Indian J.Sci.Res. 14 (2): 467-471, 2017 ISSN: 2250-0138 (Online)

A DATA MINING TECHNIQUES FOR CAMPUS PLACEMENT PREDICTION IN HIGHER EDUCATION

TANSEN PATEL[a1] AND ANAND TAMRAKAR[b]

[ab]SSIPMT, Raipur, Chhattisgarh, India

ABSTRACT

Education Data Mining is very useful in the field of education to predict campus placement of students. Campus placement is a very important concern for any educational organization and wants to improve placement. Educational data mining uses the new technique and tools for discovering the knowledge by analyzing the database to support for prediction process in educational institution. This paper analyze the different data mining techniques and implement data mining technique to enhance prediction for campus placement in any higher education institute. In this paper takes student academic data to predict the standing of placement as an input. Weka software is used for design and implements to making clusters of complete database which is classify the students according to their performance and qualification. The parameters to calculating performance of student including academic performance, communication skills, technical skills, vocational training and projects are measured to ability of a student for placement. The educational institution can predict the campus placement of each student and improve the placement of the organization.

KEYWORDS: Educational Data Mining, WEKA, Cluster, Classification, K-means

In higher educational institution one of the important facts is the growth of educational data. These data are growing fast without any advantage to the organization. The large amount of data uses new techniques and tools for processing of produced data in business processes and extracting some useful information and knowledge are required. Data mining techniques are analytical tools that can be used to extract important knowledge from large data sets. The importance of data mining in higher educational institution proposing new techniques of data mining application in education like placement system and also focused on data mining capability to improve decision making processes in placement system in education institutions.

The importance of this technology better understand through the literature survey that highlights educational system need today and data mining techniques are applied to the current educational system to predict on datasets. The applications of data mining in different fields of human life use the large volumes of data storage in various formats like text documents, images, audio, videos, and many new data formats. For better decision making on data collected from different applications need proper mechanism to extracting knowledge from large database. The main aim of data mining is to discover the useful information from large collections of data [1]. The various functionalities of data mining are applying different technique and algorithms to discover and extract pattern of collected datasets [2]. Recently data mining applications focus and implements due to its imports in decision making and it is useful part in various organizations like educational areas. Data mining fields are nowadays applying in each fields of human life such that some fields are financial data analysis, biological data analysis, retail industry, telecommunication industry, and other scientific applications.

The remainder of this paper is organized as follows. Section II describes overview of some data mining techniques. Section III, we explained some clustering techniques of data mining. The experimental results of this paper are discussed in Section IV. The conclusion is given in section V.

DATA MINING TECHNIQUES

Various algorithms and techniques like Association rules, Classification, Clustering, Regression, Neural Networks, Fuzzy logic, Decision Trees, Genetic Algorithm, are used for knowledge discovery from databases [3]. Some are introduced as follows

Classification

The most commonly useful data mining technique is classification which provides work for a group of pre-classified data to develop a method that can classify the amount of large datasets. Applications like fraud detection, credit risk applications and these similar types of application are particularly analysis by this technique.

Clustering

In data mining, clustering is a technique to grouping a set of objects in such a way that objects in the similar group are more to each other than to those in other groups. Image analysis, machine learning, bioinformatics, pattern recognition, and information retrieval are applications that are applied through this technique. Some important clustering methods are

[1]Corresponding author

K-means method, hierarchical agglomerative method, density based method, farthest first method, and EM method.

Predication

One of the data mining techniques is the prediction that establishes correlation between dependent and independent variables. Predication is adopted by regression technique. There are various regression methods include linear regression, nonlinear regression, multivariate linear regression, and multivariate nonlinear regression

Association rule

Association rule is well-known data mining technique allows in large databases to discover interesting relations between different variables. For example, the rule {Milk, Sugar → Coffee} found in the sales data of a general store would shows that if a customer buys milk and sugar together, that person is expected to also buy coffee. The different association rule methods are quantitative association rule, multilevel association rule, and multidimensional association rule.

Neural Networks

A neural network is a data mining technique that is used to derive complex relationships to find patterns in data to extract patterns for continuous input and output values. Neural networks are useful on particular database having missing, partial, and noisy data. Analyze handwritten character recognition is a one of the application implement by neural network technique.

METHODOLOGY

Clustering is a separation of data into groups of related objects. Cluster consist each group of objects that are similar between themselves and unrelated to objects of other groups. Various clustering techniques have appeared in order to discover interconnected groups in datasets. In this section the techniques for clustering are introduced.

