ANUSANDHAN- AISECT University Journal
Vol. 06, Issue No. 13, March-2018, P-ISSN 2278-4187, E-ISSN 2457-0656

Improved Rule Discovery Using FP Growth Algorithm in Educational Data Mining

Rajeev Sharma[1]* Dr. Sitendra Tamrakar[2]

[1]Asst. Prof., Dept of CSE, MIMT, Gwalior (M.P.) India

[2]Dept of CSE, CCET, Durg (C.G.) India

Abstract – In the educational data, the secret information can be extorted from the great quantity of data. This data can be used for the development in the educational system and making it high class education. The methods of data mining are useful for the extraction of the reality of the educational system. FP Growth is a regular unremitting continuous example mining, which produces visit itemset without candidate innovation. It impacts usage of tree based to structure. The learning methodologies of the academic for particular university with their related information are used. Each student's performance is evaluated from the database and it should be that reliable to resist the changes in the academic record. After that we have changed over the general arrangement into a changed relationship for appropriateness of utilizing FP Growth. In this paper, the useful rules are generated for showing the relationship among different number of attributes. FP Growth used in this paper for the appealing rules generation and extort the efficient rules. These rules are helpful to show the achievement of the each and every student. The teaching methodologies are taken into consideration to achieve the better academic of the students and it can be generated by taking different parameters.

Index Terms—Data Mining, ARM, EDM, FP Growth, etc.

----------♦----------

I. INTRODUCTION

Data mining is a technique of extricating interesting taking in or designs from large databases. There are a few procedures that have been utilized to find such sort of knowledge, the vast majority of them coming about because of machine learning and measurements. Despite the fact that this learning might be useless if it doesn't offer some sort of surprisingness to the end user. The assignments achieved in the data mining depend upon what kind of know-how someone needs to mine (Jain, et. al., 2013). The most essential sorts of undertaking done by methods for DM strategies are Classification, Dependence Modeling, Clustering, Regression, Prediction and Association. Clustering task search for the information able of compute the cost of a formerly depicted reason trademark in light of different qualities and is regularly frequently addressed by IF-THEN rules.

II. ASSOCIATION RULE MINING

ARM, a standout among the most basic and pleasantly researched systems of data mining. It centers to isolate intriguing associations, frequent patterns, associations or casual structures among sets of items in the transaction databases or differing data vaults. Let $I=I_1, I_2, \ldots, I_m$ be a course of action of m specific properties, T be transaction that combines a settled of contraptions to such a degree, to the point that $T \subseteq I$, D be a database with particular transaction records. An association rule is a proposal as $X \Rightarrow Y$, where X, $Y \subseteq I$ are sets of items called itemsets, and $X \cap Y = \emptyset$. Here, X is called forerunner while Y is recommended as coming about, the manage strategy X infers Y (Mihir & Dabhi, 2015). To choose interesting statutes from the arrangement of every single conceivable manage, necessities on different measures of essentialness and intrigue can be used. The best-known goals are insignificant edges on support and confidence.

Support (X=Y) = XUY. Count/n

Confidence of an association rule is portrayed as the proportion/fraction of the wide blend of transactions that join XUY to the entire assortment of bits of knowledge that consolidate X, where if the rate outperforms the edge of self belief a stimulating association rule $X \Rightarrow Y$ can be delivered.

Confidence (X=Y) = Support (XUY) /Support (X)

Drawbacks and solutions

In the ARM zone, the vast majority of the examination endeavors went inside the first area to enhancing the algorithmic general execution (Ceglar & Roddick, 2006) and inside the second region into diminishing the output set by techniques for empowering the chance to express essentials on the pined for results. Over the earlier decade spreads of computations that change in accordance with those issues through the refinement of look for methodologies, pruning strategies and data frameworks had been delivered (Goethals, et. al., 2006).

