lu and Yao [1998],have proposed an alternative multi-database mining technique that selects relevant databases and searches only the set of all relevant databases. Their work has focused on the first step in multi-database mining, which is the identification of databases that are most relevant to an application. A relevance measure was thus proposed to identify relevant databases for mining with an objective to find patterns or regularity within certain attributes. This can overcome the drawbacks that are the result of joining all databases into a single huge database upon which existing data mining techniques or tools are applied. The approach is effective in reducing search costs for a given application.

Zhong[1999] have proposed a way of mining peculiarity patterns from multiple statistical and transaction databases based on previous work. A peculiarity pattern is discovered from the peculiar data by searching the relevance among the peculiar data. A data item is peculiar if it represents a peculiar case described by a relatively small number of objects and is very different from other objects in a data set. Although it looks like an **exception pattern** from the viewpoint of describing a relatively small number of objects, the peculiarity pattern represents a well-known fact with common sense, which is a feature of the general pattern. Wu and Zhang[2001] have advocated an approach for identifying patterns in multi-database by weighting .Kargupta [2001], have built a collective mining technique for distributed data. Grossman have built a system, known as Papyrus, for distributed data mining. Existing parallel mining techniques can also be used to deal with multi-databases.

K-means is the simplest and the most popular clustering technique that is widely used in various fields of science and technology. The medical industry is also increasing with the data for aids patients. It is difficult for classifying and finding the DNA pattern of the AIDS documents. We use pattern matching and/or document clustering analysis in the research area of artificial intelligence and data mining. Its fundamental task is to utilize the alphabets to compute the percentage of related relationship between the records or the documents and to accomplish automatic classification without earlier knowledge. Document clustering is to utilize clustering technique to gather the documents of high resemblance collectively by computing the documents resemblance.

There are several pattern matching and clustering approaches available in the literature to fetch the relevant data, record or the document in distributed environment. But most of the existing mining techniques suffer from a wide range of limitations. The existing mining

approaches face the issues like practical applicability, very less accuracy, scalability, more classification time etc. Thus a novel approach is needed for providing significant accuracy with less classification time. Also, mining need to mine the peculiar data from the dataset. Whenever we use data mining techniques it gives the 80% relevant and 20% irrelevant data from the dataset, but there is no peculiarity. Here the specialty and the main objective of the thesis is bring the peculiar data from the dataset.

3. Proposed work:

The main aim of this work is to develop an improved peculiarity mining technique with very high classification accuracy. Peculiarity rules are a new class of rules which can be discovered by searching relevance among a relatively small number of peculiar data. Peculiarity oriented mining in multiple data sources is different from, and complementary to, existing approaches for discovering new, surprising, and interesting patterns hidden in data.Within the proposed framework, we give a formal interpretation and comparison of three classes of rules, namely, association rules, exception rules, and peculiarity rules, as well as describe how to mine interesting peculiarity rules in multiple databases. Peculiarity represents a new interpretation of interestingness, an important notion long identified in data mining.

Peculiarity, unexpected relationships/rules, may be hidden in a relatively small number of data. Peculiarity rules are a typical regularity hidden in many scientific, statistical, and transaction databases. They may be difficult to find by applying the standard association rule mining method due to the requirement of large support. In contrast, peculiarity oriented mining focuses on some interesting data (peculiar data) in order to find novel and interesting rules (peculiarity rules). The second keyword is multiple databases, which are the objects of discovery and learning. Mainstream KDD (Knowledge Discovery and Data Mining) research is limited to rule discovery in a single universal relation or an information table.

Multidatabase mining is to mine knowledge in multiple related information sources. By considering the two related issues of peculiarity and multiple databases, we propose a framework of peculiarity oriented mining in multi databases. The identification of peculiarity rules, as well as algorithms of mining peculiarity rules, will enhance the effectiveness of data mining and extend its domain of applications. Studies on peculiarity oriented mining can be divided into three phases:

1. Developing methods of peculiarity oriented mining,
2. Extending peculiarity oriented approaches to multiple data sources, and
3. Enabling peculiarity oriented mining in a distributed and cooperative mode.

There are various problems associated with the existing data mining approaches. Existing data mining algorithms suffer from problems of practical applicability. The accuracy of the existing DM approaches is a major concern. The time taken for active DM is more in large databases. The main aim of this work is the development of an improved peculiarity mining technique with very high classification accuracy. Peculiarity rules are discovered from peculiar data evaluated using unified knowledge-based statistical criteria. The main task of mining peculiarity rules is the identification of peculiar data. Peculiar data are a subset of objects in the database and are characterized by two features:

1) very different from other objects in a data set and

2) consisting of a relatively small number of objects

- **Relevance among Peculiar Data**

A peculiarity rule is discovered by searching the relevance among peculiar data. Let X(x) and Y (y) be peculiar data found in two attributes X and Y, respectively. We deal with the following two cases:

If both X(x) and Y (y) are symbolic data, the relevance between X(x) and Y (y) is evaluated by:

R1 = P(X(x)|Y (y)) P(Y (y)| X(x) 1.