Simple K Means

Simple K Means clustering technique[4] is a simplest unsupervised learning technique. In this technique partition x observations into k clusters with the nearest mean value in which each observation belongs to the cluster. In the beginning, k centroids need to be selected. The next stage is to take instances relates to a data set and correlate them to the nearest centers. After finding k new centroids, a new binding has to be done between the same data set points and the nearest new center. This is repeated until all changes are done. Finally, this technique intend at minimizing intra cluster distance automatically inter cluster distance will be maximized and this is represented through squared error function.

$$\mathrm{SEF} = \sum_{i=0}^{k} \sum_{\alpha \in Ci} ||\alpha - \sigma i||^2$$

Where, σi is a mean of i th cluster, Ci is a i th cluster and σi is a point representing the object.

Farthest First

One of the heuristic based methods of clustering is farthest first [5]. Farthest first is a variant of K Means that select centroids and allocate the objects in cluster but at the point farthest from the existing cluster centre inside the data area. This algorithm provides fast clustering in most of the cases and less modification and reassignment is required.

For each $A_i = [a_{i,1}, a_{i,2}, \ldots, a_{i,n}]$ in database D that is described by n categorical attributes, $f(a_{i,j} \mid D)$ denotes the frequency count of attribute value $a_{i,j}$ in the dataset. A scoring function designed for evaluating each point is defined as

$$\mathrm{SF}(A_i) = \sum_{j=1}^{n} f(a_{i,j} \mid D)$$

Filtered Clusterer

A filter [6] is a special subset of a partially ordered set in mathematics. A filter is the power set of some set, partially ordered by set inclusion. Let S be a topological space and s a point of S. A filter base F on S is said to cluster at s if and only if each element of F has nonempty intersection with each neighborhood of s. If a filter base F clusters at s and is finer than a filter base G, then G clusters at s too. Every limit of a filter base F is also a cluster point of the base.

Hierarchical Clusterer

Hierarchical Clusterer [7] construct a cluster hierarchy, a tree of clusters, known as a dendrogram. One of the types is Agglomerative hierarchical cluster uses bottom up approach and follows the algorithm steps, firstly, start with 1 point. Recursively add two or more appropriate clusters. Stop when k number of clusters is achieved. Second hierarchical cluster is Divisive and reverse of agglomerative that uses top down approach with following algorithms steps. Start with a big cluster. Recursively divides into smaller clusters. Stop when k number of clusters is achieved.

Make Density Based Clusterer

Make Density Based Clusterer [8] is a class for binding a Cluster and returns a density and distribution. The normal and discrete distributions produced by the wrapped clustered will be fitted within each cluster. The algorithmic steps are consider the set of elements E, no.

of clusters K, minimum number of points and max distance for density measure. Initialize k=1, repeat the step if a_i not in cluster then C={ a_j | a_j is density-reachable from a_i. If C is a valid cluster then k=k+1; K_k=C , finally repeat until i=n.

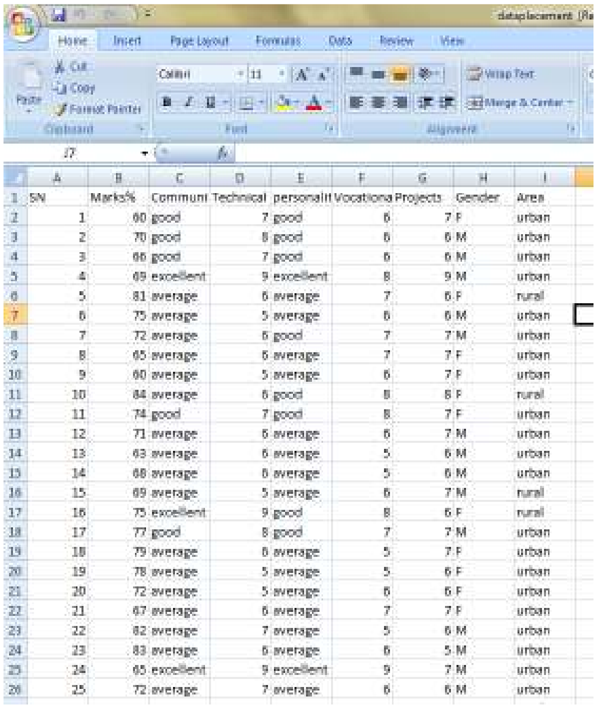
RESULTS AND DISCUSSION

In this section, the placement data of 150 final year BE students of Engineering College is collected as an input and experiments these test dataset in Weka tool for clustering.

Table I: List of attributes of datasets

Attributes	Values
Marks%	0-100(%)
Communication skills	Average, Good, Excellent
Technical skills	0-10
Personality	Average, Good, Excellent
Vocational Training	0-10
Projects	0-10
Gender	F(Female), M(Male)
Area	Rural, Urban

To implementing the various algorithm on the given data set required to convert excel file to csv file and get the required csv file for processing in weka tools.

