Performance of Data Mining Technique in Education Sector

Johina

Student of Masters of Technology, Department of Computer Science and Engineering,
JCDM college of Engineering Sirsa, GJU, Hisar, Haryana, India
ollajohina@gmail.com

Vikas Kamra

Assistant Professor, Department of Computer Science and Engineering,
JCDM College of Engineering Sirsa, GJU, Hisar, Haryana, India
kamra.vikas@yahoo.com

Abstract:

The most commonly use of data mining technique is to analysis the data from different source and finally summarizes it into useful information. This paper provides comparative study and performance analysis of commonly used data mining technique such as classification and clustering used for education system. Data mining provides many software to analysis the technique. But we are using the weka tool for this purpose. The weka is Waikato Environment for Knowledge Analysis is introduced by university of New Zealand. The weka is data mining tool and this paper describe how to use weka for these technologies. It classifies and clusters the data through various algorithms. The main objective is to find out best algorithm to produce the accuracy and easy to use for the educational data [1].

Keywords: weak tool; result for classification and clustering etc.

I. INTRODUCTION

Weka stand for Waikato Environment for knowledge analysis is a toolkit developed by Waikato University, New Zealand fig.1. Weka is mainly used to analyze the data mining algorithm .Weka is open source software and developed in java and useful for education, research and projects. The weka uses the ARFF or CSV file format. The weka can run on any platform such as widows, Linux, Mac etc. Wek is free available under the GNU (general public license). Weka is collection of algorithm for data mining task. The algorithm can be applied on a datasets or can be used with java code. By using the java connectivity weka provide access to SQL database and the result returned by the database query .Weka contains many tool for pre-processing, classification, clustering, associations rule and visualization also. WEKA consists of [2]:

- Explorer
- Experimenter
- Knowledge flow
- Simple Command Line Interface

The main user interface of weka is the Explorer, but knowledge flow and command line can be used to access the same functionality.

Fig. 1 weka tool

Explorer: The explorer provides an environment to explore the data with weka. The explorer contains the six elements as follow:

- Pre-process
- Classify
- Cluster
- Associate
- Select attributes
- Visualize

Experimenter: It provides an environment to perform experiments and allow users to run, create, modify and analyze the experiments in more comfort manner when processing individually. It contains the following elements:

- Setup
- Run
- Analyse

Knowledge flow: The knowledge flow performs the same function as Explorer but with a drag and drop option. It also performs the incremental learning.

Simple CLI: The simple CLI allow the direct execution of weka commands. It provides a simple command line interface for the operating system that does not have their own command line interface.

Steps Involve In Weka:

The weka performs the following steps as follows:

- Load the data file with .arff or CSV format in the pre-process tab.
- Select the type of algorithms such as classification, clustering or association
- Weka produces the output based upon the algorithm you select.
- Finally analyze the results produced by the weka.

II CLASSIFICATION

Classification is most commonly used data mining technique to develop class and assign each object to the particular class. The classification involves two steps first is it builds the models and model is represented by decision rules, classification rule or mathematical formulae. Seconds step is model usage. It is used to classify unknown object and check the accuracy of the model. Types of classification models:

- Bayesian classification
- Rule based Classification (decision table)
- Support vector machines (SVM)
- Decision tree
- Neural networks

A. Bayesian Networks

A Bayesian network is used show the relationship between a set of variables. Bayesian networks are represented by the directed acyclic graph in which the node represents the domain variable and the arc between the node show the dependency. The mainly use of the Bayesian network is to calculate the probability of one node, value assigned by the other node [3]. The Bayesian network performs the two tasks as follow:

- *BN learning*: the BN learning is used to get a model.
- *BN Interface*: The BN interface is used to classify the instances.

B. Decision Table (DT):

The decision table is also called as logical tree is used to show the tabular form of conditions and actions. It is used for the decision making process. The decision table represents the all possible conditions and resulting action in matrix form in Fig. 2.

Problem area	
CONDITION SET	CONDITION SPACE
ACTION SET	ACTION SPACE

Fig 2: decision table

DT contains the four things as follows:

- *Condition Set:* it contains all the condition that are used for the decision making process.
- *Condition Space*: it shows the all combination of possible conditions states and the state can be unlimited.
- *Action Set*: it contains all action like output, conclusions that a decision maker take.
- *Action Space*: it contains the all possible states of actions and the state can be unlimited.