III. EDUCATIONAL DATA MINING

EDM is rising as an examination zone with a suite of computational and mental systems and research approaches for rising how students learn. EDM in data mining recoup secured learning by applying arranged arrangement of data mining like clustering, rule mining, web based mining, test mining, neural network, baysian network, and distinctive others which gives us a last result and in the event that it require some need to changes clearly the raw data if isolated by require (Kumar, 2015).

Goals of EDM: FP Growth (Borgelt, 2003) is another basic frequent pattern basic FPM strategy, which produces visit itemset without candidate age. It uses tree based structure. The issue of Apriori algorithm was overseen, by displaying a novel, minimal data structure, called frequent pattern tree, or FP by then in light of this structure a FP pattern fragment growth technique was made (Borgelt, 2003).

FP tree is worked in two passes:

(i) Pass 1:

Scan data and check support for everything

Discard infrequent items

Sort visit things in diving demand in light of their support.

(ii) Pass 2:

Reads one exchange at any given minute and maps it to the tree.

Fixed organize is utilized with the target that way can be shared

Pointers are kept up between nodes containing same items.

Frequent things are extracted from the rundown It experiences certain.

Advantage: Preference of FP-Growth is that the developing FP Tree has awesome execution of compression, and its strategy of mining can decrease the cost of rescanning data. Also, it applies contingent FP-Tree on abstaining from producing applicant thing and testing analysing process.

Disadvantages:

Fp tree may not fit in vital memory.

Execution time is gigantic as a result of complex reduced data structure (Kulkarni & Khonde, 2017).

IV. LITERATURE SURVEY

Prajakta G. Kulkarni et al. [2017] in this paper a new scheme or algorithm is proposed that will reduce the execution time for the massive database and works efficiently on number of nodes by using Modified Apriori algorithm.

Ying Gao et al. [2017] In this paper, based on the development and application of on-board subsystem test bench for current CTCS-3 system, this paper focuses on the approach of naturally age of test sequence, takes the existing test sequences of ETCS-2 (European Train Control system level 2) as the train set existing relatively mature test sequence as the training set, to execute ARM. The whole data mining process involves data preparation (including data cleaning and data selection) firstly, providing basement for association rule, then establishes FP tree and seeks test cases with frequent pattern through implementing FP growth algorithm for the target database. Comparing the analysis results and experience, it shows that the association rule based on FP tree could play an important role on the efficiency and verification of consequently generation of test sequence.

Jih-Jeng Huang et al. [2017] In this paper, our propose an integrated framework to combine centralized calculations, for example, the FP Growth, Sequential Pattern Discovery using Equivalence classes (SPADE), and rough set algorithms, to mine decision rules in a distributed environment. In addition, our method finds some significant rules that other algorithms cannot. The experiments also demonstrate that the proposed method is well suited to finding association and sequential rules in a distributed environment.

Vaishali Patil et al. [2016] in this case, every site is interested in globally supported association rules without revealing its own local information. To fulfill this goal, we use a secure multi-party algorithm based on secure sum technique to simplify the operation of mining association rule when the database is on a horizontally apportioned among different sites. We are using a Frequent-Pattern (FP) growth algorithm to

find frequent itemsets and try to reduce total computation time.

Wenchuan Yang et al. [2016] this paper proposes an incremental line calculation styles in light of association rules, which is the progressed FP4W-Growth calculation.

Hong-Yi Chang et al. [2016] In this paper, our propose a novel incremental data mining algorithm in light of FP-Growth, the utilization of stack tree to deal with the bother of incremental refreshing of common item sets (Chang, et. al., 2016)

V. PROPOSED METHODOLOGY

The database of each institute contains private and educational information of the students. For performing some operation over that data, we have to collect that data and apply operations on it. Different categories of operations used distinctive data so there are various techniques available to development the data. Initially pre-processing performed on the academic records by taking courses data and some other attributes like Hall Status, Retention, Abandonment and etc are in use for the better mining of the rules.