That is, the larger the product of the probabilities, the stronger the relevance between X(x) and Y (y) is. If both X(x) and Y (y) are continuous attributes, the relevance between X(x) and Y (y) is evaluated by using the method developed in the KOSI system that finds functional relationships. Equation (1.) is suitable for handling more than two peculiar data found in more than two attributes if X(x) (or Y (y)) is a granule of peculiar data. The above-stated methodology can be extended for mining from multiple databases.

The proposed approach is evaluated using the datasets namely real time Data set, AIDS patient's data set, collected from **AIDS counseling centers**. There are various ways used to improve the performance of the proposed approach. The parameters used for evaluating performance are Time, Accuracy.

4. Implementation

The medical AIDS patient's data consists of multiple records in multiple date with peculiar cases. The only peculiarity data is been mined using association rule, exception rule and peculiarity rule. Then finally the performance is evaluated according to the three approaches [rules] which are used. The medical data is collected from the Government Hospital AIDS counseling centers.

4.1 Algorithm for finding peculiar data

1. Initialize the p (Pattern)
2. Retrieve a record R and read the origin field O
3. string s <- O
4. assign count <- 0
5. for i=1 to len(s)
6. if (p.equals(substring(i, len(p),s)))
7. count++
8. end for
9. if count >=50 then R is peculiar data and display R
10. else
11. display count
12. end if
13. step 2 until EOF
14. stop

In a single system we developed a small code in dot net framework which is connected with SQL and Ms-access data storage. The data are inserted and retrieved using the peculiarity mining technique. It is nothing but while retrieving the data which says the peculiarity by counting the number of occurrences of the pattern in the origin field of the each record. The length of the origin field is more than 500 characters. It is a combination of AGCT molecule of the DNA. where the pattern indicates the diseases of Malaria, Flue, AIDS etc., In this work we are very peculiarly about the pattern which retrieve the HIV-I, disease based record.

Whenever we mine the data it says the number of occurrences only. When reach above 50% it retrieve the record and says it is a peculiar record. From the above algorithm we can find out that the pattern will be compared in the origin string from the first letter to the last letter. if it occurs the count is incremented. Here the key value is the count variable. According to the count variable value we can get the peculiar record.

5. Experimental results

From the experimental results, the this approach namely **Peculiarity rules** is to produce very good accuracy of about 99.6%, less classification time of about 0.57 seconds, better convergence in only about 20 iterations and better efficiency

.

Sr.No	Technique	Efficiency
1	K-Means Clustering	4 .0
2	Clustering with Pattern Matching	5.22
3	Peculiarity Mining	7.44

Table 1: Clustering and peculiar mining

The proposed algorithm is compared with the other techniques. Table 1 shows the various mining techniques and their efficiency.

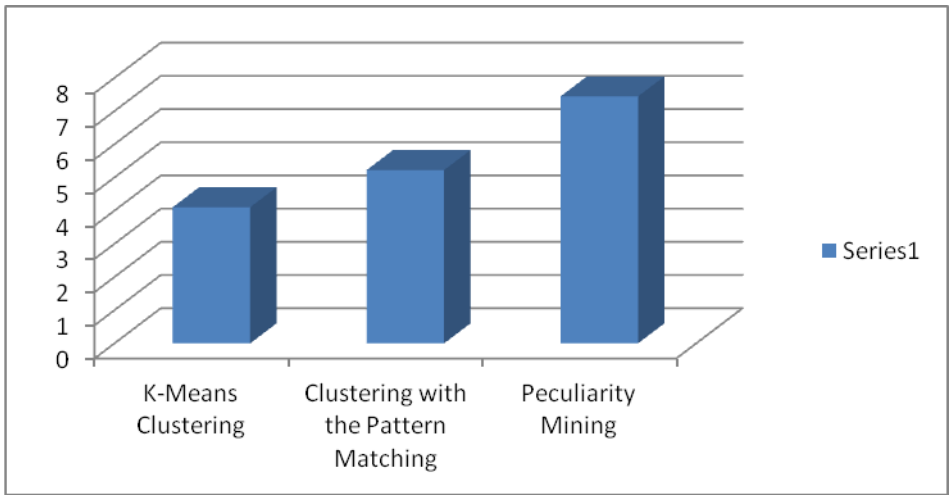

Figure 5.1 : Algorithm efficiency

From the above table Table1 and Figure 5.1,we can conclude and find out that the peculiarity mining approach is giving more efficiency than the other techniques. Where in the first approach the we are using 80 records with 5 fields and using centroid method. Also it is checked in C Language for Time complexity and efficiency.

The Second approach is done with 90 records with 7 fields and the fields are not unique. It is implemented and checked in JAVA. Finally the Peculiarity in Single system is having more than 10000 records and more than 25 Data bases. It is implemented and checked with two kinds of databases as Ms-Access and SQL server. The efficiency and complexity is much better than the other. It is find out through the coding developed in ASP.NET with C#.net. It is checked in Single system as well as in LAN. It is very good in performance level in web based Mining also. The web based mining is given for Multiple Databases.

5.1 Sample output:

The following is the output obtained from the AIDS patient's data set. Figure 5.2

showsthe peculiar data and figure 5.3 shows the number of occurrences of the particular data.