C. LAD Tree:

Lad tree is logical analysis of data and make classifier for binary variables based upon the logic that can difference between the negative and positive data in the dataset. Lad is method which combines the optimization, Boolean function. For a given dataset the lad model construction generate a pattern and then select a subset of them and assure that that selected pattern are according to need in term of prevalence and homogeneity .

D. J48 Tree:

It builds the decision tree from labelled training data set using information gain and examines the same that results from choosing an attribute for splitting the data. To make the decision the attribute with highest normalized information gain is used. Then the algorithm recurs on smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then the leaf node is created in a decision tree telling to choose that class [4].

III COMPARISON OF DIFFERENT CLASSIFICATION TECHNIQUES

Here we compare different-2 techniques of classification in term of their mean absolute error, root mean squared error, relative absolute error.

A. Comparison For Mean Absolute Error:

According to this parameter we find out the mean absolute error rate by the different technique. The table show the weka output of different technique. The mean error rate of different technique is according to different datasets. Now we are describe the experimental results which is obtained from the various classification techniques and comparison with each other.

Table No 1: Comparative result of classification techniques.

Technique/Datasets	Sample dataset 1	Sample dataset 2	Sample dataset 3	Average Error
Bayesian networks	0.0222	0.004	0.3537	0.126633
Decision table	0.0509	0.0111	0.3596	0.140533
Lad tree	0.0099	0.0012	0.3046	0.105233
J48 tree	0.0096	0.0005	0.3435	0.117866

The best techniques identified from each classifier then compared with other classifiers to discover what classifier is best to be used for classification of different dataset. We used here these techniques for the comparison on the different dataset and find the best techniques.

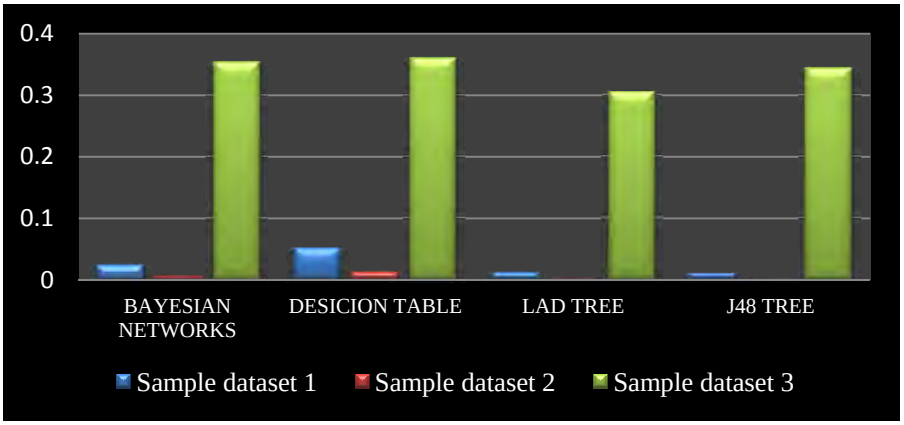

Fig.3: Comparison between parameters for Mean Absolute Error.

Based on the above Figure No. 3 and Table No.1 we can clearly see that the highest Mean Absolute Error is 0.3537 and lowest accuracy is 0.004 in the Bayesian Network classifiers. And the highest Mean Absolute Error is 0.3596 and lowest Mean Absolute Error 0.0111 in the Decision Table classifiers and the highest Mean Absolute Error is 0.3046 and lowest is 0.0012 in the Lad tree classifiers. The highest Mean Absolute Error is 0.3435 and lowest is 0.0005 in the J48 tree classifiers.

1) *Average Error Rate*:

(Error in Sample Dataset1 + Error in Sample Dataset 2 + ... Error in Sample Dataset N)/N

Fig.4: Average Error Rate for Mean Absolute Error.

From the above Fig. 4we can clearly see that the average error rate of LAD tree classifier is the best among these four classifier techniques.

B. Comparison For Root Mean Squared Error:

According to this parameter we find out the root mean squared error by the different technique. The table show the weka output of different technique. The root mean squared error of different technique is according to different datasets.

