Study of Association Rule Mining Technique for Faculty Appraisal System

Mr. Pramod B Deshmukh[1], Mr. Vinod B Bharat[2], Mr. Suraj V Kurde[3], Mr. Rahul Y Pawar[4]
Miss. Deepali A Ghanwat[5]

[1] Asst. Professor, Department of Computer Engineering, D Y Patil College of Engineering, Akurdi, Pune, MH, India.
[2] Asst. Professor, Department of Computer Engineering, D Y Patil School of Engineering Academy, Ambi, Pune, MH, India.
[3] Asst. Professor, Department of Computer Engineering, D Y Patil School of Engineering Academy, Ambi, Pune, MH, India.
[4]Asst. Professor, Department of Computer Engineering, D Y Patil College of Engineering, Akurdi, Pune, MH, India.
[5] Asst. Professor, Department of Electronics& Tele-communication Engineering, Shri Chhatrapati Shivaji Maharaj COE, Nepti, Ahmadnagar, MH, India.
MH, India

ABSTRACT

Data mining is the process of finding correspondence or patterns among dozens of fields in large relational databases. It deepens the data analysis, also is able to mine the interesting mode hiding in mass data automatically. As a new data analysis technology, it provides for researchers of various fields with new intelligence means to realize and use data and has been put to many fields. One of its novel applications has been in the teaching sector, where it has been extensively used for the evaluation and analysis of Student and Faculty performance. This works provides a comparative analysis of the use of data mining rules in different works and gives some corrective suggestions for future works.

Keywords: - *Teaching Appraisal, Data Mining, Association Rules, Clustering, A Priori.*

1. INTRODUCTION

As a basic link of the teaching practice -the quality assessment is an evaluation of teaching process and results by using the technical and theory of teaching appraisal, which aims to improve the teaching quality and to give standards of some kind of qualification of the estimated object. Teaching appraisal is an important part of the teaching management and the primary means of teaching performance assessment. It is a process of value judgment and improvement to teaching effect and the realization degree of teaching target according to the teaching target and teaching standards It can control, guide and promote the teaching practice and also has a strong orientation. Ever since a long time, the evaluation of teacher works only focused on quantity, not on quality, which lead to unscientific evaluation and non rational analysis. If the data mining technology is used in the teaching appraisal, the defects of traditional analysis methods in it can be amended. We can further explore the relationship between the teaching quality and the teachers' age and title, also between the classroom teaching effectiveness and teachers' personal qualities. Based on the above mentioned method, by reasonably deploying the teachers of a class, students will maintain an interest and motivation to learn hard; by providing the decision-making information to teaching management, teaching work will be promoted and teaching quality will be greatly improved

1.1 Data Mining

The international definition of data mining is: Data mining is the extraction of implicit, previously unknown from application data of abundant, incomplete, obscure, random, but it is the process that have potentially useful information and knowledge [1] [2].Data mining can be operated on any type of stored information with abundant data sources, such as relational database, data warehouse, text and multimedia database, transaction database, www, etc. At present, the data mining technology has been widely used in market basket analysis, financial risk prediction, telecommunication, molecular biology, genetic engineering research, the discovery of access mode to Internet site, and information retrieval, etc[3].

1.2 The Process of Data Mining

Knowledge Discovery in Database (KDD) is a complicated processes that mining effective, novel, potentially useful and final understandable mode from database. Data Mining (DM) is the core of knowledge discovery and an important step, which uses special algorithm to extract mode from the data [4], as shown in figure 1.

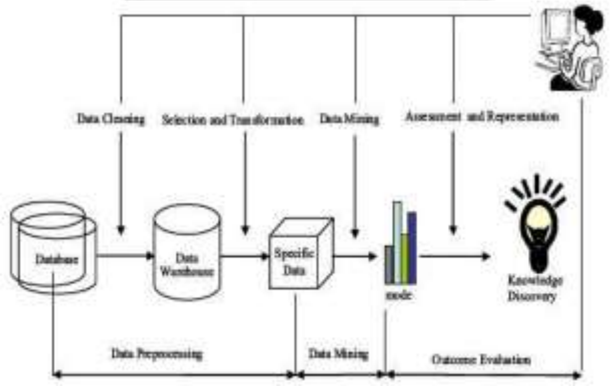

Fig-1: Data Mining is a Step of the KDD

The process of KDD is made up of following steps:

- Data Cleaning: To improve data quality by eliminating inaccurate, defective incomplete, inconsistent tuples from the original data set [5].
- Data Integration: The data used for Knowledge Discovery may come from multiple fields and systems, and it is necessary to extract relevant information from various data sources to build new data sets, as well as to eliminate redundancy of attribute.
- Data Selection: To retrieve and analyze data related to the task from a database, and abandon data unrelated to data mining.
- Data Transformation: Data have been transformed or united into suitable form for mining.

