CLUSTERING CRACKERJACK STUDENTS USING DATA MINING APPROACH

Ashok M V[1], Apoorva A[2], Dr. G Suganthi[3]
[1]Head of the department, Dept. of Computer Science,Teachers Academy, Bangalore
[2]Assistant Professor, Department of MCA, Global Institute of Management Sciences, Bangalore
[3]Associate Professor, Dept. of Computer Science, Women's Christian College,
Nagrcoil, Tamil Nadu
Email: ashokmv@ymail.com[1], a.apoorva89@gmail.com[2]

Abstract
Educational Data Mining is a part where in a combination of techniques such as data mining, machine Learning and statistics, is applied on educational data to get valuable information. The objective of this paper is to cluster crackerjack students (skillful) among the students of the educational institution to predict placement chance. Data mining approach is used for Clustering. The technique used to accomplish this task is k-means algorithm based on KSA (knowledge, Communication skill and attitude). To assess the performance of the algorithm, a student data set from an institution in Bangalore were collected for the study as a synthetic data. A model is proposed to arrive at the result. Results obtained from K-means algorithm are compared with results of K-medoids and K-means fast Algorithms with respect to accuracy, precision and recall. From the results obtained it is found that the clusters obtained using K-means algorithm are better in comparison with other algorithms.

Keywords: Educational data mining, crackerjack student, k-means algorithm, K-medoids and K-means fast

1. INTRODUCTION

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

An institution consists of many students. For the students to get placed, he/she should have good score in KSA. KSA is nothing but Knowledge, communication skills and attitude. This is one of the very important criteria for selection of student while placing him. It is also a known fact that better placements results in good admissions. All the students will not have high KSA score. Hence it is necessary to identify those students who possess good KSA and who don't. Thus there is a need for clustering to eliminate students who are not competent to be placed.

2. PROBLEM STATEMENT

Normally hundreds of students will be there in institutions. It is a tedious task and time consuming to predict placement chance for all students and it is not necessary also to predict placement chance for those students who are incompetent academically. Hence there is a need for clustering the crackerjack students having good KSA score whose placement chance can be predicted.

3. 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[10]. Khan, 2005 [11], Baradwaj and Pal,

2011 [12], Bray [13], 2007, S. K. Yadav et al.,2011[14]. K-means is one of the best and accurate clustering algorithms. This has been applied to various problems. K-means approach belongs to one kind of multivariate statistical analysis that cut samples apart into K primitive clusters. This approach or method is especially suitable when the number of observations is more or the data file is enormous .Wu, 2000[1]. K-means method is widely used in segmenting markets. (Kim et al., 2006[2]; Shin &Sohn, 2004 [3]; Jang et al., 2002[4]; Hruschka& Natter, 1999[5]; Leon Bottou et al., 1995 [6]; Vance Fabere et al., 1994[7].