Table No.2: Comparative result of classification techniques

Technique/ Datasets	Sample dataset 1	Sample dataset 2	Sample dataset 3	Average Error
Bayesian networks	0.1066	0.0177	0.4228	0.18236667
Decision table	0.1086	0.0141	0.4191	0.1806
Lad tree	0.0913	0.227	0.4084	0.24223333
J48 tree	0.0922	0.0226	0.4144	0.1764

Based on the Fig. 5 and Table No.2 we can clearly see that the highest Root Mean Squared error is 0.4228 and lowest Root Mean Squared error is 0.1066 in the Bayesian Network classifiers. And the highest Root Mean Squared error is 0.4191 and lowest Root Mean Squared error is 0.0141 in the Decision Table classifiers and the highest Root Mean Squared error is 0.4084 and lowest is 0.0913 in the Lad tree classifiers. The highest Root Mean Squared error is 0.4144 and lowest is 0.0226 in the J48tree classifiers.

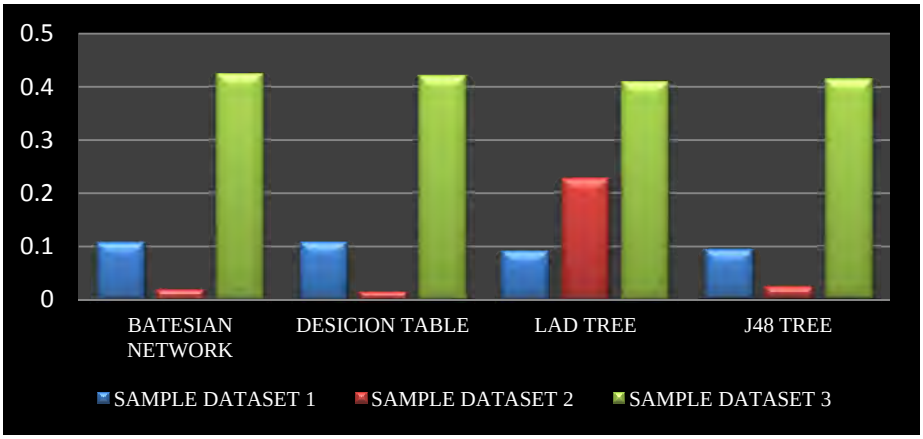

Fig. 5: Comparison between parameters for Root Mean Squared Error

1) Average *Error Rate*:

(Error in Sample Dataset1 + Error in Sample Dataset 2 + ….. Error in Sample Dataset N)/N

Fig. 6 Average Error Rate for Root Mean Squared Error

From the above graph Fig. 6 we can clearly see that the Average Error Rate of Root Mean Squared error rate of J48 tree classifier is the best among these four classifier techniques.

C. Comparison For Relative Absolute Error:

According to this parameter we find out relative absolute error rate by the different technique. The table show the weka output of different technique. The relative absolute error rate of different technique is according to different datasets.

Table No.3:.Comparison between parameters for Relative Absolute Error

Technique/Datasets	Sample Dataset 1	Sample Dataset 2	Sample Dataset 3	Average Error
Bayesian Networks	0.105627	0.012626	1.02679	0.381681
Decision Table	0.241876	0.034878	1.043783	0.440179
LAD Tree	0.046811	0.003729	0.884381	0.31164
J48 Tree	0.045665	0.001616	0.997042	0.348108

