IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 4.295

(Volume2, Issue6)

Available online at: www.ijariit.com

Educational Data Mining: Recognising and Forming Groups of Competent Students for Contests

Mr. ASHOK M V[1]
Head of the Department
Dept. of Computer Science
Teachers Academy, Bangalore
ashokmv@ymail.com

APOORVA A[2]
Assistant Professor,
Department of MCA
Global Institute of Management Sciences, Bangalore
a.apoorva89@gmail.com

Dr. G Suganthi[3]
Associate Professor,
Dept. of Computer Science
Women's Christian College,
Nagrcoil, Tamil Nadu

Abstract: Educational Data Mining is an area where in a combination of techniques such as data mining, machine Learning and statistics, is applied on educational data to get valuable information. The main objective is to recognize competent students based on marks are using clustering (X-means algorithm); then the subjects studied by them are classified into different categories and finally better combination of students as groups or teams are chosen to represent college for contests using association rules. To assess the performance of the proposed model, a student dataset of MCA from a college in Bangalore were collected for the study as a synthetic data. The accuracy of the results obtained from the proposed model was found to be promising. It was found from the study that 3 groups of 2 teams per group emerged as better combinations.

Keywords: Educational data mining, competent student, Apriori algorithm, X-means algorithm.

1. INTRODUCTION

Educational Data Mining (EDM) is the application of Data Mining (DM) techniques to educational data, and so, its objective is to analyze these types of data in order to resolve educational research issues.

1.1 PROBLEM STATEMENT

Normally hundreds of students will be there in institutions. Many inter-collegiate activities happen simultaneously. Teams of students need to be selected to represent the activities. If best students form a team and if they participate in one college then there won't be any strong teams available to represent other colleges. Hence the problem is to select better teams having the combination of good, better, and best students so that the representation is uniform and chances of winning is more.

2. RELATED WORKS

Performance appraisal system is basically a formal interaction between an employee and the supervisor or management conducted periodically to identify the areas of strength and weakness of the employee. The objective is to be consistent about the strengths and work on the weak areas to improve performance of the individual and thus achieve optimum process quality [8].(Chein and Chen,2006 [9]Pal and Pal ,2013[11]. Khan, 2005 [12], Baradwaj and Pal, 2011 [13], Bray [14], 2007, S. K. Yadav et al.,2011[15];X**-means clustering** is a variation of k-means clustering that refines cluster assignments by repeatedly attempting subdivision, and keeping the best resulting splits, until some criterion is reached. Dan pelleg, Andrew Moree, X-means: Extending K-means with Efficient Estimation of the Number of Clusters[2],Thomas Laloe, Remi Servien, The X-Alter algorithm : a parameter-free method to perform unsupervised clustering[3];**Association rules** are if/then statements that help uncover relationships between seemingly unrelated data in a transactional database, relational database or other information repository. Discovering association rules is one of the most important task in data mining. Many efficient algorithms have been proposed. Close algorithm by Nicolas Pasquier et al., (1999)[9]**,** an algorithm that incorporates buyer management and novel estimation and pruning techniques by Rakesh Agrawal et al., (1993) [5], an approach based on decision support system designed for business users who make use of association rules Rok Rupnik et al.,2007[6], to name a few and also it can be applied for effective decision making AkashRajak et al., 2007[7].In our problem association rules has been applied to recognize student buying pattern.