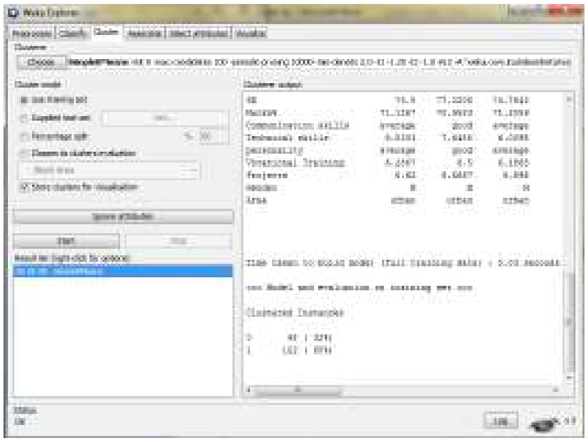


	A	B	C	D	E	F	G	H	I
1	SN	Marks%	Communi	Technical	personalit	Vocationa	Projects	Gender	Area
2	1	60	good	7	good	6	7	F	urban
3	2	70	good	8	good	6	6	M	urban
4	3	66	good	7	good	6	6	M	urban
5	4	69	excellent	9	excellent	8	9	M	urban
6	5	81	average	6	average	7	6	F	rural
7	6	75	average	5	average	6	6	M	urban
8	7	72	average	6	good	7	7	M	urban
9	8	65	average	6	average	7	7	F	urban
10	9	60	average	5	average	6	7	F	urban
11	10	84	average	6	good	8	8	F	rural
12	11	74	good	7	good	8	7	F	urban
13	12	71	average	6	average	6	7	M	urban
14	13	63	average	6	average	5	6	M	urban
15	14	68	average	6	average	5	6	M	urban
16	15	69	average	5	average	6	7	M	rural
17	16	75	excellent	9	good	8	6	F	rural
18	17	77	good	8	good	7	7	M	urban
19	18	79	average	6	average	5	7	F	urban
20	19	78	average	5	average	5	6	F	urban
21	20	72	average	5	average	6	6	F	urban
22	21	67	average	6	average	7	7	F	urban
23	22	82	average	7	average	5	6	M	urban
24	23	83	average	6	average	6	5	M	urban
25	24	65	excellent	9	excellent	9	7	M	urban
26	25	72	average	7	average	6	6	M	urban

Figure 1: CSV file of data placement

After the collection of desired data we apply various algorithms on the student's placement data set using weka tool. We experiment Simple K means, Farthest First, Filtered Clusterer, Hierarchical Clusterer and Make Density Based Clusterer algorithms on the student placement data set using weka tool.

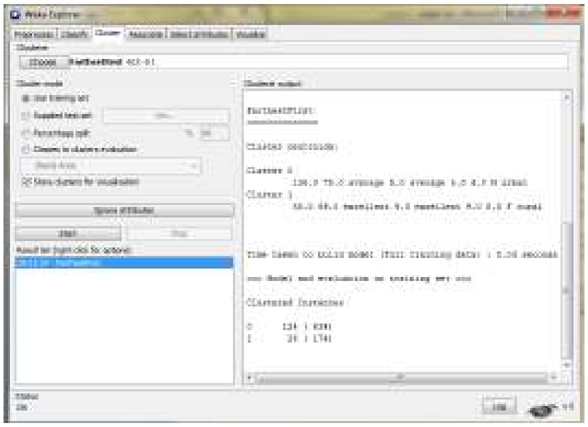
Experimental performances of each clustering algorithm are performed in weka tools with the following results.



SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance

```
SN                        [illegible]  [illegible]  [illegible]
Marks%                    [illegible]  [illegible]  [illegible]
Communication skills      average      good         average
Technical skills          [illegible]  [illegible]  [illegible]
personality               average      good         average
Vocational Training       [illegible]  [illegible]  [illegible]
Projects                  [illegible]  [illegible]  [illegible]
Gender                    M            M            M
Area                      urban        urban        urban


Time taken to build model (full training data) : [illegible] seconds

=== Model and evaluation on training set ===

Clustered Instances

0       48 ( 32%)
1      102 ( 68%)
```