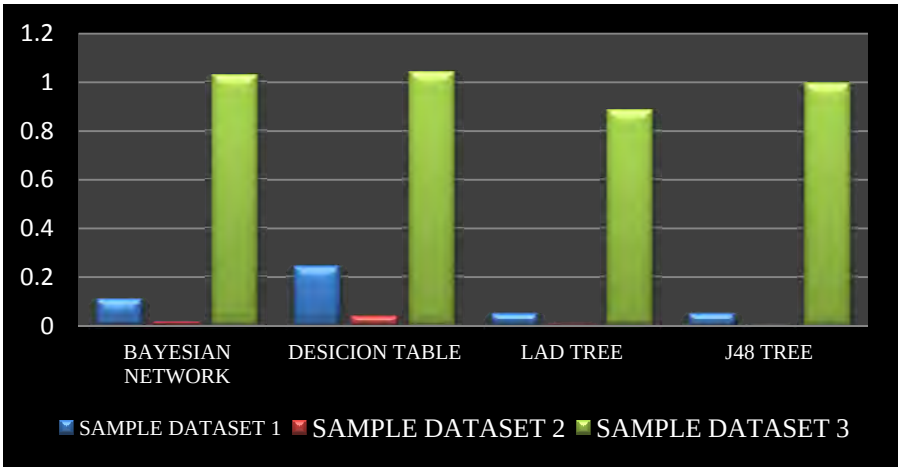

Fig.7: Comparisons between Parameters for Relative Absolute Error

Based on the Fig. 7 and Table No.3 we can clearly see that the highest Relative absolute error is 1.02679 and lowest is 0.012626 in the Bayesian Network classifiers. And the highest Relative absolute error is 1.043783 and lowest is 0.034878 in the Decision Table classifiers and the highest Relative absolute error is0. 884381 and lowest is 0.003729 in the Lad tree classifiers and for J48 Tree the highest 0.997042 and lowest is 0.001616.

1) Average Error Rate:

(Error in Sample Dataset1 + Error in Sample Dataset 2 + ….. Error in Sample Dataset N)/N

Fig. 8: Average Error Rate for Relative Absolute Error

From the above graph Fig.8 we can clearly see that the average error rate of Relative absolute error rate of Lad tree classifier is the best among these four classifier techniques.

IV CLUSTERING

Clustering means to group the data object according to their similarity. First of all the data set are divided into the group according to similarity and then assign the label to the small number of group. Various clustering algorithm are developed for a good performance on different dataset for cluster formation.Several clustering techniques are

- Partitioning Clustering (K-Mean)
- Hierarchical Clustering
- Density Based Clustering(DBSCAN)
- Farthest First Clustering

A. *K-Means Clustering*

It is a centroids based technique. This algorithm takes the input parameters k and partition a set of n objects into k clusters that the resulting intra-cluster similarity is high but the inter-cluster similarity is low. The method can be used by cluster to assign rank values to the cluster categorical data is statistical method. K mean is mainly based on the distance between the object and the cluster mean. Then it computes the new mean for each cluster [5].

B. *Hierarchical Clustering*

The hierarchical clustering creates a tree structure of a given dataset. This technique produces a sequence of in which all inclusive cluster at the top and cluster of individual points are at bottom. There are two types of hierarchical clustering first is top down and second is bottom down. The top down start with all object in a cluster and spilt into smaller cluster. The bottom down hierarchical clustering uses the many clusters as objects and then these cluster are combined until only one cluster remains [6].

C. *Farthest First Clustering*

The farthest first clustering is most suitable for the large dataset. It is the modified version of the K-mean. In this algorithm the distance one centroid is maximum from the other and not finds the mean for calculating the centroid. The Fig 9 show the farthest first clustering [7].

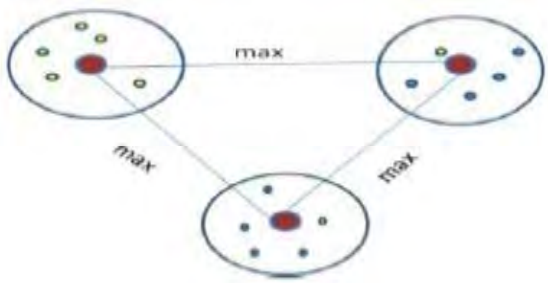

Fig. 9: Show the Farthest First Clustering

D. Density Based Clustering

One of the most used density based clustering is a DBSCAN. In this technique the clustering is based on the density of data point in a region and each cluster has a higher density of point than out side of the higher. The DBSCAN divide the data point in to three classes such as [8]:

- Core point: the interior point of a cluster is known as core point.
- Border points: These points are falls with in neighborhood of a core point.
- Noise point: the point which is not core point and border point are noise points.

The DBSCAN can find the arbitrarily shaped cluster. The DBSCAN uses the two parameters are Eps and Minpts. The basic idea of DBSCAN algorithm is that for each object of a cluster, the neighbourhood of a given radius (*Eps*) has to contain at least a minimum number of objects (*MinPts*).

V. COMPARISON OF DIFFERENT CLUSTERING TECHNIQUES

Here we compare different-2 techniques of clustering in term of no. of iteration to build the model and clustered instances.

A. According To No. Of Iteration.

According to this parameter we find out that no. of iteration to build the model by the different technique. The table show the weka output of different technique. The no. of iteration taken by different technique is according to different datasets.