3. PROPOSED MODEL

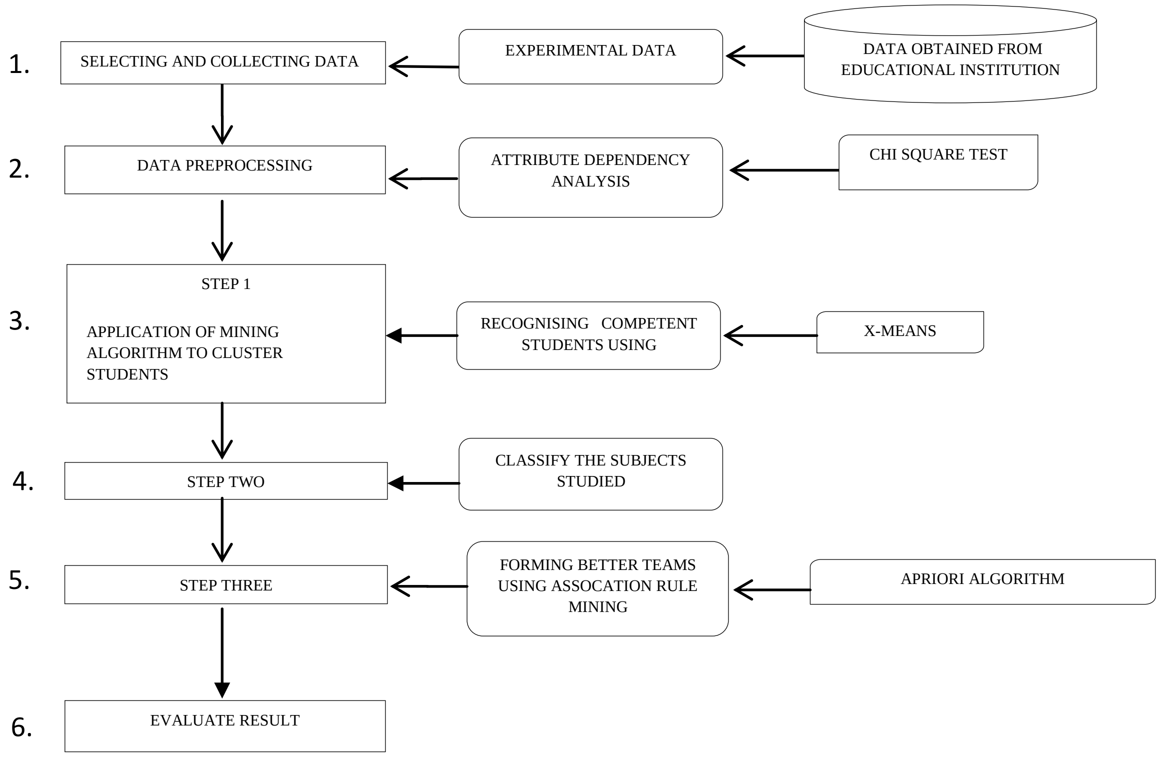
1.
SELECTING AND COLLECTING DATA
EXPERIMENTAL DATA
DATA OBTAINED FROM EDUCATIONAL INSTITUTION
2.
DATA PREPROCESSING
ATTRIBUTE DEPENDENCY ANALYSIS
CHI SQUARE TEST
3.
STEP 1
APPLICATION OF MINING ALGORITHM TO CLUSTER STUDENTS
RECOGNISING COMPETENT STUDENTS USING
X-MEANS
4.
STEP TWO
CLASSIFY THE SUBJECTS STUDIED
5.
STEP THREE
FORMING BETTER TEAMS USING ASSOCATION RULE MINING
APRIORI ALGORITHM
6.
EVALUATE RESULT

Fig 3.1: FLOWCHART OF THE PROPOSED MODEL

4. DATA DESCRIPTION

Table 4.1: Database description

Variables	**Description**	**Possible Values**
Stu_id	Id of the student	{Int}
Name	Name of the student	{Text}
Sub	Subject name	{Text}
S_MARKS	Marks scored in each subject	{1, 2, 3, 4, 5...100}
T_MARKS	Total marks	{ 1% - 100% }
Com	(Communication+Attitude) score out of 10	{1, 2, 3, 4, 5...10}
Min	Minimum marks for passing a subject	32
Max	Maxmum marks for passing a subject	100

Stu_Id:– ID of the student. It can take any integer values.
Name: - Name of the student.
Sub: – represents the name of the subject. It can take only text values ranging from A-Z.