In the starting level, we performed the processing on the courses of the institution. As distinctive things have been attempted with the student "Data of the branch of Computer Science and Engineering in BUET, we have analyzed all aides inside the curriculum which must be taken to finish the BSc degree. A scholar has to take 68 departmental publications and non-departmental courses in overall. Among them there are 40 theory courses (25 departmental and 15 non-departmental) and 28 sessional distributions (20 departmental and 7 non-departmental). We determine academic performance and impact of other factors on the basis of these course's last grades, indications of attendance, class tests, and term last answer substance add up to marks and so on.

Mainly the database contains personal information of the students and this database is known as universal database. This table also includes the academic record of the student. The particular course contains some grades and the number of attendance to calculate the efficiency of the student. Hall status, Student id, Gender, Grades and attendance are stored in this database. Now the data transformation performed by converting it from continuous to discrete form to increase the reliability of the data. The transformation of CGPA into different classifications such as poor, average, good, very good and excellent. Likewise, all the attributes are converted into discrete form.

Proposed Algorithm:

Start the process

(ii) Open the dataset and taken it as input

(iii) Choose Nominal to Binary from pre-process

(iv) Apply FP Growth in the preprocessed data

(v) Database examine performed to decide the support of each object, discard the rare objects and type the frequent items in lowering order

(vi) Scan the data set one exchange at an opportunity to make the FP-tree. For every transaction:

If it's miles a unique transaction shapes a new path and set the counter for each node to at least one.

If it shares a commonplace prefix itemset then increment the common itemset node counters and create new nodes if wished

(vii) Continue this till every transaction has been mapped unto the tree

(viii) Stop the process

VI. RESULT ANALYSIS

WEKA Explorer has used in this paper for the simulation of the proposed work and performed the implementation of the different number of parameters.

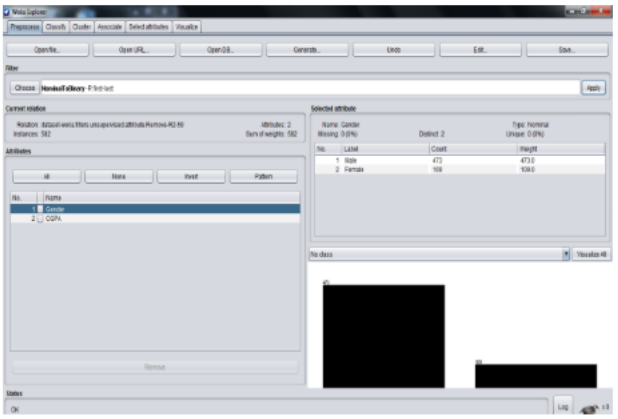

Fig.1 WEKA Explorer for Input File

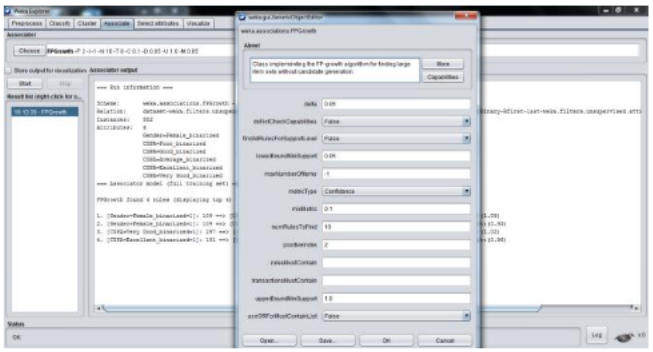

Fig. 2 Associate tab of FP Growth

In the result section, the impact of various parameters is shown with their generated rules. Different rules are generated in this section by considering minimum support and confidence value.

The following impacts are shown below:

Influence of Gender

The influence of gender has been originate in the general academic execution. The country's situations might show the sign of socio economist. Mainly there are male candidates who live in the residence of the university (BUET). There are different elements that influence the educational environment and college student's academic achievement. The female students of the university have high chance of achieving good CGPA which is demonstrated in the table below. The reason behind the good academic records is mostly influenced by usual societal issues of the country. Hence the male candidates are inferior to female candidates in the overall performance of the academic. In the table below, we used "G" for Gender, "M" for Male and "F" for Female.