Table No.4: comparisons according to No. of Iteration to build the model

Technique/ Datasets	Sample Dataset 4	Sample Dataset 5	Sample Dataset 6
K-Means Clustering	16	7	3
Hierarchical Clustering	-	-	-
DBSCAN Clustering	-	-	-
Farthest First Clustering	-	-	-

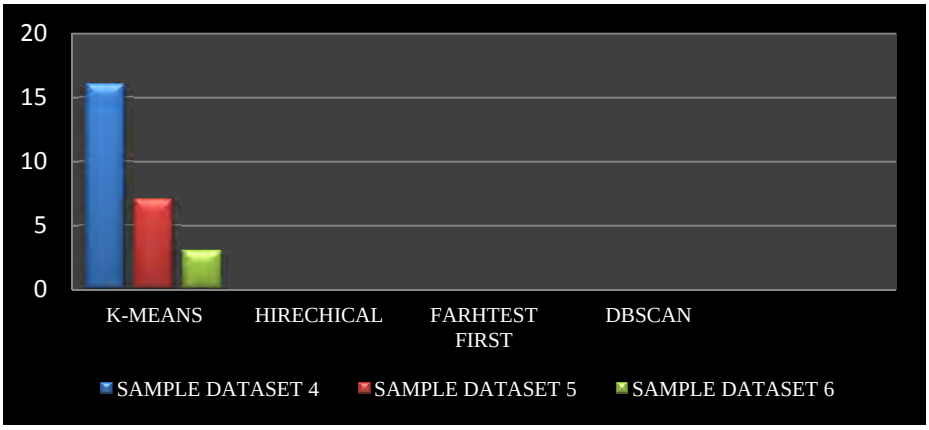

Fig 10: Compared According To No. of Iteration

According to this result, the highest no. of iteration taken to build the model by K-Mean is 16 and lowest is 3 and the other three technique takes 0 iteration for all the datasets. Based upon the result shown in Table No.4 and Fig 10 we can say k-mean takes more iteration to build the model than the others techniques.

B. *According To Clustered Distribution*:
According to this parameter we find out that no. of cluster are created by the different technique and partition of data according to different cluster. The table show the weka output of different technique.

1) Instances in Cluster 1:

The data grouped in cluster 1 by different technique is shown in table no.5.

Table No 5: Comparison According To Instances in Cluster 1

Technique/ Datasets	Sample Dataset 4	Sample Dataset 5	Sample Dataset 6
K-Means Clustering	70%	49%	64%
Hierarchical Clustering	92%	100%	100%
DBSCAN Clustering	0	0	0
Farthest First Clustering	82%	64%	57%

Based upon the Table No.5.and Fig 11 the highest instance in cluster1 by K-Mean is 70% and lowest is 49% and for hierarchical the highest is 100% and lowest is 92% and for DBSAN the instance in cluster1 is 0 but for the farthest first the highest clustered instances is 82% and the lowest is 57% So based are upon these results we can say that the hierarchical technique is very much better than the others.

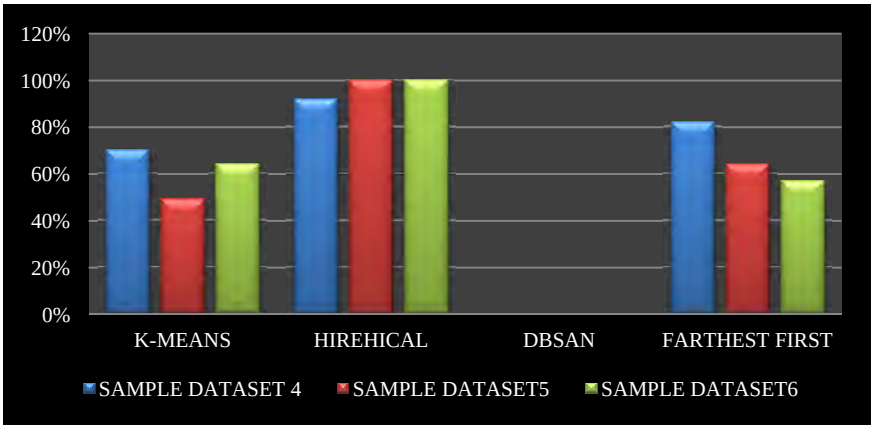

Fig11: Compared According To Instances in Cluster 1

2) Instances in Cluster 2

The data grouped in cluster 2 by different technique is shown in table no.