S_MARKS:–various subject marks scored by a student. It can take only the numeric values from 0 to 100.
T_MARKS: – total marks scored by each student represented in the form percentage i.e., 1% to 100%.
Com: –Communication and attitude score out of 10
Min:-Minimum marks for passing a subject
Max: - Maximum marks for passing a subject

5. METHODOLOGY

Step 1: Data collection

Input Table contains Student name, student id, Subject, Minimum marks, maximum marks, Subject marks, Total marks and communication skill as fields. Marks scored in selected subjects of a student over a period of three years of MCA i.e., from June, 2011 to April, 2014 is considered and collected from a college in Bangalore.

Step 2: Data preprocessing:

After Preprocessing the input table following table is obtained.

Table 5.1: Preprocessed table

Stu_id	1	2	3	4	…
Sub	**S_MARKS**	**S_MARKS**	**S_MARKS**	**S_MARKS**	**S_MARKS**
Ca	20	98	45	92	…
Bi	23	98	69	83	…
Java	24	97	67	74	…
Se	25	96	89	92	…
Cf	26	95	88	88	…
Db	28	90	56	81	…
…	…	..	…	…	…
T_MARKS	624	1910	1416	1482	…
Com	7	9	7	8	…

Preprocessing is done using following statistical technique.

Chi-square test: is applied to remove the useless variable that doesn't contribute to the result. Name, max and min marks were removed.

5.1 Proposed model

Step 1: Clustering using X-means algorithm.

Step 1.1: Preprocessed table will be the input for X-means.
Step 1.2: Cluster competent student segment [CCS] and determine the exact number of clusters. The value of X is found using heuristic method incrementing the value of 'X' by one in each step and the results are shown below.
Partition of CCS is done initially by taking X= 2
After Applying X- means clustering with X= 2, we have

Table 5.1.1: Partial view of clusters of students, for X= 2

Cluster 1	1	10	11	12	13	16	18	21	22	24	26	27	29	30		
Cluster 2	2	3	4	5	6	7	8	9	14	15	17	19	20	23	25	28

The above table shows the grouping of students into two groups. .

Table 5.1.2: Difference between clusters for X= 2

Cluster	**Cluster1**	**Cluster2**
Custer 1	0	0.22933004097360
Custer 2	0. 22933004097360	0

For X = 2, the distance between the groups are labeled; in this 0.23 is the minimum value.
For X= 3 applying X- means clustering, we have the following results

Table 5.1.3: Partial view of three clusters, for X= 3

Cluster 1	1	10	11	12	13	16	18	21	22	26	27	29	30		
Cluster 2	9	24													
Cluster 3	2	3	4	5	6	7	8	14	15	17	19	20	23	25	28

The above table indicates the partial view of 3 -clusters.

Table 5.1.4: Difference between clusters

Cluster	**Cluster1**	**Cluster2**	**Cluster3**
Custer 1	0	0.11664247116703	0.16588031292846
Custer 2	0.11664247116703	0	0.15421606581176
Custer 3	0.16588031292846	0.15421606581176	0

For X = 3, the distance between the groups are labeled, in this 0.12 is the minimum value.
For X= 4: we have the following results

Table 5.1.5: Partial view of four clusters, for X= 4

Cluster 1	1													
Cluster 2	9	10	11	12	13	16	18	21	22	24	26	27	29	30
Cluster 3	3	4	5	6	7	8	14	15	17	19	20	23	25	28
Cluster 4	2													

Table 5.1.6: Comparison of distance between the clusters

Cluster	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**
Custer 1	0	0.10453703703704	0.18785185185185	0.3428333333333
Custer 2	0. 10453703703704	0	0.083314814814815	0.2382962962963
Custer 3	0.18785185185185	0.083314814814815	0	0.15498148148148
Custer 4	0.3428333333333	0.2382962962963	0.15498148148148	0

Comparison table given above compares the two clusters in terms of distance between them. Cluster 2- cluster 1 =0.104537 given in row 1 column 3.Similarly the other values are calculated. This table is the resultant of application of X-means, incrementing value of 'X' in every step by 1.