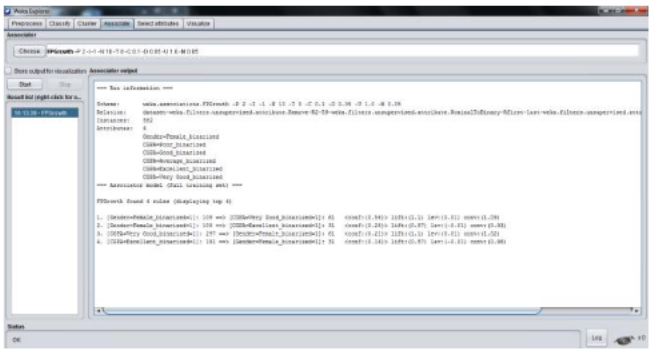

Fig.3 Rules for Gender

S.No.	Useful Rules Produced	Min. Support	Confidence
1.	G=F ==> CGPA=Very Good	5%	56%
2.	G=F ==> CGPA=Excellent	5%	28%
3.	CGPA=Very Good ==> G=F	5%	21%
4.	CGPA=Excellent ==> G=F	5%	16%

(b) Influence of Residence

There are two types of students in the university such as student live at their home and at institution's hall. In the university, large amount of students live in the hall of the institution. Both types at students are taken into consideration for calculating the impact of the residence. From the table below, it can be demonstrated that the non-resident students performed better than resident students. Non-resident student can concentrate more in their academic and get much better CGPA. This can be understood that student can achieve better score to provide the full concentration in their study.

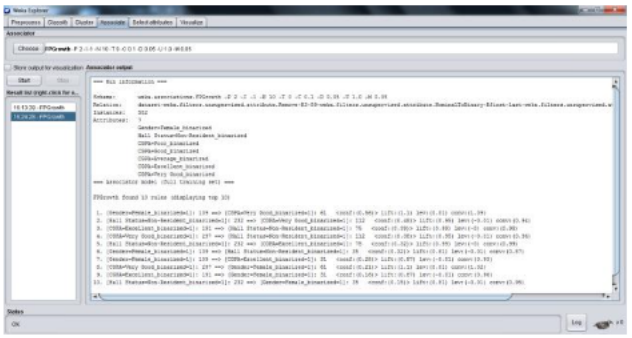

Fig.4 Rules for Residence

S.No.	Useful Rules Produced	Min. Support	Confidence
1.	G=F ==> CGPA=Very Good	5%	56%
2.	Hall Status=Non-Resident ==> CGPA=Very Good	5%	48%
3.	CGPA=Excellent ==> Hall Status=Non-Resident	5%	39%
4.	CGPA=Very Good ==> Hall Status=Non-Resident	5%	38%
5.	Hall Status=Non-Resident ==> CGPA=Excellent	5%	32%
6.	G=F ==> Hall Status=Non-Resident	5%	32%
7.	G=F ==> CGPA=Excellent	5%	28%
8.	CGPA=Very Good ==> G=F	5%	21%
9.	CGPA=Excellent ==> G=F	5%	16%
10.	Hall Status=Non-Resident ==> G=F	5%	15%

(c) Correlation between Courses

From this section, it can be show that the different courses affect the performance of other subjects. From the 1st rule, if student in CSE801 subject get poor grade then also get poor grade in CSE601 subject. From the last rule, if student in CSE701 subject get poor grade then also get poor grade in HUM275 subject.