4. PROPOSED MODEL

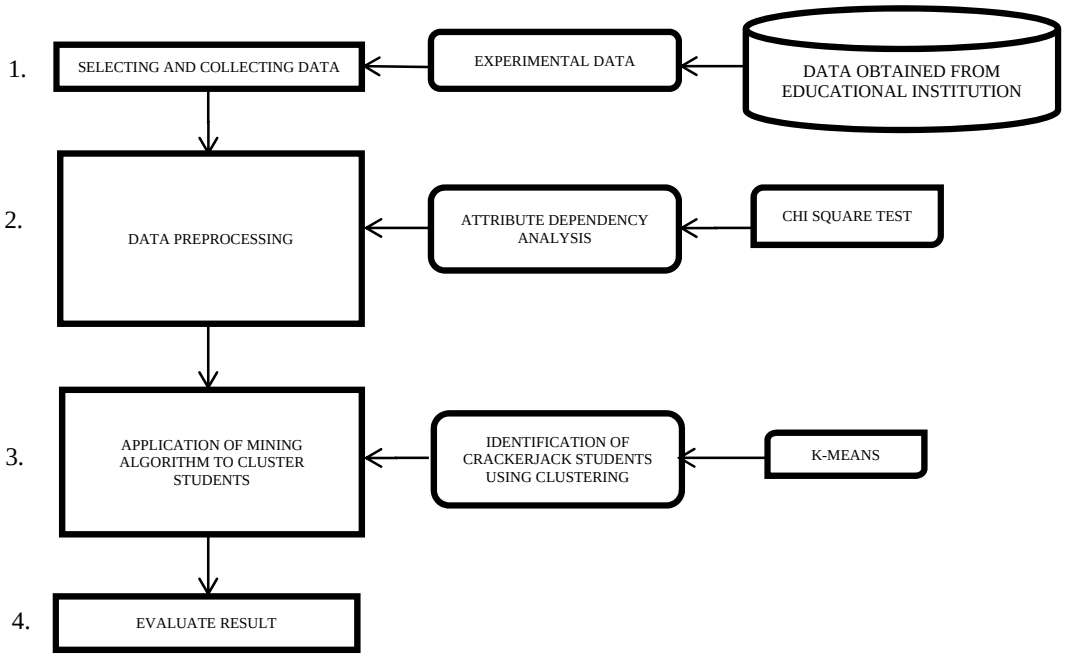

Fig 4.1: Flowchart of the Proposed Model

5. METHODOLOGY

Concept and research framework

The methodology along with its computational processes for determining the crackerjack student, is outlined below:

Step 1: Data collection.

Knowledge represented in terms of Marks scored in selected subjects of a student over a period of three years i.e., from June, 2011 to April, 2014 is considered and collected from an institution in Bangalore.

Step 2: Data preprocessing

Preprocessing was done using chi-square test for the goodness of fit to remove the attributes which doesn't contribute to the result.

Step 3: K-means clustering technique

This step clusters crackerjack students among all the students of the institution using K-means clustering algorithm.

Step 4: Evaluate the result

6. DATA DESCRIPTION

Table 6.1: Database description

Variables	Description	Possible Values
Stu_id	Id of the student	{Int}
Name	Name of the student	{Text}
sub	Subject name	{Text}
M1,M2,M3,M4 …	Marks scored in each subject	{1, 2, 3, 4, 5...100}
T	Total marks	{ 1% - 100% }
Com	(Communication skills+Attitude) score out of 10	{1, 2, 3, 4, 5...10}

Min	Minimum marks for passing a subject	32
Max	Maximum 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.
M1,M2,M3… :–various subject marks scored by a student. It can take only the numeric values from 0 to 100.
T: – 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

7. EXPERIMENTAL EVALUATION

Step 1: Data collection

Table 7.1 : Input Table

stu_id			1	2	3	4	…
Name			vikas	guru	sayed	deepak	…
Sub	Min	Max	M1	M2	M3	M4	M5
Ca	32	100	20	98	45	92	…
Bi	32	100	23	98	69	83	…
Java	32	100	24	97	67	74	…
Se	32	100	25	96	89	92	…
Cf	32	100	26	95	88	88	…
Db	32	100	28	90	56	81	…
…	…	…	…	..	…	…	…
T			624	1910	1416	1482	…
Com			7	9	7	8	…

This is an extract of the student database with the fields or variables listed above. Marks scored in selected subjects of a student over a period of three years i.e., from June, 2011 to April, 2014 is considered and collected from an institution in Bangalore.

Step 2: Data preprocessing:
Preprocessing is done using following statistical technique.
Chi-square test: is applied to remove the useless variable that doesn't contribute to the result. From the above table III name, max and min were removed.

Table 7.2 : Preprocessed table

stu_id	1	2	3	4	…
Sub	M1	M2	M3	M4	M5
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	627	1910	1416	1482	…
Com	7	9	7	8	…

Steps of the k-means algorithm is explained below:

Step 1: Clustering using k-means algorithm.

Step 1: Preprocessed table will be the input for k-means.

Step 2: Cluster crackerjack student segment [PCS] and determine the exact number of clusters. The value of K is incremented in each step and the results are shown below.

Partition of PCS is done initially by taking k=2

After Applying k means clustering with k=2, we have

Table 7.3: Partial view of clusters of students, for k=2

Cluster1	1	10	11	12	13	16	18	21	22	24	26	27	29	30		
Cluster2	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 7.4 : Difference between clusters for k=2

Cluster	**Cluster1**	**Cluster2**
Custer 1	0	0.229
Custer 2	0. 229	0

For k = 2, the distance between the groups are labeled, in this 0.23 is the minimum value.

For k=3 applying k means clustering, we have the following results

Table 7.5 : Partial view of three clusters, for k=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 7.6 : Differences between clusters

Cluster	**Cluster1**	**Cluster2**	**Cluster3**
Custer 1	0	0.116	0.165
Custer 2	0.116	0	0.154
Custer 3	0.165	0.154	0

For k = 3, the distance between the groups are labeled, in this 0.12 is the minimum value

For k=4: we have the following results.

Table 7.7 : Partial view of four clusters, for k=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 7.8 : Comparison of distance between the clusters

Cluster	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Custer 1	0	0.104	0.187	0.342
Custer 2	0. 104	0	0.083	0.238
Custer 3	0.187	0.083	0	0.154
Custer 4	0.342	0.238	0.154	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 k-means, incrementing value of k in every step by 1.

