Study of Education Patterns in Rural and Urban India using Association Rule Mining : Implementation

Sk Althaf Hussain Basha*
Professor and Head, Department of MCA, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana , India

ABSTRACT

During past two decades, several statistical techniques/methods student analysis were used. Academic accuracy from different perspectives. Over a time period availability of education within the rustic (rural) areas has improved. Further, the developments in urban areas in different sectors have resulted in educational environment changes. Our pursuit in this paper has been to find the strong rules from the available data by applying association rule mining, and there by find the relevance to the student performance associated with the educational environment in which they study. We have identified the association between different attributes of educational environment i.e., the location of the college, type of the college, different social groups, different courses etc., and thereby extract strong association rules. The processed the available data has found the unknown rules and analysis of these rules offering a suitable and build testimonial academic planning's within higher institutions learnings to heighten their deciding process .They are also helpful for a proper understanding of the educational environment aids to the course of study construction and other enhancements for readily rising students theoretical performance. Through this document we use data mining technique of association rule mining to extract strong rules in education environment that identifies students' success patterns in different colleges in different social groups and also presents, implementation is done using Java Programing and Oracle Software, Further we have processed the available data to find the pattern of support for these rules from time to time.

Key Words: Academic Student Performance, Data Mining, Higher Education, Association Mining Apriori algorithm

I. INTRODUCTION

In the last few years, in accordance with the fast developed technology the total data has been developing technology the total data has been developing large in every field. The discovery of novel and the required information from entire sum of data has been increased. The usage of data mining with different mining techniques used on various application domains such as education, banking, retail sales, bio-informatics and telecommunications and etc., In order to take out useful data to fulfill the requirements of the industry. The large amount of data is stored in databases, files and other repositories, growth fully it is important, if not required, modifying powerful means to analyze and perhaps interpreting data for the extraction of interesting knowledge and it will be helping in making decisions, It is necessary to gain important information i.e., already not known from these data by applying data mining techniques.

In this continuously changing environment, the requirement is very high for the educated work force. During these days, the foremost aim is to power up the universities and educational processes.

For the most available data, data mining considered as the most relevant technology in providing additional details to the lecturer, student, alumni, manager and the remaining educational staff and acting as an active automated assistant in helping to make better decisions in respective educational activities.

The Graduates education is one of the important part of national education system, it's the backbone of national level competitiveness and innovation. The man intensity of higher educational system is to purify the quality of education. Usage of data mining technology will be helping to fill gaps in higher education. The patterns, association and anomalies, can be used to improve the speed of the processes. These improvements will be bringing lots of advantages to the higher educational system such as increasing educational system efficiency, in all the required students statistics (promotion rate, retention, drop out, success, learning…etc) and it will be reducing the cost of system processes. To improve the above quality, we need a data mining system that will provide needed knowledge and internal views for the decision makers in the higher educational system.

In this paper we are using association rule mining to bring out the strong rules hidden in the data. Apriori algorithm is the most preferred one for his purpose. Here we have used seven years of Under Graduate data of Kakatiya University, Warangal, Telangana, India from 2000 to 2006. The data has been preprocessed to suit the needs of our mining activity and we have done the implementation by Java Programing and Oracle Software.

In this document, we are using association rule mining to get the strong rules which are hidden in the data. Apriori is one of the most suitable algorithm for this purpose. Here, we have used seven years of under graduate data of kakatiya university, Warangal, Telangana, India during 2000-2006. The entire data will be preprocessed to suit the needs of our mining activity and we will be implementing through Java Programming and Oracle software.

II. RELATED WORK

Association Mining Problems originates from the research of Market Basket problems, where the Association Mining in Educational domain has been studied by many researchers. In Minaei-Bidgoli et. Al. proposed an approach to classify students in order to predict their final year grade based on the features extracted from logged data in an educational web-based system was reported. Data mining classification process was used in conjunction with genetic algorithm to improve the prediction accuracy [1]. Also, Talavera et. al in student data was mined to characterize similar behavior groups in unstructured collaboration using clustering algorithms[2]. The relationship between students' university entrance examination results and their success was studied using cluster analysis and k-means algorithm techniques in [3]. Fuzzy logic concept was not behind in the field of educational data mining [4,5,6,7], for instance a two-phase fuzzy mining and leaning algorithm was described in [8], this is an hybrid system of association rule mining Apriori algorithm with fuzzy set theory and inductive learning algorithm to find embedded information that could be fed back to teachers for refining or reorganizing the teaching materials and test. Association rule mining technique has also been used in several occasions in solving educational problems and to perform crucial analysis in the educational environment. This is to enhance educational standards and management such as investigating the areas of learning recommendation systems, learning material organization, student assessments, course adaptation to the students' behavior and evaluation of educational web sites [9,10,11,12 13]. In [11] a Test Result Feedback (TRF) model that analyses the relationships between students' learning time and the corresponding test results was introduced. Knowledge Discovery through Data Visualization of

Drive Test Data was carried out in [14]. Genetic algorithm as Ai technique was for data quality mining in [15] Association rule mining was used to mining spartial Gene Expressing [16] and to discover patterns from student online course usage in [13] and it is reported that the discovered patterns from student online course usage can be used for the refinement of online course. Robertas, in [17] analysed student academic results for informatics course improvement, rank course topics following their importance for final course marks based on the strength of the association rules and proposed which specific course topic should be improved to achieve higher student learning effectiveness and progress.

Different types of rule-based systems have been applied to predict student academic performance (mark prediction) in an e-learning environment systems (using fuzzy association rules) [18]; to predict learner performance based on the learning portfolios compiled (using key formative assessment rules) [19]; for prediction, monitoring and evaluation of student academic performance (using rule induction) [20]; to predict final grades based on features extracted from logged data in an education web-based system (using genetic algorithm to find association rules) [21]; to predict student grades in LMSs (using grammar guided genetic programming) [22]; to predict student performance and provide timely lessons in web-based e learning systems (using decision tree) [23]; to predict online students' marks (using an orthogonal search-based rule extraction algorithm) [24]

In view of the literature survey, it is observed that different analysis has been done on students' results database but the student academic performance in social groups categories in urban and rural areas in Private sector and Government colleges different courses wise in isolation has never been analyzed for hidden and important patterns, which could be of a great importance to academic planners in enhancing their decision making process and improving student Academic performance. The generated association rules are analyzed to make useful and constructive recommendations to the academic planners. This promised to enhance academic planner's sense of decision making and aid in the curriculum structure and modification which in turn improve students' performance and to improve graduate rate. In This paper, we have studied the application of the Association rules Analysis in the curricula of graduate education and how the student performance differ among various social groups between rural and urban areas, between Government and private sector colleges in different courses. We have taken Kakatiya University as base region and studied the results from seven years of data.