Table 5.1.7: Cluster distance table

Number of cluster	The short cluster distance
Cluster 2	0.2293
Cluster 3	0.1658
Cluster 4	0.3428
Cluster 5	0.3133

The first value 0.2293 in the shorter cluster distance field represents the distance between the cluster 1 and 2, similarly the second value viz., 0.1658 represents the distance between 1 and 3. The other values in the table can be interpreted similarly.

From the above table it can be observed that, values in the ‘ shorter cluster distance’ attribute starts increasing by greater extent i.e., from 0.1658 to 0.3428, after cluster 2..Hence it can be concluded that the maximum number clusters that can be formed is 3. So we choose X= 3 and 3rd cluster because the centroid of the third cluster is nearest to maximum marks of the subjects i,e., 2000(20 subjects).

Step 2: Classification of subjects studied by competent students

Step 2.1: Choosing the cluster

When ‘X’ takes value 3 i.e.,X= 3, the 3rdcluster is chosen as the best cluster as the centroid value of the third cluster is nearest to maximum marks of the subjects i,e., 2000(20 subjects).

The objective is to classify the subjects of those students recognized in the 3rd cluster of the step 1.

Table 5.1.8: Number of subjects studied by each student of the cluster 3

Ca	bi	java	se	cf	Db	cn	php	ooad	c++
asp.net	data mining	data warehousing	dbms lab	java lab	cloud computing	asp.net project	java project	mysql	Cobol

The above table indicates the number of subjects studied by each student of the cluster 3

Step 2.2: Identifying the elements of the cluster.

Table 5.1.9: Elements of cluster 3

Cluster 3	2	3	4	5	6	7	8	14	15	17	19	20	23	25	28

The table above represents the elements of the best cluster recognized in the step 1

Step 2.3: Identifying the category of the subject and toppers.

The total subjects studied should be classified into 3 categories i.e. Programming, database and general subjects of MCA along with the top 6 students with their respective percentages in each category.

5.1.10 Categorized Subjects

Programming	Java	Php	Cobol	asp.net	java lab	java project	asp.net project	C++
DB	data mining	data warehousing	dbms lab	cloud computing	Mysql			
General	Ca	bi	Se	Cf	ooad	cn		

Step 1: Each subject in the excel sheet is compared with each elements of 3 databases namely i,e,. Programming, Database and general subjects. If the processed subject in the excel sheet is found to be in the programming subject listing, its marks is considered and if the next subject that is processed in the excel sheet is also found in the programming subject list, the marks gets added to the previous marks and the process repeats until all the programming subjects are scanned. Finally percentage of the entire programming subjects of a particular student is taken.

The same process is repeated for database and general subjects to find the respective percentages of each student.

Step 2: Listing of three classified categories of top 6 students with their percentages and id's as shown in the table below.

Table 5.1.11: Output table after classification

General			Programming			Database		
Toppers	Stu_id	Percent	Toppers	Stu_id	Percent	Toppers	Stu_id	percent
1	19	95	1	2	95	1	2	94
2	2	94	2	20	75	2	17	88
3	17	83	3	17	75	3	23	86
4	20	81	4	28	74	4	5	82
5	3	75	5	8	73	5	28	79
6	28	75	6	4	72	6	19	79

Step 3: Choosing the best two teams using association rules.

The algorithm used to accomplish the above task is Apriori. The application of the steps of Apriori is explained below.

Step 3: Generating 3-item set with its corresponding support count value
Step 3.1.1 Item sets are obtained by taking Cartesian product of programming, Database and general categories as displayed above i.e., table number.

Table 5.1.12: Extract of Cartesian product table.