Table 7.9 : 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 1and 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 great 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 k=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: Choosing the cluster

When k takes value 3i.e., k=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).

Step 3: Identifying the elements of the cluster.

Table 7.10 : 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 identified.

8. RESULTS:

From the deduction the cluster 3 is found to be the best cluster having the number of crackerjack students given below:

Table 8.1 : Elements of Cluster 3

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

Results obtained from K-means algorithm are compared with results of K-medoids and K-means fast Algorithms with respect to accuracy, precision and recall.

Table 8.2: list of elements placed correctly and incorrectly in class A and B

ALGORITHM	TA	FA	TB	FB	PRECISION	ACCURACY	RECALL
k-means	429316	34478	495081	68586	0.92	0.8996906	0.86
k-medoids	437508	50440	480293	59220	0.89	0.89327089	0.88
k-means fast	397940	61802	397940	16977 9	0.86	0.77460848	0.81

Above Table represents the elements placed correctly and incorrectly in class A and B with the use of different algorithms. TA and TB represent the elements that are placed correctly in class A and B, whereas FA and FB represent elements that are placed incorrectly in class A and B. Precision and recall are calculated using the formula written above. By the data available in Table it can be observed that the precision and recall values are comparatively better for k-means algorithm.

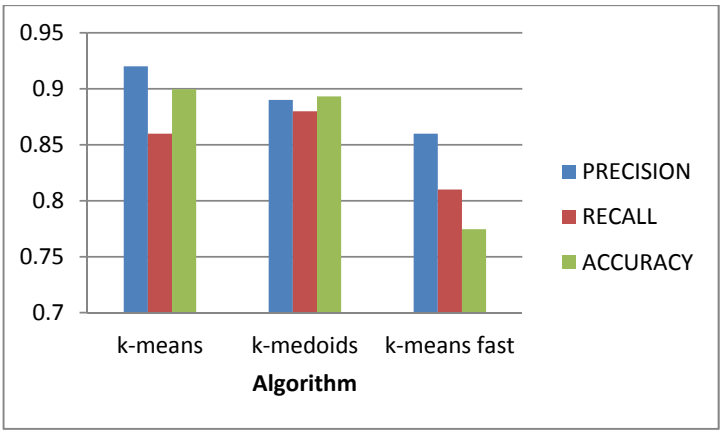

Fig 8.1: Bar graph representing the comparison of Accuracy, Precision and Recall of different algorithms

9. CONCLUSION

The main objective was to identify crackerjack students by clustering using Data mining approach. K-means algorithm was used to accomplish the same. It was found that cluster 3 having crackerjack students emerged among the students of the institution as the best cluster. Results obtained from K-means algorithm are compared with results of K-medoids and K-means fast in terms of accuracy, precision and recall. K-means algorithm stood out with an accuracy of 89%. Thus the solution for the above problem was found successfully.

REFERENCES

[1] Kim, S. Y., Jung, T. S., Suh, E. H., & Hwang, H. S. (2006). Student segmentation and strategy development based on student lifetime value: A case study. Expert Systemswith Applications, 31(1), 101–107.

[2] Shin, H. W., &Sohn, S. Y. (2004). Product differentiation and market segmentation as alternative marketing strategies. Expert Systems with Applications, 27(1), 27–33.

[3] Jang, S. C., Morrison, A. M. T., & O'Leary, J. T. (2002). Benefit segmentation of Japanese pleasure travelers to the USA and Canada: Selecting target markets based on the profitability and the risk of individual market segment. Tourism Management, 23(4), 367–378.

[4] Hruschka, H., & Natter, M. (1999). Comparing performance of feed forward neural nets and k-means of cluster-based market segmentation. European Journal of Operational Research, 114(3), 346–353.

[5] Leon Bottou, YoshuaBengio, "Convergence Properties of the K-Means Algorithms", Advances in Neural Information Processing Systems 7, 1995.

[6] Vance Fabere, "Clustering and the Continuous k-Means Algorithm", Los Alamos Science, 1994.

[7] Rakesh Agrawal, Tomasz Imielinski, Arun Swam, "Mining Association Rules between Sets of Items in Large Databases", ACM SIGMOD Record, 1993

[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] 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.

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

[12] 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.

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

[14] 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.

[15] Ji Wentian, Zhong Sheng,Improved K-medoids Clustering Algorithm under Semantic Web, ICCSEE 2013.

Raied Salman1 ,Vojislav Kecman, Qi Li, Robert Strack and Erik Test, FAST K-MEANS ALGORITHM CLUSTERING , IJCNC, Vol.3, No.4, July 2011.