Student's Academic Performance Analysis by using Apriori Technique

Rekha Thakillapathi[1] G Ananthnath[2]
[1]Student [2]Assistant Professor
[1,2]Department of Computer Applications
[1,2]KMM Institute of PG Studies, Tirupati, India

Abstract— Educational data mining (EDM) is a multi-disciplinary research area that examines statistical modeling, artificial intelligence and data mining with the data generated from an educational institution. In this mainly focus on prediction of student's academic performance. Some statistical tools are also used for predicting student's performance. In this paper we implemented Apriori and k-means clustering algorithm is use for predicting the student's result. Apriori algorithm more efficient and less time over whelming for predicting the student results. The main objective of exploitation Apriori and k-means clustering algorithm is for predicting the students' academic performance and also improving students' academic performance. During this academic student performance is evaluated supported some attributes are selected which generate rules by suggests that of association rule. In this paper we used k-means algorithm for calculating the test anxiety (stress) and time management of the students at the time of examination. It will help to improve the student's tutorial performance. Experiment is conducted usingWeka and real time knowledge set on the market within the college premises , collected from the students and the result which released by the university.

Key words: Data Mining, Association Rule Mining, Apriori Algorithm, Student Dataset

I. INTRODUCTION

In the information technology, huge amount of data is stored from different areas. This huge data exists knowledge. so that it can be utilize in the knowledge discover process. This knowledge discovery process obtained by various techniques such as Decision trees, data mining techniques (classification, clustering, association), Bayesian classifier and so on. In this paper consider association rule and try to optimize this technique by applying modified Apriori and k-means algorithm.Data mining is the process of discovering patterns in large amount data sets involving methods at the intersection of machine learning, , database systems and statistics, . Data mining is an interdisciplinary subfield of statistics and computer science with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use of Data in the data base. The ultimate goal, assuming a large mining is the analysis step for "knowledge discovery in databases" process, or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. The difference between data mining and data analysis is that data analysis is to summarize the history such as analyzing the effectiveness of a marketing campaign, in contrast, data mining focuses on using specific statistical models and machine learning to predict the future and discover the patterns among data. In academic institutions like colleges, universities data increases day by day and storage becomes difficult. Data like student's attendance, internal marks, external marks, his personal details, health details will be coming and this entire data has to be stored in a database. So, order to store data there are data bases in which various tuples are created and data is stored into them. Problem of storing is solved but now a problem arises that, how to accesses the stored data. When we want a particular data to be extracted how to extract it from data base. For this purpose mining of the data is useful. This knowledge discovery process obtained by various techniques such as Decision trees, data mining techniques (classification, clustering, association), Bayesian classifier and so on. In this paper consider association rule and try to optimize this technique by applying modified Apriori and k-means algorithm. The complete description about Data Mining (DM) Association rule technique and Apriori and k-means clustering algorithmic given in next sections. In the concept of Big Data (BD), Data Mining (DM) is a technology by which knowledge Discover (KD) is processed by DM tasks such as Summarization, classification, association and statistical techniques so on. Data Mining is a intersection of machine learning, Artificial Intelligence (AI), database systems. Association Rules was introduced in the year 1993. It is one of the tasks in Data Mining Process. In the Association rule it follows the Apriori algorithm. On the item sets apriori algorithm is used to find out frequent items (Association rule). Association rule is an expression of A=>C, where 'A' and 'C' are the item sets and antecedent, consequent respectively. The meaning of the above expression is, in the Dataset(D) and contains N tuples(assume) and for instance T be the transaction in the data set then C consists high probability when the item A is purchased. This probability is also considered as "confidence". This confidence is represents in the transaction (T) item B be the highest chance when purchase the item A. In the mathematical expression it can be write as p {(C belongs to T)/(A belongs to T)}. One of the best real time examples is present market scenario that is butter have the highest probability to purchase. Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in the large amount databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. rule-based approach also generates new rules as it analyzes more data enough dataset, is to help a machine mimic the human brain's feature extractions and abstract association capabilities from new uncategorized data.