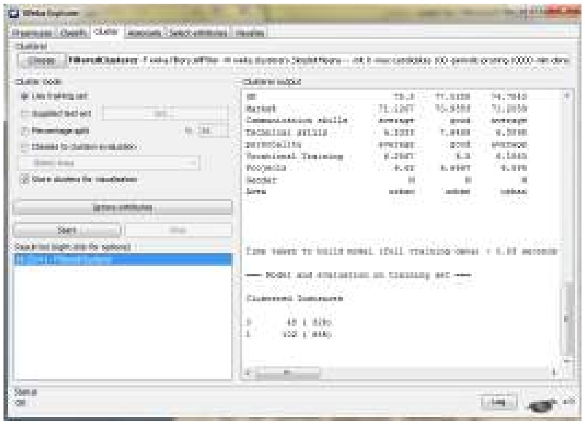
Figure 2: Experimental result of Simple K Means cluster



FarthestFirst -N 2 -S 1

```
FarthestFirst
==============

Cluster centroids:

Cluster 0
        138.0 75.0 average 5.0 average 5.0 6.0 M urban
Cluster 1
        53.0 65.0 excellent 9.0 excellent 8.0 8.0 F rural



Time taken to build model (full training data) : [illegible] seconds

=== Model and evaluation on training set ===

Clustered Instances

0      124 ( 83%)
1       26 ( 17%)
```

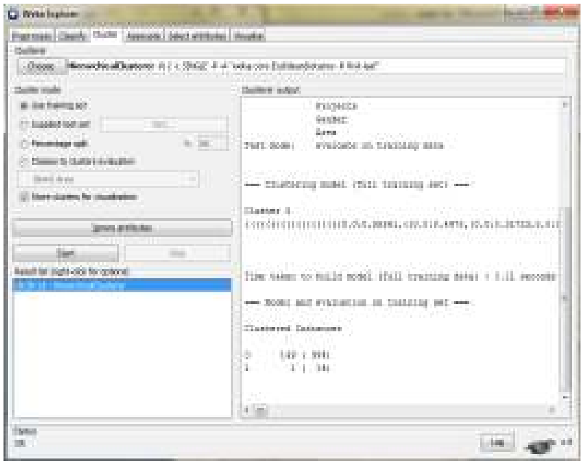
Figure 3: Experimental result of Farthest First



FilteredClusterer -F weka.filters.AllFilter -W weka.clusterers.SimpleKMeans -- -init 0 -max-candidates 100 -periodic-pruning 10000 -min-dens

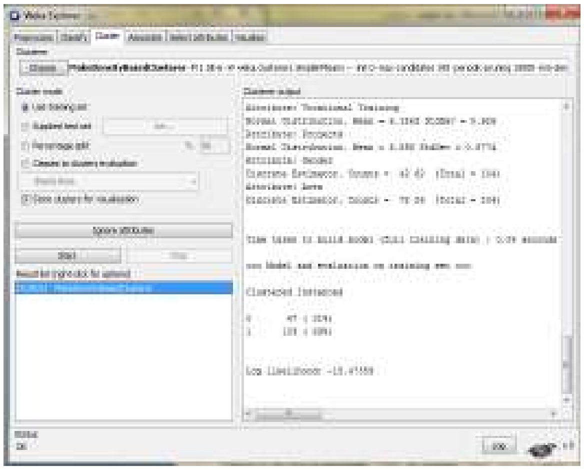
```
SN                        [illegible]  [illegible]  [illegible]
Marks%                    [illegible]  [illegible]  [illegible]
Communication skills      average      good         average
Technical skills          [illegible]  [illegible]  [illegible]
personality               average      good         average
Vocational Training       [illegible]  [illegible]  [illegible]
Projects                  [illegible]  [illegible]  [illegible]
Gender                    M            M            M
Area                      urban        urban        urban


Time taken to build model (full training data) : [illegible] seconds

=== Model and evaluation on training set ===

Clustered Instances

0       48 ( 32%)
1      102 ( 68%)
```

Figure 4: Experimental result of Filtered Clusterer



```
              Projects
              Gender
              Area
Test mode:    evaluate on training data


=== Clustering model (full training set) ===

Cluster 0
[illegible]



Time taken to build model (full training data) : [illegible] seconds

=== Model and evaluation on training set ===

Clustered Instances

0       [illegible] ( 99%)
1         1 (  1%)
```

Figure 5: Experimental result of Hierarchical Clusterer



```
Attribute: Technical Training
Normal Distribution. Mean = [illegible] StdDev = [illegible]
Attribute: Projects
Normal Distribution. Mean = [illegible] StdDev = [illegible]
Attribute: Gender
Discrete Estimator. Counts = [illegible] (Total = [illegible])
Attribute: Area
Discrete Estimator. Counts = [illegible] (Total = [illegible])


Time taken to build model (full training data) : [illegible] seconds

=== Model and evaluation on training set ===

Clustered Instances

0       47 ( 31%)
1      [illegible] ( 69%)


Log likelihood: -15.[illegible]
```

Figure 6: Experimental result of Make Density Based Clusterer

After implementing the above clustering algorithm we analysis the clustering algorithms performance in the term of execution time of complete datasets and clustered instances in the form of percentage. Finally, we compare the clustering algorithms Simple K Means, Farthest First, Filtered Clusterer, Hierarchical Clusterer and Make Density Based Clusterer.