III. ASSOCIATION RULE MINING

In data mining techniques, association rules are one of the most preferred techniques, to identify hidden patterns. The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on inference on available data.

One of the popular descriptive data mining techniques is Association rule mining (ARM), owing to its extensive use in marketing and retail communities in addition to many other diverse fields. Mining association rules is particularly useful for discovering relationships among items from large databases.

Association rule mining deals with market basket database analysis for finding frequent item sets and generate valid and important rules. Various association rules mining algorithms have been proposed in 1993 by Aggrawal et. al. [25,26] viz. Apriori, Apriori-TID and Apriori Hybrid. Association rule mining (Aggarwal et.al. al [25].,

1993) is one of the important problems of data mining. The goal of the Association rule mining is to detect relationships or associations between specific values of categorical variables in large data sets. This is a common task in many data mining projects. Suppose I is a set of items, D is a set of transactions, an association rule is an implication of the form X=>Y, where X, Y are subsets of I, and X, Y do not intersect. Each rule has two measures, support and confidence.

Let □ = {*i*1, *i*2 … *im*} be a universe of items. Also, let *T* = {*t*1, *t*2 …tn} be a set of all transactions collected over a given period of time. To simplify a problem, we will assume that every item *i* can be purchased only once in any given transaction *t*. Thus *t* □ □ ("t is a subset of omega"). In reality, each transaction *t* is assigned a number, for example a transaction id (TID). Let now *A* be a set of items (or an itemset). A transaction *t* is said to contain *A* if and only if *A* □ *t*. Now, mathematically, an association rule will be an implication of the form *A* □ *B* Where both *A* and *B* are subsets of □ and *A* □□*B* = □ ("the intersection of sets A and B is an empty set").

Support : The *support* of an item set is the fraction of the rows of the database that contain all of the items in the item set. Support indicates the frequencies of the occurring patterns. Sometimes it is called *frequency*. Support is simply a probability that a randomly chosen transaction *t* contains both item sets *A* and *B*. Mathematically, *Support* (*A* □ *B*) *t* = *P* (*A* □ *t* □ *B* □ *t*) We will use a simplified notation that *Support* (*A* □□*B*) = *P* (*A* □ *B*)

Confidence: Confidence denotes the strength of implication in the rule. Sometimes it is called *accuracy*. Confidence is simply a probability that an item set *B* is purchased in a randomly chosen transaction *t* given that the item set *A* is purchased. Mathematically, *Confidence* (*A* □□*B*) *t* = *P* (*B* □□*t* | *A* □□*t*). We will use a simplified notation that *Confidence* (*A* □ *B*) = *P* (*B* / *A*)

In general, a set of items (such as the antecedent or the consequent of a rule) is called an item set. The number of items in an item set is called the length of an item set. Item sets of some length k are referred to as k-item sets. Generally, an association rules mining algorithm contains the following steps:

- ✓ The set of candidate k-item sets is generated by 1-extensions of the large (k -1)- item sets generated in the previous iteration.
- ✓ Supports for the candidate k-item sets are generated by a pass over the database.
- ✓ Item sets that do not have the minimum support are discarded and the remaining itemsets are called large k-item sets. This process is repeated until no more large item sets are found[27].

IV. PROBLEM STATEMENT

In this paper we wanted to study the students' academic performance in different social group categories in rural and urban areas in Government and private sector colleges, different courses using association rule mining. This is to discover the hidden relationships that exist between different student categories in different college environments. Data has been collected from Kakatiya University (about 298 affiliated under graduate colleges to the University) over a period of seven year from 2000 to 2006 and also Implementation is done using Java Programing and Oracle Software.

There is a popular belief that the education in urban institutions is far better than in rural area. As the availability of educational facilities in rural areas has improved we believed that the quality of education has improved here as well. In order to prove the point we wanted to identify the strong rules with the association of different attributes like the location of the college, type of the college, different social groups, different courses etc., using the Apriori algorithm. Using this algorithm over the data, we wanted to study the performance of different category of students in different environments. We also want to identify the trend of a specific category of students over a period of time.

V. APPROACH

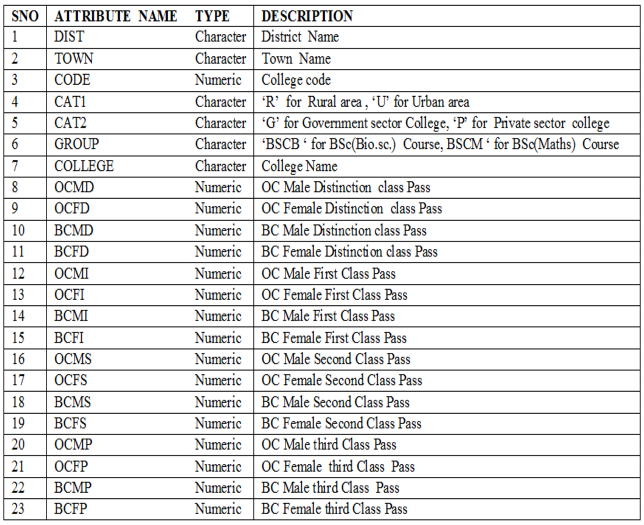
Six years of data has been collected from examinations branch of Kakatiya University for this study purpose. The collected data was purely associated with examinations and hence further several other data was also needed to be collected related to the student's social status, the type and location of the college, etc. The overall activities are broadly categorized in to the following steps:

1. Data collection and Data set preparation.
2. Data preprocessing.
3. Data processing.
4. Implementation of Apriori Algorithm
5. Results & Analysis.

5.1 Data Collection and Data set Preparation:

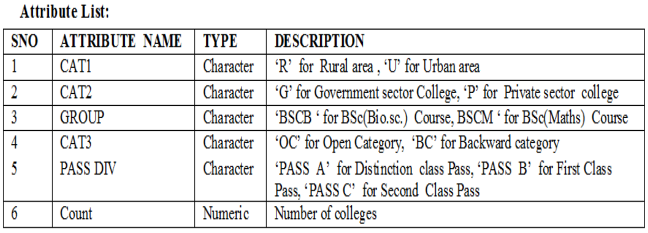
We have collected student data set from 294 affiliated colleges of Kakatiya University from 2000 to 2006. The data set contains the result and marks for B.Sc.(M), B.Sc.(B) Courses from these colleges. There are approximately 5,00,000 records in this data set. Further the personal data of the students containing their social status has been collected from the colleges. The data related to the type of the college and the location (Rural/Urban) is added to this data. After combining all these data sets the resultant database record contains forty eight attributes such as different social groups the different social groups' categories, rural and urban areas and Government and Private sector colleges in different courses wise of each student. As the data collected is from the different sources there needs to be a proper cleaning of data such as filling in missing values; smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Then, the cleaned data are transformed into a form of table that is suitable for data mining model.