Table No 6: Comparison According To Instances in Cluster 2

Technique/ Datasets	Sample Dataset 4	Sample Dataset 5	Sample Dataset 6
K-Means Clustering	30%	51%	36%
Hierarchical Clustering	8%	0%	0%
DBSCAN Clustering	-	-	-
Farthest First Clustering	18%	36%	43%

Based upon the Table No.6 and Fig 12 the highest instance in cluster 2 by K-Mean is 51% and lowest is 30% and for hierarchical the highest is 8% and lowest is 0% and for the farthest first the highest clustered instances is 43% and the lowest is 18% .So based are upon these results we can say that the hierarchical technique is very much better than the others.

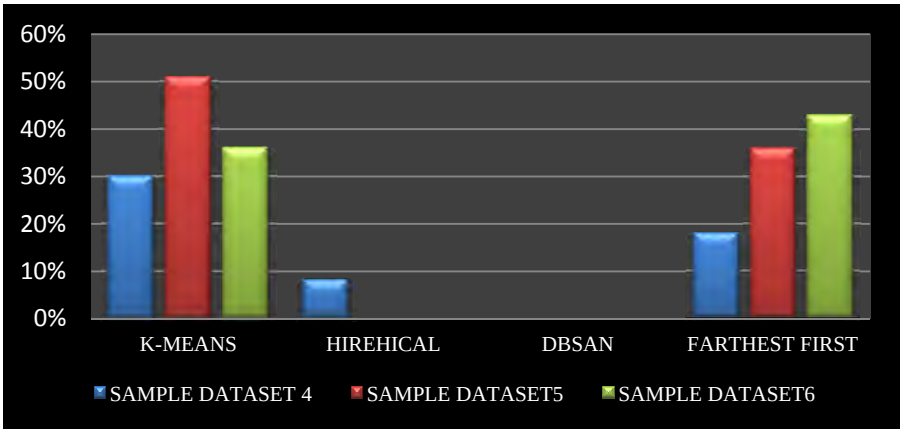

Fig 12: Compared According To Instances in Cluster 2

VI. FINAL RESULT FOR CLASSIFICATION TECHNIQUE

To analyze the performance of the selected classification methods or algorithms namely as Bayesian Network, Decision table, Lad tree and J48 tree. We use the 3 different datasets for this purpose. Here we find performance of Accuracy, Mean Absolute Error, Root mean squared error, Relative absolute error etc. from different education datasets as follows.

Table No. 7: Comparison for different parameters of classification technique

Technique/ Parameters	Average Mean Absolute Error	Average Root Mean Squared Error	Average Relative Absolute Error	Average Correctly classified
Bayesian Networks	0.12663333	0.18236667	0.381681	373.6667
Decision Table	0.14053333	0.1806	0.440179	392.3333
LAD Tree	0.10523333	0.24223333	0.31164	394.6667
J48 Tree	0.11786667	0.1764	0.348108	390