1st Element of combination		2nd Element of combination		3rd Element of combination		Support count
Id	Percent	Id	percent	Id	percent	
2	94	19	95	2	95	3
17	88	19	95	2	95	4
.	.	.	.	.	.	.
.	.	.	.	.	.	.
5	82	2	94	28	74	10
28	79	2	94	28	74	11

.	.	.	.	.	.	.
23	86	20	81	17	75	10
5	82	20	81	17	75	11
.	.	.	.	.	.	.
17	83	28	74	23	86	10
.	.	.	.	.	.	.
3	75	2	95	5	82	10

Above table is an extract of the resultant of Cartesian product of programming, database and general subjects listing along with support count. The illustration of calculation of support count is as follows.

Step 3.1.2 Calculation of Support count.

Support count=$\sum_{i=1}^{i=6}$ (topper value of programming[i]+ topper value of database[i]+ topper value of general[i])

Sum of all values of toppers of all three categories headed by topmost student i,e,. 1in toppers column which implies 1+1+1=3

Sum of all values of toppers of all three categories headed by row 6 in toppers column which implies 6+6+6=18

Step 3.1.3 to find the threshold value.

Threshold value=absolute [(minimum support count+ maximum support count)/2]

=absolute [(3+18)/2] =10

Table 5.1.13: Elements with threshold value <=10

1st Element of combination		2nd Element of combination		3rd Element of combination		Support count
Id	Percent	Id	percent	Id	percent	
19	95	2	95	17	88	4
9	95	2	95	23	86	5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
2	94	20	75	5	82	8
.	.	.	.	.	.	.
17	83	18	74	23	86	10
.	.	.	.	.	.	.
.	.	.	.	.	.	.
2	94	28	74	23	86	9
.	.	.	.	.	.	.
17	83	20	75	5	82	9
.	.	.	.	.	.	.
2	94	28	74	5	82	10
.	.	.	.	.	.	.
17	83	20	75	23	80	8
.	.	.	.	.	.	.
28	75	20	75	17	88	10
28	75	17	75	2	94	10

The above table contains extracts of elements with threshold value <=10.

Step 3: Formation of teams

Step 3.1: Generating two teams of three students per team

Generate two teams as a combination of 3 students per team which are mutually exclusive.

Table 5.1.14: listing of teams with communication values

Team	Stu_id	com	Stu_id	Com	Stu_id	Com	Tot_com_stu	Total_com_team
Team 1	19	6	2	9	23	9	24	51
Team 2	17	10	20	8	5	9	27	
Team 1	19	6	2	9	23	9	24	51
Team 2	17	10	20	8	28	9	27	
.	.		.		.		.	.
Team 1	2	9	20	8	5	9	26	54
Team 2	17	10	28	9	23	9	28	
Team 1	2	9	20	8	19	6	23	51
Team 2	17	10	28	9	23	9	28	
Team 1	2	9	28	9	23	9	27	54
Team 2	17	10	20	8	5	9	27	
Team 1	2	9	28	9	5	9	27	54
Team 2	17	10	20	8	23	9	27	
.	.		.		.		.	.
Team 1	2	9	8	6	23	9	24	51
Team 2	17	10	20	8	28	9	27	

The above table represents combination of two teams (3 students per team) along with communication values represented by com. Tot_com_stu indicates sum total of individual communication marks of each student eg. Tot_com_stu[Team1(row1)]=6+9+9=24. Tot_com_team in the above table shows the sum total of com values of two teams in a group i,e,. tot_com_team=24+27=51.Similarly other values in the table can be explained.