II. LITERATURE SURVEY

In this the concepts of combination of association rules with Apriori and k-means algorithm .apriori algorithm is used for Mining of frequent item sets in transactions database. It is the

core in many tasks of data mining that try to find interesting patterns from datasets, such as association rules, classifier, clustering and correlation, etc. Many algorithms are proposed to find frequent item sets in the database

A. Apriori Algorithm for Association rules

There are several mining algorithms of association rules. One of the most popular algorithms is Apriori algorithm that is used to finding frequent item sets from large database and getting the association rule for discovering the knowledge. Based on this apriori algorithm, this paper indicates the limitation of the original Apriori algorithm of wasting time for scanning the whole data searching in the database for the frequent item sets, and presents an improvement on Apriori algorithm by reducing that wasted time depending on scanning only some transactions. This paper shows by experimental results with several groups of transactions, and with several values of minimum support that applied on the original Apriori algorithm and our implemented improved Apriori that our improved Apriori algorithem reduces the time consumed by 67.38% in comparison with the original Apriori, and makes the Apriori algorithm more efficient and less time consuming.

B. Mining Frequent Item sets Using Apriori Algorithm

Data Mining required data from voluminous Data has been recognized as one of the most challenging problems in data mining approach. In many real world scenarios, the data is not extracted from single data source it will extracted from multiple data sources but from distributed and heterogeneous data sources. Discovered knowledge is expected comprehensive so that can be better fit in the business environment Enterprise data mining applications involve dealing with complex data such as data from multiple heterogeneous data sources, extracting data in single step from such data sources such data sources is space and time consuming. So effective approaches are needed to decrease the sapce as well as time. Here we use Apriori Algorithm for discovering informative patterns in complex data sets from the data base.

C. Efficient enhanced k-means clustering algorithm

In k-means clustering algorithm, we are given a set of n data points in d-dimensional space $\mathbb{R}d$ and an integer k and the problem is to determine a set of k points in $\mathbb{R}d$, called centers, so as to minimize the mean squared distance from each data point to its nearest center point. In this paper, we present a efficient and simple clustering algorithm based on the k-means algorithm, which we call enhanced k-means algorithm. This algorithm is easy to implement, requiring a simple data structure to keep some information in each iterations to be used in the next iteration. Our experimental results demonstrated that our scheme can improve the computational speed of the k-means algorithm by the magnitude in the total number of distance calculations and the overall time of computation.

III. PROPOSED ALGORITHM

The approach begins from the collection of data of graduate students in the university's, and later the pre-processing procedures are tested to the dataset. The data pre-processing Apriori algorithm approach is utilized to make the data more deserved for data mining. Choosing of trait is mostly utilized to minimize the data dimensionality in the datasets. The principle thought of choosing the elements is to acknowledge a subset of info factors by disposing of or ending highlights with no prescient or little data which is important in model improvement and to deliver better execution then the privies models.

A. Algorithm

Apriori algorithm is an algorithm is used to find frequent item set mining and association rule learning over transactional databases. It proceed to identify frequent individual items in the database and extending them to larger items and larger item sets as long as those item sets appear sufficiently often in the database with this we can easily identify frequent itemsets in the database. The frequent item sets determined by using Apriori algorithm can be used to determine association rules which highlight general trends in the database we can reduced search the data in the data base : this has applications in domains such as market basket analysis, etc. Apriori algorithm uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate item sets of length from item sets of length then it extend the candidates which have an infrequent sub pattern. According to the downward closure proposition, the candidate set contains all frequent length item sets. After that, it scans the all transaction database to determine frequent item sets among the candidates. The pseudo code for the algorithm is given below for a transaction database, and support threshold of. Usual set theoretic notation is employed; though note that is a multistep. Is the candidate set for level? At each step, the algorithm is assumed to generate the candidate sets from the large item sets in the database for the preceding level, heeding the downward closure proposition. Accesses a field from data structure that represents candidate set, which is initially assumed to be zero. Many details are omitted below, usually the most important part of the implementation is the data structure used for storing the candidate sets, and counting their frequencies of the data strode in the database.