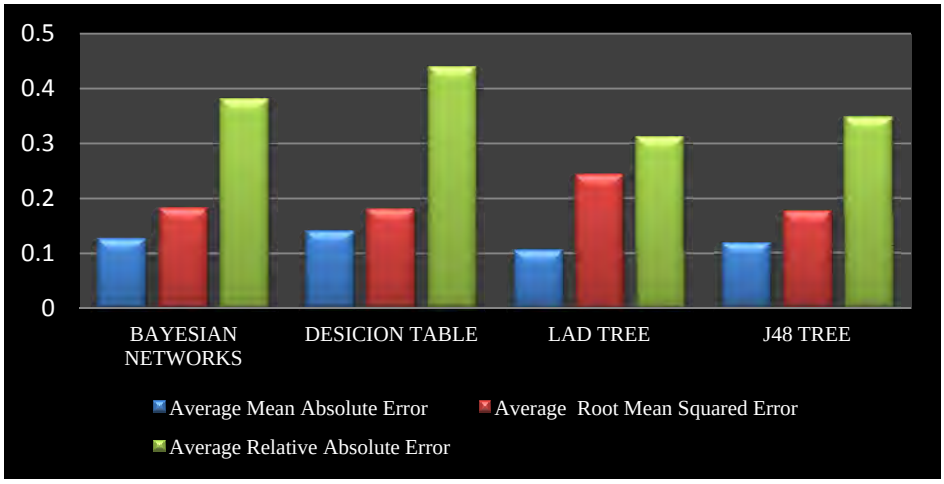

Fig 13: Average error of different classification technique

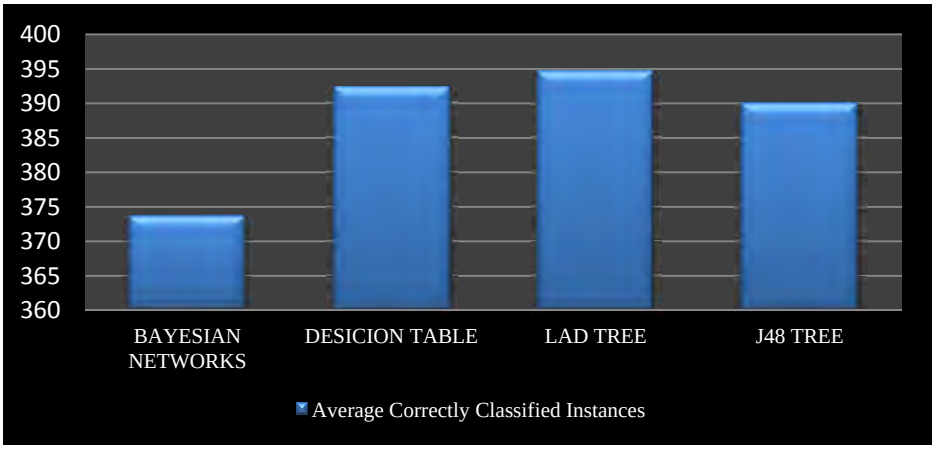

Fig 14 :Average correctly classified Instance

From the above graph Fig 13 we can clearly see that the according to average error rate of mean absolute error relative absolute error the LAD tree best and From Fig 14 we can see Average correctly classified instances of the LAD tree classifier is the best among these four classifier techniques. So that we can say that the LAD tree classifier is the best for the education datasets.

VII. RESULT FOR CLUSTERING TECHNIQUE

To analyze the performance of the selected clusters methods or technique namely as K-mean, Hierarchical, DBSCAN and Farthest First clustering. We use the 3 different datasets for this purpose. Here we have taken different types of sample datasets to find performance no of iteration and correctly clustered instances from different education datasets.

Table No. 8: Comparison for different parameters of clustering technique

Technique/ Parameters	Maximum No. of Iteration	Average Instances In Cluster 1	Average Instances In Cluster 2
K-Means Clustering	16	61%	39%
Hierarchical Clustering	0	97.33333%	2.666667%
DBSCAN Clustering	0	0%	0%
Farthest First Clustering	0	67.66667%	32.33333%