Step 3.2.1: Fixing minimum confidence threshold

Table 5.1.15: Confidence threshold table

Team	Stu_id	Com	Stu_id	Com	Stu_id	Com	Tot_com_stu	Total_com_team
Team 1	2	9	20	8	5	9	26	54
Team 2	17	10	28	9	23	9	28	
Team 1	2	9	28	9	23	9	27	54
Team 2	17	10	20	8	5	9	27	
Team 1	2	9	28	9	5	9	27	54
Team 2	17	10	20	8	23	9	27	

The minimum confidence threshold is taken as 52. From table 3.2.1 it is observed that teams with tot_com_team< minimum confidence threshold(i.e.,52) has been eliminated resulting in final best combinations of two teams per group which are mutually exclusive.

6. RESULTS

- It is found that the number of competent students recognized is 15, obtained by using clustering.
- It was found that 3 groups with two teams each per group were formed as resultants which are considered to be the best combination that could be sent for contests as shown below.

Table 6.1: Final output table

Group	Team	Stu_id	Stu_id	Stu_id
Group A	Team 1	2	20	5
	Team 2	17	28	23
Group B	Team 1	2	28	23
	Team 2	17	20	5
Group C	Team 1	2	28	5
	Team 2	17	20	23

7. CONCLUSION

The main objective was to identify the competent students by clustering using X-means algorithm and then segregating the subjects studied by the student and last but not the least forming better teams to represent the institution. This was achieved using association rule mining. It was found that 15 students emerged as competent students, 3 groups with two teams each per group were formed as a resultant which is considered to be the best combination that could be sent for contests and proposed methodology has an accuracy of 89%. Thus the problem considered was solved.

REFRENCES

[1] U. Kaymak “Fuzzy target selection using RFM variables”, in: Proceedings of the IFSA World Congress and 20th NAFIPS International Conference, vol. 2, 1038–1043, 2001

[2] Dan pelleg, Andrew Moree,"X-means: Extending K-means with Efficient Estimation of the Number of Clusters"

[3] Thomas Laloe, Remi Servien, "The X-Alter algorithm : a parameter-free method to perform unsupervised clustering"

[4] Vance Fabere, “Clustering and the Continuous k-Means Algorithm”, Los Alamos Science, 1994.

[5] RakeshAgrawal, Tomasz Imielinski, Arun Swam, “Mining Association Rules between Sets of Items in Large Databases”, ACM SIGMOD Record, 1993

[6] RokRupnik, MatjažKukar, ‘Data Mining Based Decision Support System to Support Association Rules’, Elektrotehniškivestnik 74(4): 195-200, 2007

[7] AkashRajak and Mahendra Kumar Gupta, ‘Association Rule Mining: Applications in Various Areas’, International Conference on Data Management,2007

[8] Archer-North and Associates, “Performance Appraisal”, http://www.performance-appraisal.com, 2006, Accessed Dec, 2012.

[9] Chein, C., Chen, L., "Data mining to improve personnel selection and enhance human capital: A case study in high technology industry", Expert Systems with Applications, In Press (2006).

[10] Sanders, W. L., & Horn, S. P., Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. Journal of Personnel Evaluation in Education, 12, 247-256, 1998.

[11] K. Pal, and S. Pal, “Analysis and Mining of Educational Data for Predicting the Performance of Students”,(IJECCE) International Journal of Electronics Communication and Computer Engineering, Vol. 4, Issue 5, pp. 1560-1565, ISSN: 2278-4209, 2013.

[12] Z. N. Khan, “Scholastic achievement of higher secondary students in science stream”, Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87, 2005.

[13] B.K. Bharadwaj and S. Pal. “Mining Educational Data to Analyze Students’ Performance”, International Journal of Advance Computer Science and Applications (IJACSA), Vol. 2, No. 6, pp. 63-69, 2011.

[14] M. Bray, The shadow education system: private tutoring and its implications for planners, (2nd ed.),UNESCO, PARIS, France, 2007.

[15] S. K. Yadav, B.K. Bharadwaj and S. Pal, “Data Mining Applications: A comparative study for Predicting Student’s Performance”, International Journal of Innovative Technology and Creative Engineering (IJITCE), Vol. 1, No. 12, pp. 13-19, 2011.