5.2 Data Preprocessing: The data collected and brought together is very huge and contains a lot of unwanted details. Hence it is further processed and the following attributes have been identified. The Attribute list below:



SNO	ATTRIBUTE NAME	TYPE	DESCRIPTION
1	DIST	Character	District Name
2	TOWN	Character	Town Name
3	CODE	Numeric	College code
4	CAT1	Character	'R' for Rural area , 'U' for Urban area
5	CAT2	Character	'G' for Government sector College, 'P' for Private sector college
6	GROUP	Character	'BSCB ' for BSc(Bio.sc.) Course, BSCM ' for BSc(Maths) Course
7	COLLEGE	Character	College Name
8	OCMD	Numeric	OC Male Distinction class Pass
9	OCFD	Numeric	OC Female Distinction class Pass
10	BCMD	Numeric	BC Male Distinction class Pass
11	BCFD	Numeric	BC Female Distinction class Pass
12	OCMI	Numeric	OC Male First Class Pass
13	OCFI	Numeric	OC Female First Class Pass
14	BCMI	Numeric	BC Male First Class Pass
15	BCFI	Numeric	BC Female First Class Pass
16	OCMS	Numeric	OC Male Second Class Pass
17	OCFS	Numeric	OC Female Second Class Pass
18	BCMS	Numeric	BC Male Second Class Pass
19	BCFS	Numeric	BC Female Second Class Pass
20	OCMP	Numeric	OC Male third Class Pass
21	OCFP	Numeric	OC Female third Class Pass
22	BCMP	Numeric	BC Male third Class Pass
23	BCFP	Numeric	BC Female third Class Pass

This file contains the data related to an individual details and hence cannot be used directly from processing further. Hence this file is processed further to segregate the information regarding social status wise gender wise course wise and the college wise data. This data is represented in the following form:



Attribute List:

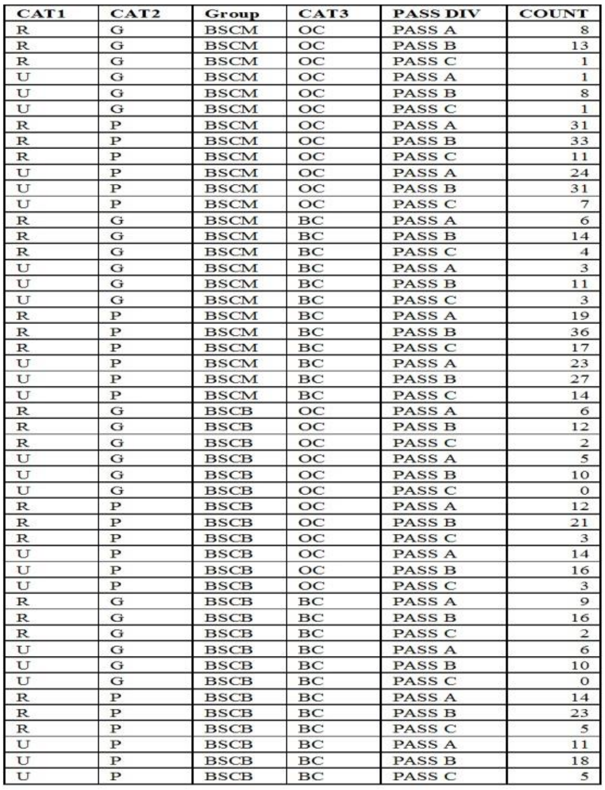
SNO	ATTRIBUTE NAME	TYPE	DESCRIPTION
1	CAT1	Character	'R' for Rural area , 'U' for Urban area
2	CAT2	Character	'G' for Government sector College, 'P' for Private sector college
3	GROUP	Character	'BSCB ' for BSc(Bio.sc.) Course, BSCM ' for BSc(Maths) Course
4	CAT3	Character	'OC' for Open Category, 'BC' for Backward category
5	PASS DIV	Character	'PASS A' for Distinction class Pass, 'PASS B' for First Class Pass, 'PASS C' for Second Class Pass
6	Count	Numeric	Number of colleges

The data contained in this form is chosen for further processing.

5.3 Data Processing:

We use association rule mining to extracting strong association rules for the specified attributes. As stated earlier we will use Apriori algorithm to extract these rules.

Step 1: By using Attribute Oriented Induction concept we have removed unimportant attributes. For the purpose of our analysis use have considered 2002 data.



CAT1	CAT2	Group	CAT3	PASS DIV	COUNT
R	G	BSCM	OC	PASS A	8
R	G	BSCM	OC	PASS B	13
R	G	BSCM	OC	PASS C	1
U	G	BSCM	OC	PASS A	1
U	G	BSCM	OC	PASS B	8
U	G	BSCM	OC	PASS C	1
R	P	BSCM	OC	PASS A	31
R	P	BSCM	OC	PASS B	33
R	P	BSCM	OC	PASS C	11
U	P	BSCM	OC	PASS A	24
U	P	BSCM	OC	PASS B	31
U	P	BSCM	OC	PASS C	7
R	G	BSCM	BC	PASS A	6
R	G	BSCM	BC	PASS B	14
R	G	BSCM	BC	PASS C	4
U	G	BSCM	BC	PASS A	3
U	G	BSCM	BC	PASS B	11
U	G	BSCM	BC	PASS C	3
R	P	BSCM	BC	PASS A	19
R	P	BSCM	BC	PASS B	36
R	P	BSCM	BC	PASS C	17
U	P	BSCM	BC	PASS A	23
U	P	BSCM	BC	PASS B	27
U	P	BSCM	BC	PASS C	14
R	G	BSCB	OC	PASS A	6
R	G	BSCB	OC	PASS B	12
R	G	BSCB	OC	PASS C	2
U	G	BSCB	OC	PASS A	5
U	G	BSCB	OC	PASS B	10
U	G	BSCB	OC	PASS C	0
R	P	BSCB	OC	PASS A	12
R	P	BSCB	OC	PASS B	21
R	P	BSCB	OC	PASS C	3
U	P	BSCB	OC	PASS A	14
U	P	BSCB	OC	PASS B	16
U	P	BSCB	OC	PASS C	3
R	G	BSCB	BC	PASS A	9
R	G	BSCB	BC	PASS B	16
R	G	BSCB	BC	PASS C	2
U	G	BSCB	BC	PASS A	6
U	G	BSCB	BC	PASS B	10
U	G	BSCB	BC	PASS C	0
R	P	BSCB	BC	PASS A	14
R	P	BSCB	BC	PASS B	23
R	P	BSCB	BC	PASS C	5
U	P	BSCB	BC	PASS A	11
U	P	BSCB	BC	PASS B	18
U	P	BSCB	BC	PASS C	5

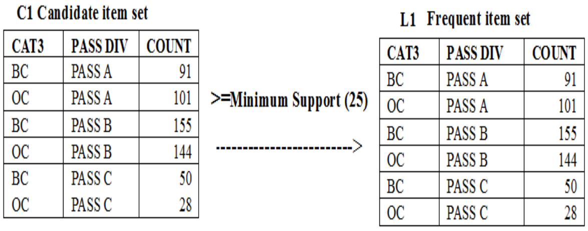
Figure 1. 2002 year Data table

Step 2: In the second step, algorithm scans the table in figure 1.Candidate item set C1 is formed from the combination of CAT3 and PASS DIV attributes. This process is called as joining. In order to get the support values sort the attributes and calculate the sum of each combination. Tabulate the values as shown below.