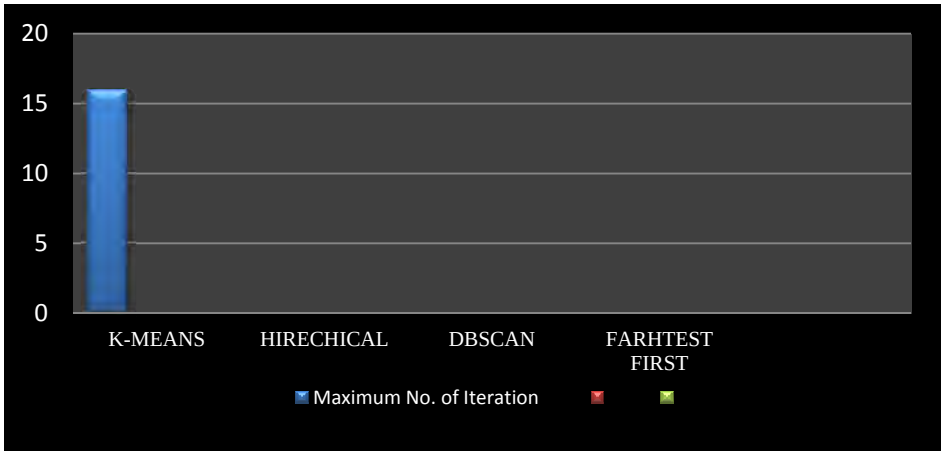

Fig 15 Maximum No. of Iteration

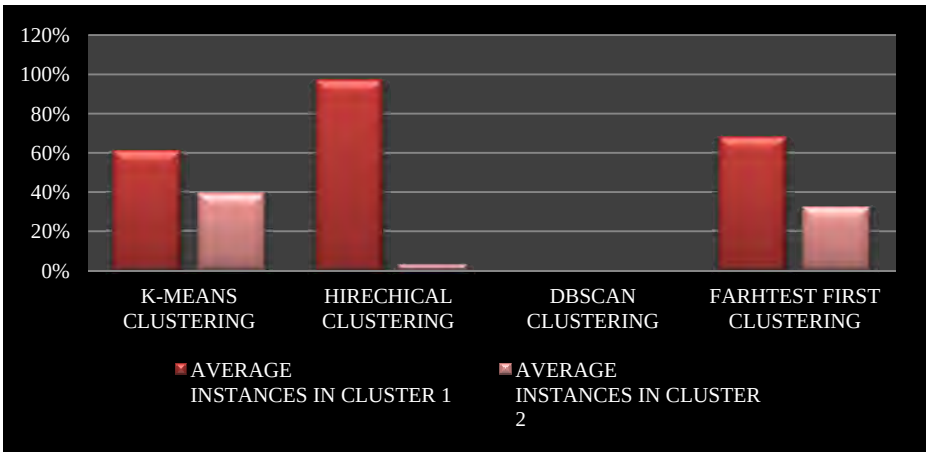

Fig 16: Average Instances According To Clusters

From the above graph Fig 16 according to maximum no of iteration taken by k-mean and Fig 16 we can clearly see that Average clustered instances the are best in the hierarchical clustering and also takes no iteration so that it is the best among these four clustering techniques for education datasets.

VIII CONCLUSION

As a conclusion, we have met our objective which is to evaluate and investigate four Classification and Clustering technique based on Weka. The best algorithm based on education dataset is LAD tree classifier amongst these four classification techniques according to Mean Absolute Error and Relative Absolute Error and also give the best accuracy among all the four technique s and hierarchical is best among clustering methods. These results suggest that among the machine learning techniques tested, LAD Tree classifier and hierarchical has the potential to significantly improve the conventional classification and clustering methods for use in general education field.

REFERENCES:

[1] .M.Vijayakamal, Mulugu Narendhar "A Novel Approach For Weka & Study On Data Mining Tools" International Journal Of Engineering And Innovative Technology (Ijeit) Volume 2, Issue 2, August 2012

[2] Swasti Singhal, Monika Jena "A Study On Weka Tool For Data Preprocessing, Classification And Clustering" International Journal Of Innovative Technology And Exploring Engineering (Ijitee) Issn: 2278-3075, Volume-2, Issue-6, May 2013

[3] Sunita Beniwal*, Jitender Arora "Classification And Feature Selection Techniques In Data Mining" International Journal Of Engineering Research & Technology (Ijert) Vol. 1 Issue 6, August – 2012 Issn: 2278-0181

[4] Trilok Chand Sharma1, Manoj Jain2 "Weka Approach For Comparative Study Of Classification Algorithm" International Journal Of Advanced Research In Computer And Communication Engineering Vol. 2, Issue 4, April 2013

[5] Pallavi #, Sunila Godara *" A Comparative Performance Analysis Of Clustering Algorithms" Pallavi , Sunila Godara / International Journal Of Engineering Research And Applications (Ijera) Issn: 2248-9622 Vol. 1, Issue 3, Pp.441-445

[6] Aastha Joshi" A Review: Comparative Study Of Various Clustering Techniques In Data Mining" International Journal Of Advanced Research In Computer Science And Software Engineering" Volume 3, Issue 3, March 2013 Issn: 2277 128x

[7] Deepshree A. Vadeyar1,Yogish H.K" Farthest First Clustering In Links Reorganization" International Journal Of Web & Semantic Technology (Ijwest) Vol.5, No.3, July 2014

[8] Narendra Sharma 1, Aman Bajpai 2, Mr. Ratnesh Litoriya 3" Comparison The Various Clustering Algorithms Of Weka Tools" International Journal Of Emerging Technology And Advanced Engineering (Issn 2250-2459, Volume 2, Issue 5, May 2012)