B. Association rule in Apriori Algorithm:

Association rule mining is usually split into two separate steps:

1) We have to find the minimum support is applied to the all frequent item sets in a database.
2) In this we have to find frequent item sets and the minimum confidence constraint are used to form rules.
3) We want to find a group of items that tend to occur together frequently in item set.
4) The association rules are often written as $X \rightarrow Y$ meaning that whenever X appears Y also tends to appear in the item set. X and Y may be single items or sets of items in a database.
5) Support indicates the frequency of the pattern.

Minimum support is necessary if an association is going to be finding some value.

Support(X) = no. of transactions which contain the item set X / total no. of transactions Confidence denotes the strength of the association between X and Y in the item sets.

Confidence(X →Y) =Support (X∪Y)/Support(X)

1) Support:

The support of an itemsetsx, sup(x) is the proportion of transaction in the database in which the item x appears. Itsignifien the popularity of an item set

$$supp(x) = \frac{\text{Number of transaction in which } X \text{appears}}{\text{Total number of transactions}}$$

The algorithm starts by collecting all the frequent C1-itemsets in the first pass based on the minimum support. It uses this set called L1 to generate the candidate sets to be frequent in the next pass called C2 by combing L1 with itself. Any item that is in C1 and not in L1 is eliminated from C2itemsets. This is achieved by call in a function called 'apriori-gen'. It reduces the item size drastically. The algorithm continues in the same way to generate the Ck, where k is size of the large item sets of k-1, then reduces the candidate set by eliminating all those items in k-1 with support count is less than minimum support. The algorithm terminates when there are no candidates to be counted in the next pass.

C. Data Collection and Preparation

In our study, we have considered student's data that are pursuing Master of Computer Application (MCA) degree from University. On the basis of the data collected some attributes have been considered to predict student's performance in university examination. The Variables used for the prediction students' performance in university results are Graduation%, Attendance%, Assignment%, UnitTest%,pass/fail, seminar grades, and UniversityResult%.

Attributes	Description	Values
Graduation%	Percentage of marks obtained in graduation.	Good, Avg, Poor
Attendance	Attendance of the student.	Good, Avg, Poor
Assignment	Assignment performance given during the semester.	Good, Avg, Poor
UnitTest Performance	Percentage marks obtained by a student in Unit Test.	Good, Avg, Poor
University Result	Percentage marks obtained by the student in university examination.	Good, Avg, Poor
Seminar	Grades obtained in seminars	A,B,C,D
Result	Pass /fail in university exams	Pass=P,Fail=F

Table 1: Attributes and Its Possible Values

IV. Results and Discussion

The dataset of 70 students from MCA course was obtained from M.C.A department of kmmips, svu University. In this paper we find various association rules between attributes like students graduation percentage, Attendance, Assignments, Unit test marks, seminor marks and how these attributes affect the student's university result. Number of association rule can be found for different confidence values on the the dataset. The analysis for generating association rules is as follows:

A. The rules generated for 90% confidence and 0.1 supports are:

1) Attendance%=Good Assignment%=Poor ==> UnitTest%=Poor<conf :(1)> lift :(1.28) lev :(0.02) conv :(1.3)
2) Attendance%=GoodAssignment%=Poor==>University Result=Poor<conf :(1)> lift :(1.76) lev :(0.04) conv :(2.6)

B. Rules for confidence 85% confidence and 0.1 supports are:

1) Attendance%=Poor Assignment%=Good==> UnitTest%= Poor UniversityResult%=Poor <conf :(0.86)> lift :(1.9) lev :(0.05) conv :(1.93)
2) Assignment= Poor ==> UnitTest= Poor UniversityResult= Poor <conf :(0.82)> lift :(1.82) lev :(0.07) conv :(2.02)