C1 Candidate item set

CAT3	PASS DIV	COUNT
BC	PASS A	91
OC	PASS A	101
BC	PASS B	155
OC	PASS B	144
BC	PASS C	50
OC	PASS C	28

Step 3: In this step, with the help of minimum support which is determined as 25, L1 item set is obtained which includes the most frequent single items.



C1 Candidate item set

CAT3	PASS DIV	COUNT
BC	PASS A	91
OC	PASS A	101
BC	PASS B	155
OC	PASS B	144
BC	PASS C	50
OC	PASS C	28

>=Minimum Support (25)
------------------------>

L1 Frequent item set

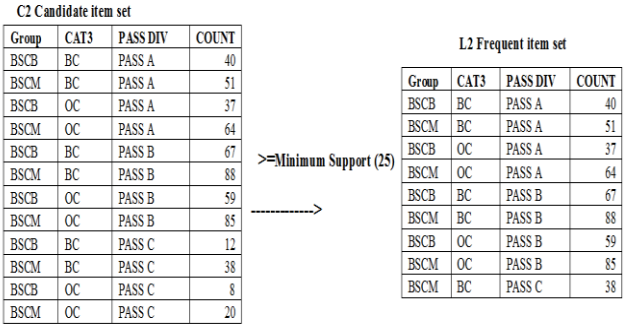
CAT3	PASS DIV	COUNT
BC	PASS A	91
OC	PASS A	101
BC	PASS B	155
OC	PASS B	144
BC	PASS C	50
OC	PASS C	28

Step 4: In this step, a new candidate item set C2 is formed from the combination of GROUP, CAT3 and PASS DIV attributes. According to the summary code of C2 pruning procedure is necessary to be done after candidate set is formed. In pruning, the existence of subsets of C2 candidate item sets in L1 set is controlled and if a subset of any item doesn't take place in L1, the related item is excluded from evaluation and it is deleted from C2 candidate item set. But, here pruning step is not necessary; the reason is all the subsets of C2 take place in L1 set.

C2 Candidate item set

Group	CAT3	PASS DIV	COUNT
BSCB	BC	PASS A	40
BSCM	BC	PASS A	51
BSCB	OC	PASS A	37
BSCM	OC	PASS A	64
BSCB	BC	PASS B	67
BSCM	BC	PASS B	88
BSCB	OC	PASS B	59
BSCM	OC	PASS B	85
BSCB	BC	PASS C	12
BSCM	BC	PASS C	38
BSCB	OC	PASS C	8
BSCM	OC	PASS C	20

Step 5: In this step, from the C2 candidate item set, L2 frequent item set is obtained according to the values that are equal to minimum support.



C2 Candidate item set

Group	CAT3	PASS DIV	COUNT
BSCB	BC	PASS A	40
BSCM	BC	PASS A	51
BSCB	OC	PASS A	37
BSCM	OC	PASS A	64
BSCB	BC	PASS B	67
BSCM	BC	PASS B	88
BSCB	OC	PASS B	59
BSCM	OC	PASS B	85
BSCB	BC	PASS C	12
BSCM	BC	PASS C	38
BSCB	OC	PASS C	8
BSCM	OC	PASS C	20

>=Minimum Support (25)
------------>

L2 Frequent item set

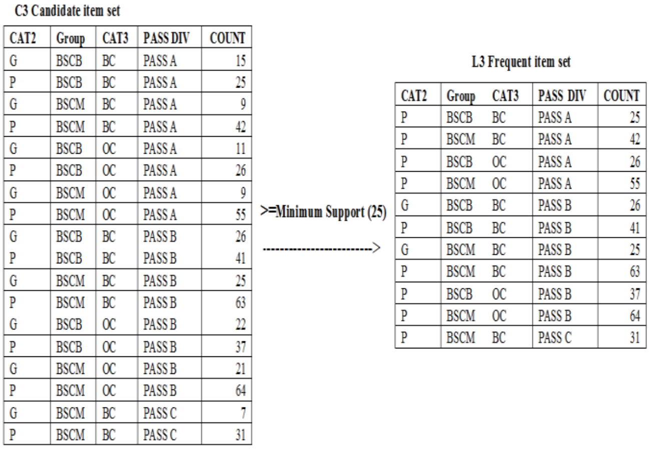
Group	CAT3	PASS DIV	COUNT
BSCB	BC	PASS A	40
BSCM	BC	PASS A	51
BSCB	OC	PASS A	37
BSCM	OC	PASS A	64
BSCB	BC	PASS B	67
BSCM	BC	PASS B	88
BSCB	OC	PASS B	59
BSCM	OC	PASS B	85
BSCM	BC	PASS C	38

Step 6: In 6th step of application, a new candidate item set C3 is formed from the combination of CAT2, GROUP, CAT3 and PASS DIV attributes. According to the summary code of apriori algorithm, pruning step is required after C3 candidate item set has been formed. If a subset of any item does not take place in L2 set, the related item is excluded from the evaluation and is deleted from C3 candidate item set. All items at the end of pruning process can be excluded from the evaluation.

C3 Candidate item set (after pruning)

CAT2	Group	CAT3	PASS DIV	COUNT
G	BSCB	BC	PASS A	15
P	BSCB	BC	PASS A	25
G	BSCM	BC	PASS A	9
P	BSCM	BC	PASS A	42
G	BSCB	OC	PASS A	11
P	BSCB	OC	PASS A	26
G	BSCM	OC	PASS A	9
P	BSCM	OC	PASS A	55
G	BSCB	BC	PASS B	26
P	BSCB	BC	PASS B	41
G	BSCM	BC	PASS B	25
P	BSCM	BC	PASS B	63
G	BSCB	OC	PASS B	22
P	BSCB	OC	PASS B	37
G	BSCM	OC	PASS B	21
P	BSCM	OC	PASS B	64
G	BSCM	BC	PASS C	7
P	BSCM	BC	PASS C	31

Step 7: In this step, from the C3 candidate item set, L3 frequent item set is obtained according to the values that are equal to minimum support.