The above four steps called data preprocessing which accounts for the largest proportion in the whole process of data mining, usually 60%[6], and Data Mining, Knowledge Explanation & Evaluation account for 10%, 30% separately. Data preprocessing can improve the quality of the data, and the accuracy and performance of subsequent mining process.

- Data Mining: To make sure the task or goal of mining firstly, and then decide use which kind of mining algorithm to implement the data mining operation, and to extract data mode. Two factors shall be taken into consideration while selecting algorithm: one is different data must have different characteristics; another is the requirements of user or actual operation system should be met.
- Mode Assessment: To analyze the extractive mode at data mining stage, eliminate the redundant or irrelevant mode. If there is no mode in accordance with the requirements of users, it will return to the previous stage, such as data reselection, adopting new method on data transformation, setting new parameter values, even using other data mining algorithms.
- Knowledge Representation: It provides for users with the mining knowledge through visualization and knowledge representation technology. Data mining will face users ultimately,

so visualization should be used in found mode as much as possible, or the result should be converted into other means which is easily accessible to the users

2. DATA MINING ANALYSIS METHOD BASED ON ASSOCIATION RULES

Association means the regularity among the values of two or more data items, which can set up simple and practical association rules of these data items [7]. The purpose of the association analysis is to dig out the meaningful relationship hidden in data, to detect hidden mode not found before automatically, and provide a very important, valuable information or knowledge for decision maker. The most famous mining method of association rules is Apriori algorithm, an effective method of mining association rules from the large scale commercial data put forward by American scholars r. Agrawal and others

2.1 Related Concept of Association Rules

- Item: As for a data sheet, each field has one or more different values. Each value of fields is an item [8]
- Item sets: K Item sets means the item set with K items; K is the number of items in item sets. Maximum item sets is a set that consist of all the items, generally is represented by symbols “I”.
- Affair: Affair is a set of item sets and is a subset of item sets. Set of affairs is called affairs sets, generally is represented by symbols “D”. Each affair has a unique identity, written as TID. Supposing X is a item set, T is a affair, and $x \subseteq T$ then T include X, marked as $x \subseteq T$
- Association Rules: Association Rules are implications like as XY, in which $X \subset I$, $Y \subset I$, and $X = \varphi$. Generally, the following two parameters are used to describe characteristic of association rules:

a) Support: In transaction database D, the support of rules X Y is the ratio between the affairs number of X,Y and total affairs number which contained in affair sets, marked as support(X→Y), namely support $\subseteq \subset |$.The support describes the probability of X item sets and Y item sets appear in all affairs at the same time.

b) Reliability: Reliability is often termed Confidence. In affair sets, the confidence of rules is the ratio between the affairs number of X,Y and the affairs number of X ,marked as confidence

(Xk-1’s Y), namely confidence(X Y) $=>\{T: X \subseteq T ,T \in D \}| /|\{T: X \subseteq T,T \in D \}|$. The reliability expresses the probability of X item sets and Y item sets appear in D affair at the same time.

In order to find out meaningful association rules, two threshold values should be given: minimum support and minimum confidence. The former is the minimum support that must be met by association rules specified by users, which shows the lowest level item sets should be met in statistics sense; the latter is the minimum confidence that must be met by association rules specified by users, which reflects minimum reliability should be met by association rules. Without factoring in support and confidence of association rules, there will be too much association rules in transaction database, however, people usually have interested in association rules which can meet certain support and confidence. Therefore, a given threshold of support and confidence can limit output the quantity of association rules of the data mining system, and to provide users with a meaningful association rules as far as possible.

- Frequent item set: If the support of an itemset is equal or greater than the threshold value of support, this itemset is called frequent itemsets. Frequency set with k items is called k-frequency set, or frequent k-itemsets

2.2 Apriori Algorithm

The basic idea of the Apriori algorithm is to break the design of association rules mining algorithm down into two steps [9]:

1) To find out all itemsets whose support is greater than the minimum support, namely find out frequent itemsets;

2) To use frequent itemsets found out in the first step to generate anticipant rules.