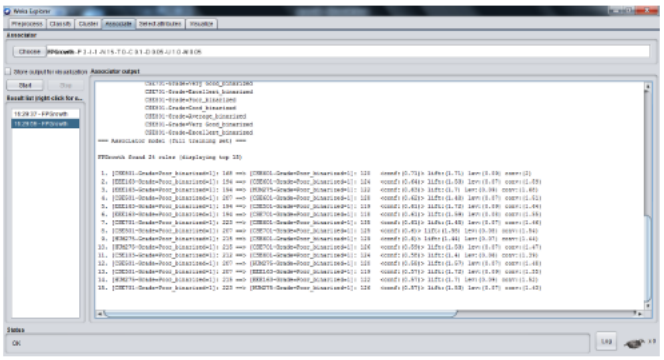

Fig.5 Rules for Courses

S. No.	Useful Rules Produced		Min. Support	Confidence
1.	CSE801-Grade=Poor CSE601-Grade=Poor	==>	5%	71%
2.	EEE163-Grade=Poor CSE601-Grade=Poor	==>	5%	64%
3.	EEE163-Grade=Poor HUM275-Grade=Poor	==>	5%	63%
4.	CSE501-Grade=Poor CSE601-Grade=Poor	==>	5%	62%
5.	EEE163-Grade=Poor CSE501-Grade=Poor	==>	5%	61%
6.	EEE163-Grade=Poor CSE701-Grade=Poor	==>	5%	61%
7.	CSE701-Grade=Poor CSE601-Grade=Poor	==>	5%	61%
8.	CSE501-Grade=Poor CSE701-Grade=Poor	==>	5%	60%
9.	HUM275-Grade=Poor CSE601-Grade=Poor	==>	5%	60%
10.	HUM275-Grade=Poor CSE701-Grade=Poor	==>	5%	59%

(d) Influence on Retention

Retention is the term which is useful to extract the students who is not able to pass the exam and have to attend that course again.

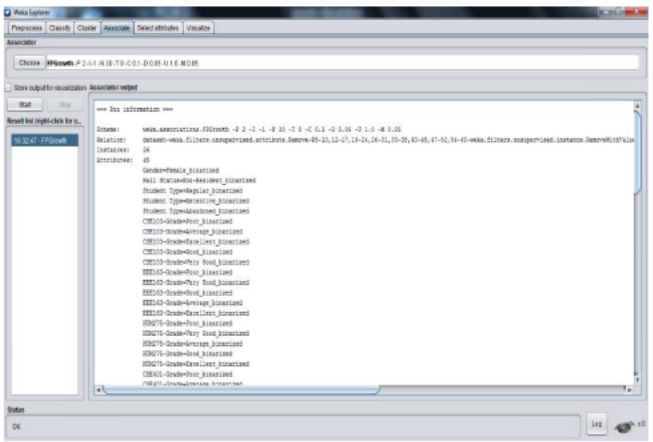

Fig.6 Rules for Retention

From the table below, it can be demonstrate that the retentive students not perform well in the exam and achieve poor grades in the exams. All the rules below show the grades of retentive students with the subjects.

S. No.	Useful Rules Produced		Min. Support	Confidence
1.	HUM275-Grade=Poor Student Type=Retentive	==>	5%	100%
2.	EEE163-Grade=Poor Student Type=Retentive	==>	5%	100%
3.	CSE701-Grade=Poor Student Type=Retentive	==>	5%	100%
4.	CSE601-Grade=Poor Student Type=Retentive	==>	5%	100%
5.	CSE103-Grade=Poor Student Type=Retentive	==>	5%	100%
6.	HUM275-Grade=Poor EEE163-Grade=Poor	==>	5%	100%
7.	EEE163-Grade=Poor HUM275-Grade=Poor	==>	5%	100%
8.	HUM275-Grade=Poor CSE103-Grade=Poor	==>	5%	100%

(e) Influence on Abandonment

Abandoned students are that who does not complete their studies and leave it in the middle of the course is known as abandoned students. From the table below, it can be show that female students are more likely is abandoned. Mainly non-resident students leave their studies without completing their course.