>=Minimum Support (25)

C3 Candidate item set

CAT2	Group	CAT3	PASS DIV	COUNT
G	BSCB	BC	PASS A	15
P	BSCB	BC	PASS A	25
G	BSCM	BC	PASS A	9
P	BSCM	BC	PASS A	42
G	BSCB	OC	PASS A	11
P	BSCB	OC	PASS A	26
G	BSCM	OC	PASS A	9
P	BSCM	OC	PASS A	55
G	BSCB	BC	PASS B	26
P	BSCB	BC	PASS B	41
G	BSCM	BC	PASS B	25
P	BSCM	BC	PASS B	63
G	BSCB	OC	PASS B	22
P	BSCB	OC	PASS B	37
G	BSCM	OC	PASS B	21
P	BSCM	OC	PASS B	64
G	BSCM	BC	PASS C	7
P	BSCM	BC	PASS C	31

L3 Frequent item set

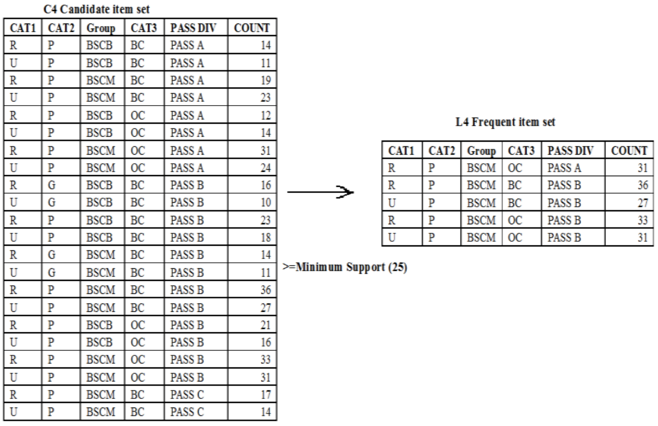
CAT2	Group	CAT3	PASS DIV	COUNT
P	BSCB	BC	PASS A	25
P	BSCM	BC	PASS A	42
P	BSCB	OC	PASS A	26
P	BSCM	OC	PASS A	55
G	BSCB	BC	PASS B	26
P	BSCB	BC	PASS B	41
G	BSCM	BC	PASS B	25
P	BSCM	BC	PASS B	63
P	BSCB	OC	PASS B	37
P	BSCM	OC	PASS B	64
P	BSCM	BC	PASS C	31

Step 8: In this step, a new candidate item set C4 is formed from the combination of CAT1, CAT2, GROUP, CAT3 and PASS DIV attributes. According to the summary code of apriori algorithm, pruning step is required after C4 candidate item set has been formed. If a subset of any item does not take in L3 set, the related item is excluded from the evaluation and is deleted from C4 candidate item set. All items at the end of pruning process can be excluded from the evaluation.

C4 Candidate item set

CAT1	CAT2	Group	CAT3	PASS DIV	COUNT
R	P	BSCB	BC	PASS A	14
U	P	BSCB	BC	PASS A	11
R	P	BSCM	BC	PASS A	19
U	P	BSCM	BC	PASS A	23
R	P	BSCB	OC	PASS A	12
U	P	BSCB	OC	PASS A	14
R	P	BSCM	OC	PASS A	31
U	P	BSCM	OC	PASS A	24
R	G	BSCB	BC	PASS B	16
U	G	BSCB	BC	PASS B	10
R	P	BSCB	BC	PASS B	23
U	P	BSCB	BC	PASS B	18
R	G	BSCM	BC	PASS B	14
U	G	BSCM	BC	PASS B	11
R	P	BSCM	BC	PASS B	36
U	P	BSCM	BC	PASS B	27
R	P	BSCB	OC	PASS B	21
U	P	BSCB	OC	PASS B	16
R	P	BSCM	OC	PASS B	33
U	P	BSCM	OC	PASS B	31
R	P	BSCM	BC	PASS C	17
U	P	BSCM	BC	PASS C	14

Step 9: In this step, from the C4 candidate item set, L4 frequent item set is obtained according to the count that are equal to minimum support or higher.


>=Minimum Support (25)

C4 Candidate item set

CAT1	CAT2	Group	CAT3	PASS DIV	COUNT
R	P	BSCB	BC	PASS A	14
U	P	BSCB	BC	PASS A	11
R	P	BSCM	BC	PASS A	19
U	P	BSCM	BC	PASS A	23
R	P	BSCB	OC	PASS A	12
U	P	BSCB	OC	PASS A	14
R	P	BSCM	OC	PASS A	31
U	P	BSCM	OC	PASS A	24
R	G	BSCB	BC	PASS B	16
U	G	BSCB	BC	PASS B	10
R	P	BSCB	BC	PASS B	23
U	P	BSCB	BC	PASS B	18
R	G	BSCM	BC	PASS B	14
U	G	BSCM	BC	PASS B	11
R	P	BSCM	BC	PASS B	36
U	P	BSCM	BC	PASS B	27
R	P	BSCB	OC	PASS B	21
U	P	BSCB	OC	PASS B	16
R	P	BSCM	OC	PASS B	33
U	P	BSCM	OC	PASS B	31
R	P	BSCM	BC	PASS C	17
U	P	BSCM	BC	PASS C	14

L4 Frequent item set

CAT1	CAT2	Group	CAT3	PASS DIV	COUNT
R	P	BSCM	OC	PASS A	31
R	P	BSCM	BC	PASS B	36
U	P	BSCM	BC	PASS B	27
R	P	BSCM	OC	PASS B	33
U	P	BSCM	OC	PASS B	31

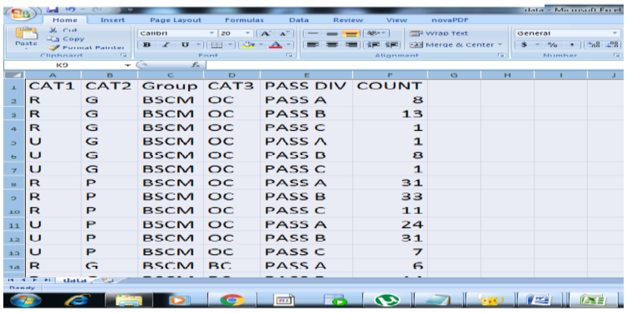
Step 10: In 10th step of application, in this situation the algorithm is finished and the items of L4 item set which are equal to determined minimum support or having higher are used in forming association rules. These steps can be repeated according to the situation of data set or can be ended before they have reached at this step.