Table II: Comparison of cluster algorithms

Cluster Algorithms	Clustered Instances (in %)		Time taken to build model (in Seconds)
	0	1	
SimpleKMeans	32	68	0.02
FarthestFirst	83	17	0.02
FilteredClusterer	32	68	0.02
HierarchicalClusterer	99	1	0.09
MakeDensityBasedClusterer	31	69	0.08

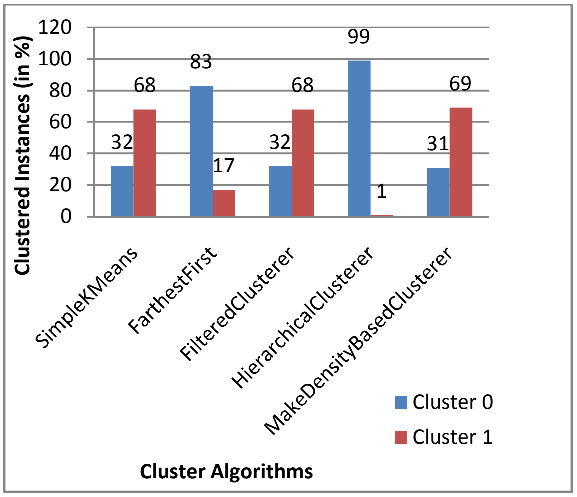


Figure 7: Clustered Instances

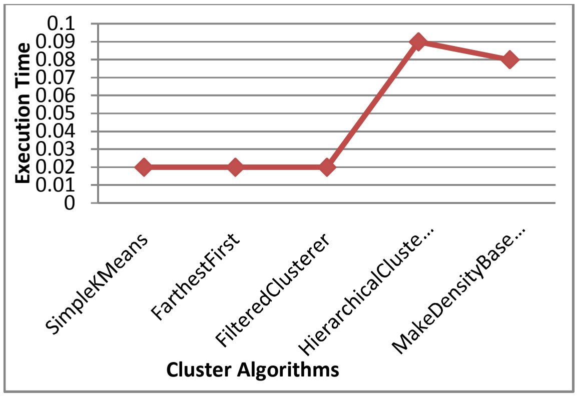


Figure 8: Execution time

CONCLUSION

The experimental results of clustering algorithms on student's placement datasets is performed here. The performance of the various clustering algorithms is compared based on clustered instances and time taken to form the estimated clusters. In the cluster 0, Hierarchical Clusterer algorithm gets highest percentage and Make Density Based Clusterer with lowest percentage. In case of cluster 1, it is reverse performance of cluster 0. Both Simple K Means and Filtered Clusterer are same clustered instances value. In the terms of execution time Hierarchical Clusterer takes maximum time as compare to others. The equal execution time taken by Simple K means, Farthest First and Filtered Clusterer algorithm for given datasets. These results can be used in future for similar type of research work.

REFERENCES

Heikki, Mannila, "Data mining: machine learning, statistics, and databases", IEEE, 1996.

Fayadd U., Piatesky -Shapiro G., and Smyth P., "Data Mining To Knowledge Discovery in Databases", AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–262 56097–6 Faya, 1996.

M. Bharati,Ramageri, "Data Mining Techniques and Applications", Indian Journal of Computer Science and Engineering, Vol. 1, No. 4, pp. 301-305, 2010.

J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques. Morgan kaufmann, 2006.

M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in Procs. of the twenty-first international conference on Machine learning, p. 11, 2004.

Priyanka Sharma, "Comparative Analysis of Various Clustering Algorithms Using WEKA". IRJET, e-ISSN: 2395 -0056, Volume 2, Issue 4, July-2015.

I. Jonyer, D. J. Cook, and L. B. Holder, "Graph-based hierarchical conceptual clustering," Journal of Machine Learning Research 2, 19-43, 2001.

Shrimani P.K. and Koti M.S. "EVALUATION OF PRINCIPAL COMPONENTS ANALYSIS (PCA) AND DATA CLUSTERING TECHNIQUES (DCT) ON MEDICAL DATA." International Journal of Knowledge Engineering 3.2, 202-206, 2012.