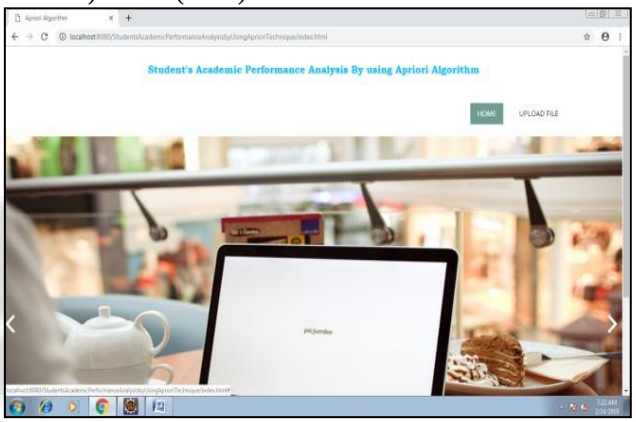

Fig. 1: Index Page

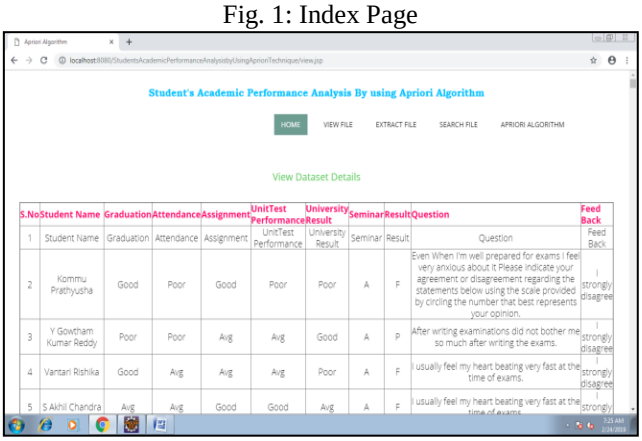

Fig. 2: View File

Above fig2 will shows view of itemsets in the data base. With the help of this screen we view all files in the data base.

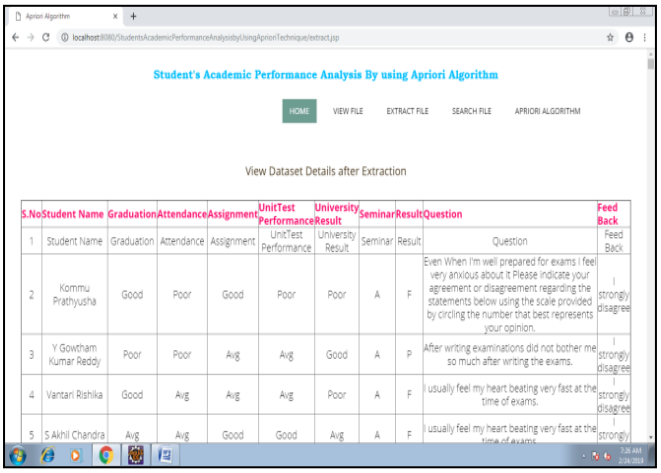

Fig. 3: Extract File

Above fig3 will shows the extraction of files from the data base.with the help of this window we can easily extract file from large amount data set in the database

Fig. 4: Search File

Above fig4 will shows searching the files from the database.

With the help of this window we can easily search file from the large amount datasets.

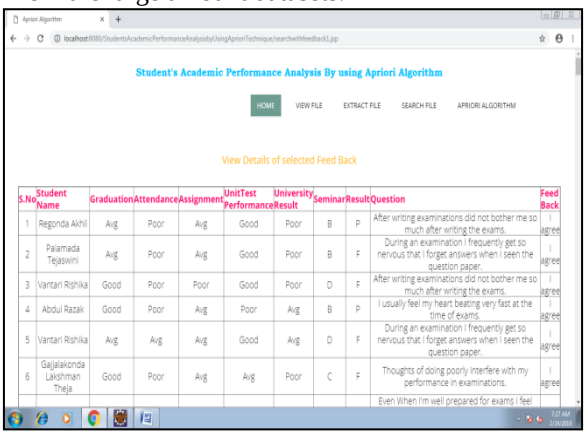

Fig. 5: Search File Output

Above fig5 will shows the search files output.we can easily find the output of the search files from the database.