Final Pattern Rules

CAT1	CAT2	Group	CAT3	PASS DIV	COUNT
R	P	BSCM	OC	PASS A	31
R	P	BSCM	BC	PASS B	36
U	P	BSCM	BC	PASS B	27
R	P	BSCM	OC	PASS B	33
U	P	BSCM	OC	PASS B	31

5.4 Implementation of Apriori Algorithm

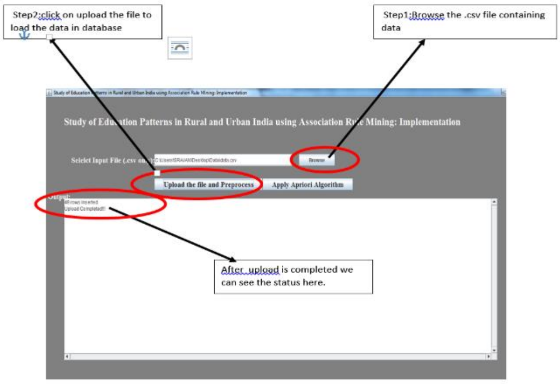
Our implementation is done by using JAVA and ORACLE software's. In this Paper first we are collecting the data, then data set preparation, Data pre-processing will be done.



CAT1	CAT2	Group	CAT3	PASS DIV	COUNT
R	G	BSCM	OC	PASS A	8
R	G	BSCM	OC	PASS B	13
R	G	BSCM	OC	PASS C	1
U	G	BSCM	OC	PASS A	1
U	G	BSCM	OC	PASS B	8
U	G	BSCM	OC	PASS C	1
R	P	BSCM	OC	PASS A	31
R	P	BSCM	OC	PASS B	33
R	P	BSCM	OC	PASS C	11
U	P	BSCM	OC	PASS A	24
U	P	BSCM	OC	PASS B	31
U	P	BSCM	OC	PASS C	7
R	G	BSCM	BC	PASS A	6

Fig. 1. Data collecting for input file .csv format

In this following of processing we are processing the data. After data processing, browse the file which containing data, then upload the file and pre-processing.


Study of Education Patterns in Rural and Urban India using Association Rule Mining: Implementation
Study of Education Patterns in Rural and Urban India using Association Rule Mining: Implementation
Selelct Input File (.csv only):
Browse
Upload the file and Preprocess
Apply Apriori Algorithm
Output:

Step2:click on upload the file to load the data in database

Step1:Browse the .csv file containing data

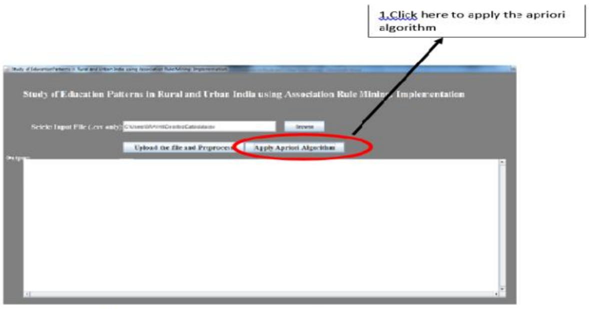
C:\Users\SRAVAN\Desktop\Data\data.csv

48 rows inserted
Upload Completed!!

After upload is completed we can see the status here.

Fig. 2. Browse and upload the input .csv file

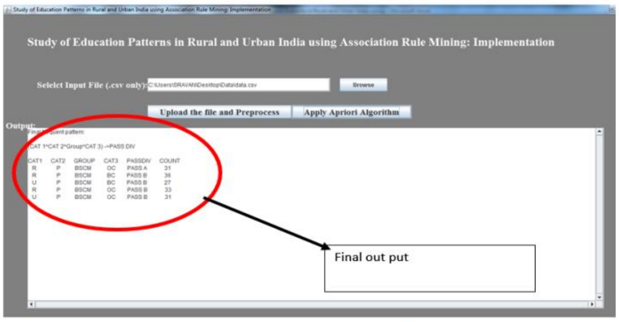
Here after uploading the file we will get a status on the output screen. Click to apply apriori algorithm, after applying apriori algorithm we get a final result get result.


Study of Education Patterns in Rural and Urban India using Association Rule Mining: Implementation
Study of Education Patterns in Rural and Urban India using Association Rule Mining: Implementation
Selelct Input File (.csv only):
Browse
Upload the file and Preprocess
Apply Apriori Algorithm
Output:

1.Click here to apply the apriori algorithm

C:\Users\SRAVAN\Desktop\Data\data.csv

Fig. 3. Apply Apriori Algorithm


Study of Education Patterns in Rural and Urban India using Association Rule Mining: Implementation
Study of Education Patterns in Rural and Urban India using Association Rule Mining: Implementation
Selelct Input File (.csv only):
Browse
Upload the file and Preprocess
Apply Apriori Algorithm
Output:

C:\Users\SRAVAN\Desktop\Data\data.csv

```
Final frequent pattern

(CAT 1^CAT 2^Group^CAT 3) ->PASS DIV

CAT1   CAT2   GROUP   CAT3   PASSDIV   COUNT
 R      P     BSCM     OC    PASS A     31
 R      P     BSCM     BC    PASS B     36
 U      P     BSCM     BC    PASS B     27
 R      P     BSCM     OC    PASS B     33
 U      P     BSCM     OC    PASS B     31
```

Final out put

Fig. 4. Displaying Association Rules

5.5 Result & Analysis:

The available data has been processed as specified above for all the years (for year 2000-2006) and the support and confidence for each of the strong rules identified has been calculated. The results are tabulated as below.