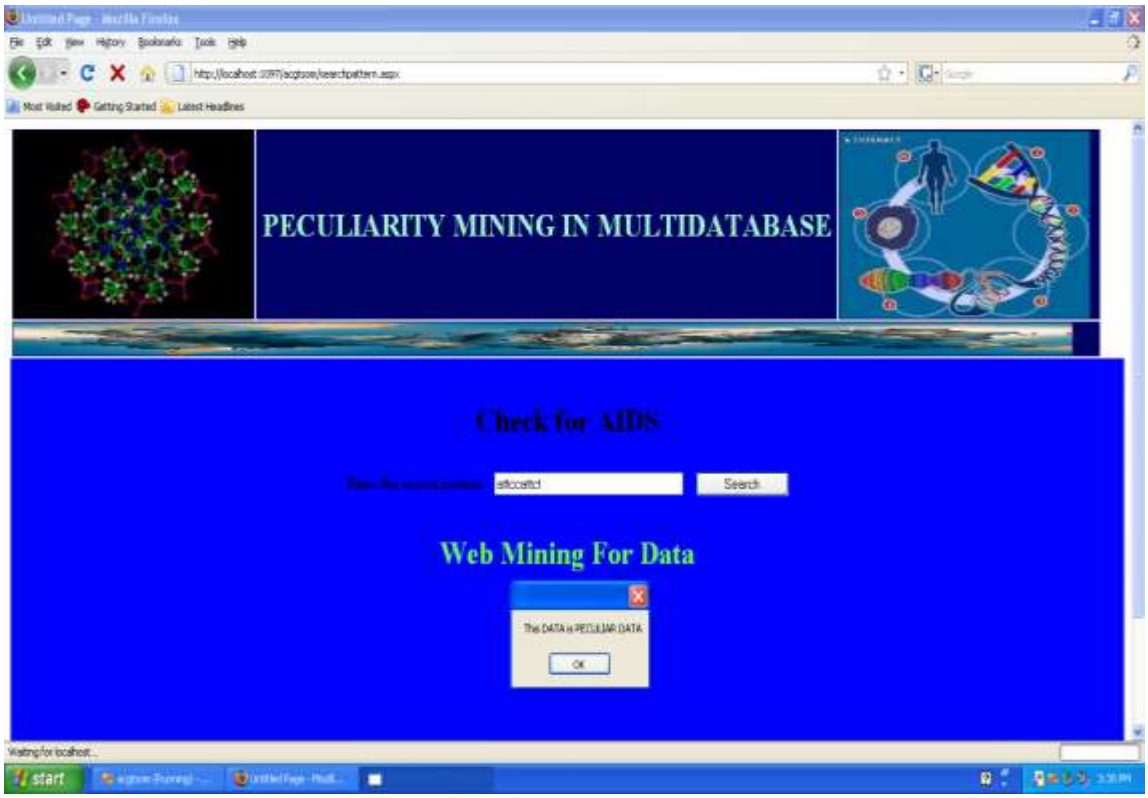

Figure 5.2: Mining for peculiar data

Figure 5.3: Number of occurrences of data

6. CONCLUSION

This paper deals with peculiarity mining with the AIDS data set, where the implementation is compared in single system, P2P system and multiple data base in web based system. In the proposed approach, the patterns are generated initially for the available vague data sets. With the help of those generated pattern, the clustering of data are carried with the help of k-means approach (modified). This proposed approach utilized a pattern matching algorithm based on multi database to search the peculiar data in the global situation.

7. REFERENCES

1. Ribeiro, K. Kaufman, and L. Kerschberg, 1995, Knowledge Discovery From Multiple databases. In: Proceedings of KDD95. 240-245.
2. S.Wrobel,1997, An algorithm for multi- Relational discovery of subgroups.In: J.Komorowski and J. Zytkow (eds.)Principle of Data Mining and Knowledge Discovery, 367-375.
3. J.Yao and H. Liu, 1997,Searching Multiple Databases for Interesting complexes. Proc. of PAKDD, 198-210.
4. H. Lu, and J.Yao, 1998, identifying Relevant Databases for Multi Database mining Proceedings of PacificAsia conf on Knowledge discover and Data mining 210–221.
5. N.Zhong,Y.Yao, and S. Ohsuga 1999 Peculiarity Oriented mining in multi Database mining Proceedings of PKDD,136-146.
6. H. Kargupta, K.Sivakumar,B.Park and S.Wang, 2000, Collective Principal Component Analysis from Distributed Heterogeneous Data. Principles of Data Mining and knowledge discovery, 452- 457.
7. S.Zhang,2001, Knowledge discovery Multi- databases by analyzing Local instances. PhD Thesis, Deakin University,
8. Kargupta,W.Huang,K. Sivakumar, And E. Johnson, 2001, distributed clustering Using collective principal component analysis. Knowledge and Information Systems, 3(4) : 422-448.
9. Zhang and S. Zhang,2002 , Association Rules Mining: Models And algorithms. Springer- Verlag Publishers in Lecture Notes on Computer Science, p. 243.