The second step only considers the situation that there is only one item on the right side of the rules. Given a frequent item sets:$Y=I1,I2,\ldots,Ik$, $k \geq 2,Ij \in I$, and then there are at most k rules of items which come from set $\{I1,I2,\ldots,Ik\}$.

These rules such as :I1,I2,…,Ii-1,Ii+1,…,Ik_Ii,1≤ i ≤ k. In these rules, only those rules whose confidence is greater than that specified by users can be remained. In order to generate all frequent item sets, Apriori algorithm uses a recursive method. The pseudo code of the recursion algorithm can be expressed as:

```
Ck: It is a set that make up of all candidate k-itemset
Lk: It is a set that make up of all frequent k-itemset
L1= {It is a set that make up of all frequent l-itemset};
For(k=2;Lk-1 = Φ;k++)do begin.
Utilize Lk-1 to generate Ck;
For each transaction in database do Increment the count of all candidates in Ck that are contained in t;
Lk=candidates in Ck with min_support;
End;
Return= U Lk
```

Fig -2: Pseudo code of the recursion algorithm

The algorithm generates set L1 of frequent 1-itemset, and then generates set L2 of frequent 2-itemset, but the algorithm will not stop until there is some r value causes the Lr empty. When the number of recurrence is k, set Ck of candidate k item set is generated in which the items is used to generate candidate item set of frequent item set, and the last frequency Item sets Lk must be a subset of the Ck[10]. If there are m frequent itemsets in Lk-1, Ck will get m(m+1)/2 itemsets from Ck is got through Lk-1's self join, in which there are few frequent k-itemsets. Therefore, it is necessary to clip after the generation of Ck. The clip strategy is introduced into algorithm based on this property: One itemset is frequent itemset if and only if it's all subset is frequent itemset. So, if one (k-1) subset of some candidate itemset in Ck does not belong to Lk-1, this itemset should be clipped without consideration. This clip process can reduce cost while calculating the support of all candidate itemsets.

3. COMPARATIVE ANALYSIS OF RELATED WORK

Lots of work has been done in the field of teaching performance analysis using data mining association rules and various alternatives and improvements have been proposed. In [11], by Pan Qingxian et al, the author(s) have taken student evaluation as example; the author has proposed a modified Apriori algorithm to build the evaluation index system. In the suggested improved Apriori algorithm, first view mechanism is used to pre-process the original data, data meeting requirements and useful property are put into view. After producing one-dimensional frequent data items collection, according to the frequent data items collections of one-dimension and two-dimensions,

the properties of database are filtered to decrease the data properties and compress the data of database, which lays the basis of enhancing the performance of mining algorithm based on association rules.

In [12], Hibadullah, et.al, similar analysis has been made on the data of Programming students. The critical point of this study is the use of classification algorithm to extract patterns which are examined from the cognitive factor specific learning style. The data set has been analyzed using Decision tree algorithm as compared to other methods previously which have used statistical analysis approach. The findings show that that student's good performance in programming courses has a visual, active and sequential learning style.

[13] by Fazilah Hibadullah et.al does a similar analysis on an increased dataset, however, a data mining algorithm called rough set is used to understand the relationship between characteristics. A dataset of Malaysian's undergraduate student from Faculty of IT, UUM were mined using rough set. Rough set is a classification framework for discovering relationship in imprecise data and can extract hidden pattern inside data. The model used is given in Fig.2.

The proposed algorithm works as follows

1) Calculate the sum of each row in the the 0-1 matrix A ,and the number of items included in all transactions

2) Select the largest number of items k (may contain more) as alternative sets for the frequent item sets.

3) Calculate every support of the alternative sets, if support is less than the minimum support (minsup),

4) Remove the set directly from the alternative sets; if the final selected alternative set is empty, the k minus 1 to, until find the frequent item sets.

5) Among them, calculate the support of itemsets that do not use the traditional links, pruning and other steps, but use of matrix calculations directly.

6) Scan the matrix and calculate the number of rows to all 1 that is the support.

The objective is not to repeat the database scanning, thereby reducing the I / 0 computation.
Another attempt at reducing the computation load is by Lanfang Lou [16] et. al where it scans the database only once and creates a logo for each data item list. The support of frequent item sets is obtained by logo list intersection. The improved algorithm is mainly improved when generating frequent item sets. The main idea is based on the intersection operation.