S No	Year	(CAT 1^CAT 2^Group^CAT 3) ->PASS DIV	Count	Support	Confidence
1	2000	(U^P^BSCM^BC)->PASS B	28	0.318182	0.58333333
2	2001	(U^P^BSCM^BC)->PASS B	27	0.303371	0.55102041
3	2001	(U^P^BSCM^OC)->PASS B	25	0.280899	0.58139535
4	2002	(R^P^BSCM^OC) ->PASS A	31	0.28972	0.56363636
5	2002	(R^P^BSCM^BC) ->PASS B	36	0.336449	0.57142857
6	2002	(U^P^BSCM^BC) ->PASS B	27	0.252336	0.42857143
7	2002	(R^P^BSCM^OC) ->PASS B	33	0.308411	0.515625
8	2002	(U^P^BSCM^OC) ->PASS B	31	0.28972	0.484375
9	2003	(R^P^BSCM^BC) ->PASS A	42	0.328125	0.6
10	2003	(U^P^BSCM^BC) ->PASS A	28	0.21875	0.4
11	2003	(R^P^BSCM^OC) ->PASS A	44	0.34375	0.59459459
12	2003	(U^P^BSCM^OC) ->PASS A	30	0.234375	0.40540541
13	2003	(R^P^BSCB^BC) ->PASS B	33	0.34375	0.58928571
14	2003	(R^P^BSCM^BC) ->PASS B	54	0.421875	0.61363636
15	2003	(U^P^BSCM^BC) ->PASS B	34	0.265625	0.38636364
16	2003	(R^P^BSCB^OC) ->PASS B	30	0.3125	0.56603774
17	2003	(R^P^BSCM^OC) ->PASS B	49	0.382813	0.60493827
18	2003	(U^P^BSCM^OC) ->PASS B	32	0.25	0.39506173
19	2004	(R^P^BSCM^BC) ->PASS A	63	0.425676	0.65625
20	2004	(U^P^BSCM^BC) ->PASS A	33	0.222973	0.34375
21	2004	(R^P^BSCM^OC) ->PASS A	56	0.378378	0.62921348
22	2004	(U^P^BSCM^OC) ->PASS A	33	0.222973	0.37078652
23	2004	(R^P^BSCB^BC) ->PASS B	45	0.405405	0.61643836
24	2004	(U^P^BSCB^BC) ->PASS B	28	0.252252	0.38356164
25	2004	(R^P^BSCM^BC) ->PASS B	70	0.472973	0.66666667
26	2004	(U^P^BSCM^BC) ->PASS B	35	0.236486	0.33333333
27	2004	(R^P^BSCB^OC) ->PASS B	37	0.333333	0.60655738
28	2004	(R^P^BSCM^OC) ->PASS B	59	0.398649	0.62765957
29	2004	(U^P^BSCM^OC) ->PASS B	35	0.236486	0.37234043
30	2005	(R^P^BSCB^BC) ->PASS A	40	0.310078	0.60606061
31	2005	(U^P^BSCB^BC) ->PASS A	26	0.20155	0.39393939
32	2005	(R^P^BSCB^OC) ->PASS A	28	0.217054	0.65306122
33	2005	(R^P^BSCB^BC) ->PASS B	56	0.434109	0.34693878
34	2005	(U^P^BSCB^BC) ->PASS B	30	0.232558	0.60869565
35	2005	(R^P^BSCB^OC) ->PASS B	46	0.356589	0.62637363
36	2005	(U^P^BSCB^OC) ->PASS B	30	0.232558	0.37362637
37	2005	(R^P^BSCM^BC) ->PASS A	64	0.426667	0.65116279
38	2005	(U^P^BSCM^BC) ->PASS A	34	0.226667	0.34883721
39	2005	(R^P^BSCM^OC) ->PASS A	57	0.38	0.66336634
40	2005	(U^P^BSCM^OC) ->PASS A	34	0.226667	0.33663366
41	2005	(R^P^BSCM^BC) ->PASS B	67	0.446667	0.60526316
42	2005	(U^P^BSCM^BC) ->PASS B	34	0.226667	0.39473684
43	2005	(R^P^BSCM^OC) ->PASS B	60	0.4	0.65217391
44	2005	(U^P^BSCM^OC) ->PASS B	32	0.213333	0.34782609
45	2006	(R^P^BSCB^BC) ->PASS A	73	0.424419	0.71568627
46	2006	(U^P^BSCB^BC) ->PASS A	29	0.168605	0.28431373
47	2006	(R^P^BSCB^OC) ->PASS A	54	0.313953	0.72173913
48	2006	(R^P^BSCB^BC) ->PASS B	85	0.494186	0.27826087
49	2006	(U^P^BSCB^BC) ->PASS B	35	0.203488	0.69230769
50	2006	(R^P^BSCB^OC) ->PASS B	71	0.412791	0.69607843
51	2006	(U^P^BSCB^OC) ->PASS B	29	0.168605	0.30392157
52	2006	(R^P^BSCM^BC) ->PASS A	83	0.439153	0.70833333
53	2006	(U^P^BSCM^BC) ->PASS A	32	0.169312	0.29166667
54	2006	(R^P^BSCM^OC) ->PASS A	71	0.375661	0.73643411
55	2006	(U^P^BSCM^OC) ->PASS A	31	0.164021	0.26356589
56	2006	(R^P^BSCM^BC) ->PASS B	95	0.502646	0.71
57	2006	(U^P^BSCM^BC) ->PASS B	34	0.179894	0.29
58	2006	(R^P^BSCM^OC) ->PASS B	71	0.375661	0.71
59	2006	(U^P^BSCM^OC) ->PASS B	29	0.153439	0.29

To obtain meaningful knowledge from this data, we have drawn charts for support and confidence for similar rules of all the years, as below. The charts show our belief that over the years the educational environment in the rural areas has improved and there by the results of the students show a tremendous improvement. Further the urban areas have started coming down in their educational quality.

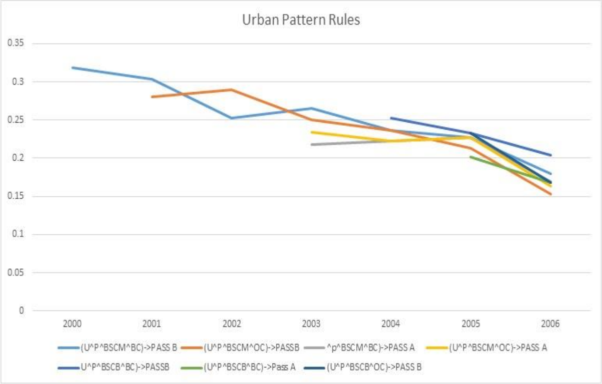


Figure 5

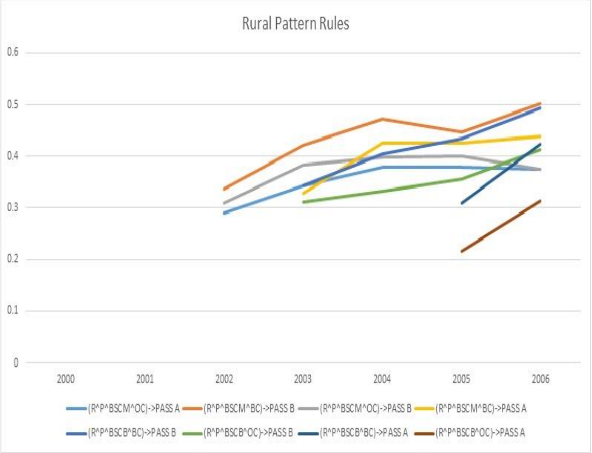


Figure 6

VI. CONCLUSIONS

The developments in urban areas in different sectors have resulted in educational environment changes and our study to find the relevance to the student performance associated with the educational environment in which they study was quite successful as the result show. We have Implemented Java Programing and Oracle Software. We have identified the association between different attributes of educational environment i.e., the location of the college, type of the college, different social groups, different courses etc., and thereby extract strong association rules. The results have proved that the performance of the students in the rural areas have been slowly but steadily increasing. We have done the implementation by Java Programing and Oracle Software. In the meantime the performance of the students in urban areas has been decreasing.