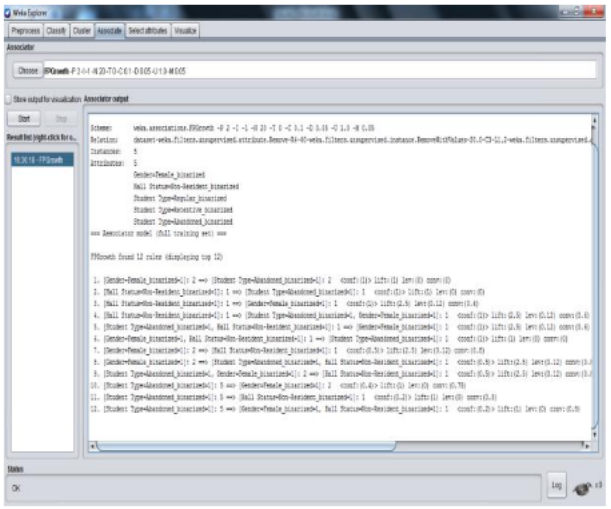

Fig.7 Rules for Abandonment

S. No.	Useful Rules Produced	Min. Support	Confidence
1.	G=F ==> Student Type=Abandoned	5%	100%
2.	Hall Status=Non-Resident ==> Student Type=Abandoned	5%	100%
3.	Hall Status=Non-Resident ==> G=F	5%	100%
4.	Hall Status=Non-Resident ==> Student Type=Abandoned, G=F	5%	100%
5.	Student Type=Abandoned, Hall Status=Non-Resident ==> G=F	5%	100%
6.	G=F, Hall Status=Non-Resident ==> Student Type=Abandoned	5%	100%
7.	G=F ==> Hall Status=Non-Resident	5%	50%
8.	Student Type=Abandoned ==> G=F	5%	40%

(f) Influence of Departmental Courses

In any institution, there are many departmental courses which are responsible for final CGPA. Grade of one subject affects the performance of other subjects. If the grade of one course is good or average then other courses also achieve better grades and vice-versa.

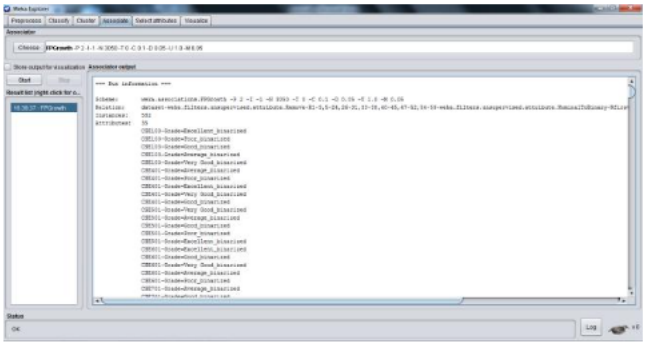

Fig.8 Rules for Departmental Courses

S. No.	Useful Rules Produced	Min. Support	Confidence
1.	CSE801-Grade=Good, CSE501-Grade=Very Good ==> CGPA=Excellent	5%	100%
2.	CSE801-Grade=Good, CSE601-Grade=Excellent ==> CGPA=Excellent	5%	100%
3.	CSE103-Grade=Average, CSE501-Grade=Very Good ==> CGPA=Excellent	5%	97%
4.	CSE501-Grade=Excellent ==> CGPA=Excellent	5%	94%
5.	CSE601-Grade=Very Good ==> CGPA=Excellent	5%	92%
6.	CSE701-Grade=Average, CSE801-Grade=Very Good ==> CGPA=Excellent	5%	90%
7.	CSE103-Grade=Average, CSE801-Grade=Average ==> CGPA=Excellent	5%	88%
8.	CSE401-Grade=Good, CSE701-Grade=Good ==> CGPA=Excellent	5%	85%
9.	CSE401-Grade=Good, CSE103-Grade=Average ==> CGPA=Excellent	5%	81%
10.	CSE501-Grade=Average ==> CGPA=Excellent	5%	75%

(g) Influence of Continuous Assessment

The evaluating of a way relies upon different components together with signs of participation, class test, and two areas of term last examination.