Some significant results have been obtained in [17] by Here the student feedback which is the test data, after preprocessing using the WEKA preprocessor, was subjected to filter where all the numeric attributes were transformed to the nominal form and the clustering algorithm was applied on the individual as well as the set of attributes/parameters and different cluster statistics were found. After the clustering has been performed the student feedback data was subjected to association rule mining. However, the A-Priori Algorithm used is the conventional A Priori which suffers from the drawbacks of computational complexity and repeated scanning. Table I gives a comparative analysis of the existing works on the topic and their assessed performance under comparable data sets.

Table -1: Comparative analysis of Existing Literature

Sr.no	Author(s)	Algorithm	Computational Complexity	Assessed Performance
1	[11]	Modified A Priori with Database compression	Fair	Fair
2	[12]	Decision Tree	Simple	Inferior under heavy dataset
3	[13]	Rough Set	Fair	Better than [11][12] under heavy datasets
4	[15]	Modified A priori using 0-1 Transaction Matrix	Fair	On Similar lines as [11], with better execution speeds
5	[16]	Modified A Priori using Intersection operation	Complex	Better than [11], [15]
6	[17]	Clustering with Association	Complex	Best Performance under tested dataset

4. CONCLUSIONS

In this paper, its novel applications have been in the teaching sector, where it has been extensively used for the evaluation and analysis of Student and Faculty performance. This works provides a comparative analysis of the use of data mining rules in different works and gives some corrective suggestions for future works.

REFERENCES

[1]. Brachman, R, Anand, T. The Process of Knowledge Discovery in Databases:A Human-Centered Approach. U. Fryyad, G. Piatetsky- Shapiro, P. Smyth, R. Uthurusamy, Menlo Park. AAAI Press. Calif. 1996. 37-58..

[2]. Usama Fayyad, Greory Piatesky-Shapiro, Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence. Calif. 1991.

[3]. Xiaowen TANG ,Qing'sheng CAI. The Application of Data Mining in the Telecommunications Industry [J]. Computer Engineering. 2004, 30(6). 36-38..

[4]. Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques. Beijing: China Machine Press, 2001.8.

[5]. Rahm,E,Do,H. H. Data cleaning:problems and current approaches. IEEE Data Engineering Bulletin. 2000,23 (4) 3-13.

[6]. R Wang, V Storey, C Firth. A Framework for Analysis of Data Quality Research. IEEE Trans Knowledge and Data Engineering. 1995, (7). 623- 640

[7]. Ruling LU. Knowledge Engineering and Scientific Knowledge at the Era of New Century [M]. Beijing: Tsinghua University Press, 2002

[8]. Ian H.Witten,Eibe Frank,Data Mining Practical Machine Learing Tools and Techniques [M]. Beijing:China Machine Press,2006.2.

[9]. Xinning SU, Jianlin YANG, Niannan JIANG.. Data Warehouse and Data Mining. Beijing: Tsinghua University Press, 2006. 4

[10]. Xiaodong Kang,Data Mining technology Based on Data Warehouse. [M]. Beijing:China Machine Press,2004.1.

[11]. Pan Qingxian, Qu Linjie, and Lou Lanfang, "Data Mining and Application of Teaching Evaluation Based on Association Rules," Proceedings of 2009 4th International Conference on Computer Science & Education, 2009.

[12]. Norwawi, N.M., Hibadullah, C.F., and Osman, J. (2005). "Factors Affecting Performance in Introductory Programming". [CDROM]. In Proceedings of the International Conference on Qualitative Sciences and Its Applications (ICOQSIA).

[13]. Mohamad Mohsin, Md Norwawi, Fazilah Hibadullah, Abd Wahab "Mining the Student Programming Performance using the Rough Set," IEEE, 2010.

[14]. Bakar, A.A.(2005). "Propositional satisfiability method in rough set classification modeling for data mining", PHd Thesis. Universiti Putra Malaysia

[15]. Hong Liu, Yuanyuan Xia,, "Teaching Evaluation System Based on Association Rule Mining," Circuits, Communications and System (PACCS), 2011 Third Pacific-Asia Conference on Page(s): 1 - 3

[16]. Lanfang Lou, Qingxian Pan, Xiuqin Qiu, "New Application of Association Rules in Teaching Evaluation System," International Conference on Computer and Information Application, ICCIA 2010.

[17]. Chandrani Singh, Arpita Gopal, and Santosh Mishra, "Extraction And Analysis Of Faculty Performance Of Management Discipline From Student Feedback Using Clustering And Association Rule Mining Techniques," 3rd International Conference on Electronics Computer Technology (ICECT), 2011 Page(s): 94 - 96.