VII. ACKNOWLEDGEMENT

This work is partially supported by University Grants Commission, New Delhi, India, under Grant No.MRP-4585/14(SERO/UGC). We would like to thankful to the University Grants Commission, New Delhi, India, for providing the financial assistance in terms of Minor Research Project in the Department of Master of Computer Applications, Gokraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India.

VIII. REFERENCES

[1]. B.Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer and, W. F. Punch."Predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA" In Proceedings of ASEE/IEEE Frontiers in Education Conference,Boulder, CO: IEEE, 2003.

[2]. Talavera, L., and Gaudioso, E. "Mining student data to characterize similar behavior groups in unstructured collaboration spaces". In Proceedings of the Arti_cial Intelligence in Computer Supported Collaborative Learning Workshop at the ECAI ,Valencia, Spain, 2004.

[3]. S. Z. ERDOĞAN, M. TİMOR . "A data mining application in a student database". Journal of aeronautics and space technologies ,volume 2 number 2 (53-57) 2005.

[4]. G.J. Hwang. "A Knowledge-Based System as an Intelligent Learning Advisor on Computer Networks" Journal of Systems, Man, and Cybernetics Vol. 2 , pp.153-158, 1999.

[5]. G.J. Hwang, T.C.K. Huang,and C.R. Tseng. "A Group-Decision Approach for EvaluatingEducational Web Sites". Computers & Education Vol. 42 pp. 65-86 , 2004.

[6]. G.J. Hwang, C.R. Judy, C.H. Wu, C.M. Li and G.H. Hwang. "Development of an Intelligent Management System for Monitoring Educational Web Servers". In proceedings of

the 10th Pacific Asia Conference on Information Systems, PACIS . 2334-2340, 2004.

[7]. G.D. Stathacopoulou, M. Grigoriadou. "Neural Network-Based Fuzzy Modeling of the Student in Intelligent Tutoring Systems". In proceedings of the International Joint Conference on Neural Networks. Washington ,3517-3521,1999.

[8]. C.J. Tsai, S.S. Tseng, and C.Y. Lin. "A Two-Phase Fuzzy Mining and Learning Algorithm for Adaptive Learning Environment". In proceedings of the Alexandrov, V.N., et al. (eds.): International Conference on Computational Science, ICCS 2001. LNCS Vol. 2074. Springer-Verlag, Berlin Heidelberg New York, 429-438. 2001.

[9]. B. Dogan, A. Y. Camurcu. "Association Rule Mining from an Intelligent Tutor" Journal of Educational Technology Systems Volume 36, Number 4 / 2007-2008, pp 433 – 447, 2008

[10]. S. Encheva , S. Tumin. " Application of Association Rules for Efficient Learning Work-Flow" Intelligent Information Processing III , ISBN 978-0-387-44639-4, pp 499-504 published Springer Boston, 2007.

[11]. H.H. Hsu, C.H. Chen, W.P. Tai. "Towards Error-Free and Personalized Web-Based Courses". In proceedings of the 17th International Conference on Advanced Information Networking and Applications, AINA'03. March 27-29, Xian, China, 99-104, 2003.

[12]. P. L. Hsu, R. Lai, C. C. Chiu, C. I. Hsu (2003) "The hybrid of association rule algorithms and genetic algorithms for tree induction: an example of predicting the student course performance" Expert Systems with Applications 25 (2003) 51–62.

[13]. A.Y.K. Chan, K.O. Chow, and K.S. Cheung. "Online Course Refinement through Association Rule Mining" Journal of Educational Technology Systems Volume 36, Number 4 / 2007-2008, pp 433 – 44, 2008.

[14]. S. Saxena, A. S.Pandya, R. Stone, S. R. and S. Hsu (2009) "Knowledge Discovery through Data Visualization of Drive Test Data" International Journal of Computer Science and Security (IJCSS), Volume (3): Issue (6) pp. 559-568.

[15]. S. Das and B. Saha (2009) "Data Quality Mining using Genetic Algorithm" International Journal of Computer Science and Security, (IJCSS) Volume (3) : Issue (2) pp. 105-112

[16]. M.Anandhavalli , M.K.Ghose and K.Gauthaman(2009) "Mining Spatial Gene Expression Data Using Association Rules". International Journal of Computer Science and Security, (IJCSS) Volume (3) : Issue (5) pp. 351-357

[17]. R. Damaševicius. "Analysis of Academic Results for Informatics Course Improvement Using Association Rule Mining". Information Systems Development Towards a Service Provision Society. ISBN 978-0-387-84809-9 (Print) 978-0-387-84810-5 (Online) pp 357- 363, published by Springer US, 2009.

[18]. Nebot, A., Castro, F., Vellido, A., Mugica, F. (2006). Identification of fuzzy models to predict students performance in an e-learning environment. In International Conference on Web-based Education, Puerto Vallarta, 74-79.

[19]. Chen, c., Chen, M., Li, Y. (2007). Mining key formative assessment rules based on learner portfiles for web-based learning systems. In IEEE International Conference on Advanced Learning Technologies, Japan, 1-5.

[20]. Ogor, E.N. (2007). Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques. In Electronics, Robotics and Automotive Mechanics Conference, Washington, DC, 354-359.

[21]. Shangping, D., Ping, Z. (2008). A data mining algorithm in distance learning. In International Conference on Computer Supported Cooperative Work in Design, Xian, 1014-1017.

[22]. Zafra, A., Ventura, S. (2009), Predicting student grades in learning management systems with multiple instance programming. In International Conference on Educational Data Mining, Cordoba, Spain, 307-314.

[23]. Chan, C.C. (2007). A Framework for Assessing Usage of Web-Based e- Learning Systems. In International Conference on innovative Computing, Information and Control, Washington, DC, 147- 151.

[24]. Etchells, T.A., Nebot, A., Vellido, A., Lisboa, P.J.G., Mugica, F. (2006). Learning what is important: feature selection and rule extraction in a virtual course. In European Symposium on Artificial Neural Networks, Bruseles, Belgium, 401-406.

[25]. Rakesh Aggarwal, Tomasz Imielinski, Arun Swami, " Mining Association Rules between Sets of Items in Large Databases" ACM Sigmod Conference Washington DC, May 1993.

[26]. Rakesh Aggarwal , Ramakrishanan Srikant, "Fast Algorithm for mining Association Rules", IBM Almaden Research Centre, Proceedings of 20th VLDB Conference, Santiago, Chile, 1994.

[27]. Sotiris Kotsiantis, Dimitris Kanellopoulos, "Association Rules Mining: A Recent Overview", GESTS International conference on computer science and engineering, Vol. 32(1) pp- 71-82, 2006.