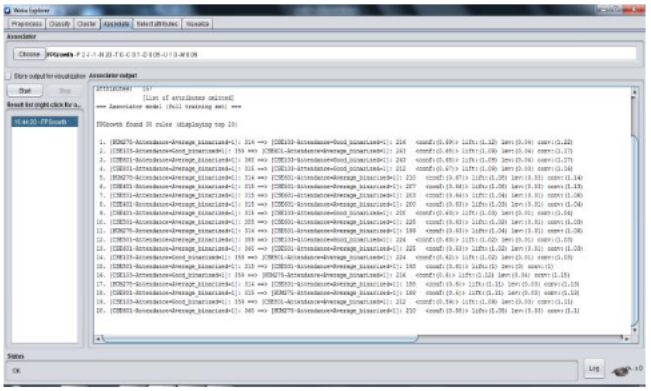

Fig.9 Rules for Continuous Assessment

In the table, there should be good attendance for each subject and if student attend class of one course then there is more chance to attend the class of other courses.

S. No.	Useful Rules Produced	Min. Support	Confidence
1.	HUM275-Attendance=Average ==> CSE103-Attendance=Good	5%	69%
2.	CSE103-Attendance=Good ==> CSE601-Attendance=Average	5%	68%
3.	CSE601-Attendance=Average ==> CSE103-Attendance=Good	5%	68%
4.	CSE801-Attendance=Average ==> CSE103-Attendance=Good	5%	67%
5.	HUM275-Attendance=Average ==> CSE601-Attendance=Average	5%	67%
6.	CSE401-Attendance=Average ==> CSE501-Attendance=Average	5%	66%
7.	CSE801-Attendance=Average ==> CSE601-Attendance=Average	5%	64%
8.	CSE401-Attendance=Average ==> CSE601-Attendance=Average	5%	63%
9.	CSE401-Attendance=Average ==> CSE103-Attendance=Good	5%	63%
10.	CSE601-Attendance=Average ==> CSE501-Attendance=Average	5%	63%

(h) Influence of Non-Departmental Courses

The generated rules are considered for showing the influence of non-departmental courses with their corresponding grades. The concluding result of particular session is based on the accomplishment of the non-departmental courses. But these courses have less influence on the last result and provide the option to improve the achievement of the academic

record. From the generated rules, it can be shown that these subjects have less influence over the final grades. Rule 1 demonstrates that subject EEE163 has average grade and final CGPA is also very good. But from rule 8, it can be shown that subject HUM275 has poor grade instead of this the final CGPA is also very good.

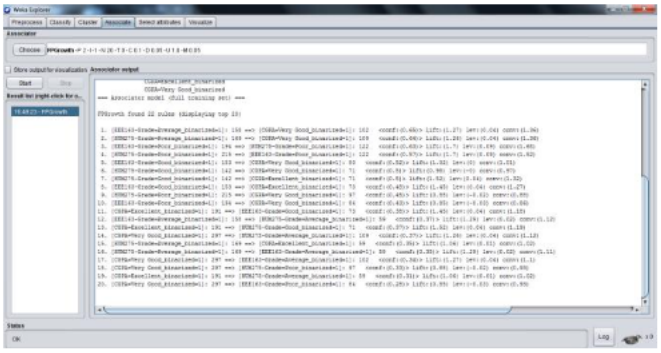

Fig.10 Rules for Non-Departmental Courses

S. No.	Useful Rules Produced	Min. Support	Confidence
1.	EEE163-Grade=Average ==> CGPA=Very Good	5%	65%
2.	HUM275-Grade=Average ==> CGPA=Very Good	5%	64%
3.	EEE163-Grade=Poor ==> HUM275-Grade=Poor	5%	63%
4.	HUM275-Grade=Poor ==> EEE163-Grade=Poor	5%	57%
5.	EEE163-Grade=Good ==> CGPA=Very Good	5%	52%
6.	HUM275-Grade=Good ==> CGPA=Very Good	5%	50%
7.	EEE163-Grade=Good ==> CGPA=Excellent	5%	48%
8.	HUM275-Grade=Poor ==> CGPA=Very Good	5%	45%
9.	EEE163-Grade=Poor ==> CGPA=Very Good	5%	43%
10.	CGPA=Excellent ==> EEE163-Grade=Good	5%	38%