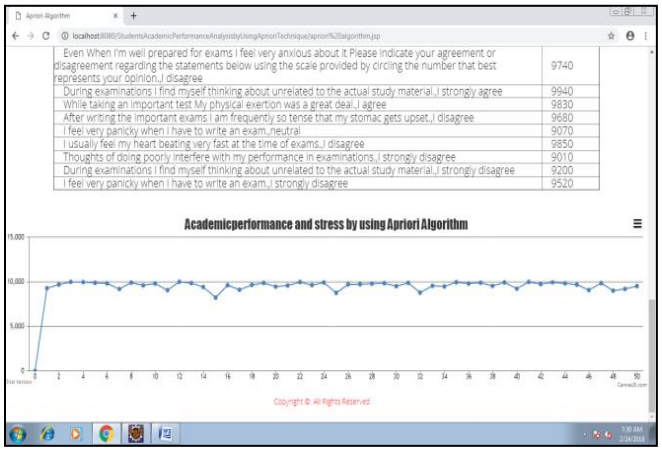

Fig. 6: Graph Page

Above fig6 will shows the graphical representation of the data in the database.with this graphical representation we can easily analysis data in the data sets.

V. CONCLUSION

The paper presented the potential use of education data mining using association rule algorithm in enhance the quality and predicting students' performances in university result. The analysis revealed that student's university performance is dependent on Unit test, Assignment, Attendance, seminar grades and graduation share. The results reveal that the student's performance level can be improved in university result by identifying students who are poor unit check, Attendance, Assignment , seminar and graduation and giving them additional guidance to the students to improve their university result. in this we conjointly find stress of the scholars at the time of examinations by conducting some surveys. By this we can reduce stress on the students by giving some guidance's.

REFERENCES

[1] Sonali Agarwal, G. N. Pandey and M. D. Tiwari, "Data Mining in Education: Data Classification and Decision Tree Approach", International Journal of e Education, e-Business, e-Management and e Vol. 2, No. 2, April 2012.

[2] Mounika Goyal and RajanVokra, "Applications of Data Mining in Higher Education", International Journal on Computer Science Issues (IJCSI), ISSN (Online): 1694-0814 , Vol 9, Issue 2, No 1, March 20

[3] M.R. Thanasekhar and N. Balaji, "Performance Analysis of Students using Data Mining techniques", International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET), ISSN (Online) : 2319 – 8753, Volume 3, Special Issue 3, March 2014..

[4] AjinkyaKunjir, Poonam Pardeshi, Karan Naik, "Recommendation of Data Mining Techniques in Higher Education", International Journal of Computational Engineering and Research (IJCER), Volume 5, Issue 3, March 2015.

[5] NaeimehDelavari, "Application of Enhanced Analysis Model for Data Mining Processes in Higher Educational System ", 2005 IEEE July 7 Juan Dolio, Dominican Republic ITHET 6th Annual International Conference.

[6] Pooja Thakar, Anil Mehta , "Performance analysis and Prediction in Educational Data Mining", International Journal of Computer Applications (IJCA) , Volume 110, Issue 5, January 2015.
[7] CeaserVialardi, Javier Bravo, Alvaro Ortigosa, "Recommendation in Higher Education Using Data Mining Techniques ", educational data mining, 2009.
[8] Q. A. AI-Radaideh, E. W. AI-Shawakfa, and M. I. AI-Najjar, "Mining student data using decision trees", International Arab Conference on Infonnation Technology (ACIT'2006), Yannouk University, Jordan, 2006.
[9] Surjeetkumar& Saurabh Pal, "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology Journal (WCSIT) Volume 2, No-2, pp 51-56, 2012.
[10] Nguyen, Nguyen T., Paul Janecek, and Peter Haddawy "A Comparative Analysis of Techniques for Predicting Academic Performance." In Proceedings of the 37th ASEE/IEEE Frontiers in Education Conference.pp. 7-12, 2007.