CONCLUSION

The academic achievement of particular student is very vital for each institution to improve the record of their institution. In general manner it can be shown that the database should be used accurately to perform the detailed study. The brightest students can also reduce the performance and the reason can be detected from the academic record. Retention and abandonment are two main causes for the performance of the students. So the data mining techniques are used to mined the useful data and generate the most influential rules to detect the main causes. FP Growth is used in this paper for the generation of rules over the pre-processed data and makes it more useful for the institution.

REFERENCES

Agrawal R. and Srikant R. (1994). "Fast algorithms for mining association rules". In Proc. Int'l Conf. Very Large Data Bases (VLDB), pages 487–499, Sept. 1994.

C. Borgelt (2003). "Efficient Implementations of Apriori and Eclat". In Proc. 1st IEEE ICDM Workshop on Frequent Item Set Mining Implementations, CEUR Workshop Proceedings 90, Aachen, Germany 2003.

Ceglar, A., Roddick, J.F. (2006). Association mining. ACM Computing Surveys, 38:2, pp. 1-42.

Goethals B., Nijssen S., Zaki, M. (2006). Open source data mining: workshop report. SIGKDD Explorations, 7:2, pp. 143-144.

Han J., Pei H., and Yin. Y. (2000). Mining Frequent Patterns without Candidate Generation, In Proc. Conf. on the Management of Data.

Hong-Yi Chang, Jia-Chi Lin, Mei-Li Cheng, Shih-Chang Huang (2016). "A Novel Incremental Data Mining Algorithm based on FP-Growth for Big Data" 2016 International Conference on Networking and Network Applications.

http://en.wikipedia.org/wiki/Educational_data _mining.

Jeetesh Kumar Jain, Nirupama Tiwari, Manoj Ramaiya (2013). "A Survey: On Association Rule Mining" International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 3, Issue 1, January -February 2013, pp.2065-2069.

Jih-Jeng Huang (2017). "An Integrated Method for Mining Association and Sequential Rules in Distributed Databases" IFSA-SCIS 2017, Otsu, Shiga, Japan, June 27-30, 2017.

Kumar, J. (2015). "A Comprehensive Study of educational Data mining". IJEECSE.

Mihir R Patel, Dipak Dabhi (2015). "An Extensive Survey on Association Rule Mining Algorithms" International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 1, January 2015).

Prajakta G. Kulkarni, Prof. Shraddha R. Khonde (2017). "HDFS Framework for Efficient

Frequent Itemset Mining Using MapReduce" 978-1-5090-4264-7/17/$31.00 ©2017 IEEE.

Vaishali Patil, Ramesh Vasappanavara, Tushar Ghorpade (2016). "Securing association rule mining with FP growth algorithm in horizontally partitioned database" 2016 International Conference on Control, Computing, Communication and Materials (ICCCCM).

Wenchuan Yang , Lei Hui, Dong Zhang 3 and Yimin F. (2016). "An Improved Incremental Queue Association Rules for Mining Mass Text" 2016 International Symposium on Computer, Consumer and Control.

Ying Gao, Qi Zhang, Lijie Chen, Kaifeng Wang, Ningning Chen, Hongjie Liu (2017). "Research on application of FP_tree based association rule mining on test sequence in train control system" Proceedings of the 36th Chinese Control Conference July 26-28, 2017.

Corresponding Author

Rajeev Sharma*

Asst. Prof., Dept of CSE, MIMT, Gwalior (M.P.) India.

E-Mail – sharmaraj2007@gmail.com