International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

A Hybrid Data Mining Approach To Evaluate Performance of Classification And Clustering Methods Implemented On Weka Platform

Erica Sethi[1], Krishan Kumar[2]
M.Tech (CSE), JCDM College of Engineering, Sirsa, India [1]
Assistant Professor (CSE), JCDM College of Engineering, Sirsa, India [2]

ABSTRACT

Data mining is the process of finding of hidden information from a huge amount of data. Data mining analyzing the data from different source and convert it into meaningful information. Data mining is a new powerful technology that helps business to focus on important information like future trends, decision making, customer choice etc. A target dataset is prepared before applying the data mining algorithm. The common source of data is the data warehouse. Pre processing is needed to analyze the data sets before applying the data mining. Data mining is also defined as the process of discovering patterns in data. The process must be automatic or (more usually) semi-automatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage. The data invariably present in substantial quantities. Different types of learning techniques can be used, including classification, association rules, clustering, attribute selection, normalization, instance based measures and decision trees. Selection of a learning technique is a difficult task that depends on the database and the types of desired results. Raw data is useless without techniques to extract information from it. Main reasons to use data mining are too much data & too little information and need to extract useful information from the data and to interpret the data.

Keywords : WEKA, Classification Technique, Data Mining, Clustering Technique

I. INTRODUCTION

Data Mining (DM) represents a set of specific methods and algorithms aimed solely at extracting patterns from raw data. Data mining sometimes is also called knowledge discovery in databases (KDD). Knowledge Discovery in Databases (KDD) is an automatic, exploratory analysis and modeling of large data repositories. KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets. The existing relationships and patterns can also be found. Data mining combines machine learning, statistics and visualization techniques to discover and extract knowledge. Student retention has become an indication of academic performance and enrollment management. Here, potential problem will be identified at earlier. The raw data was preprocessed in terms of filling up missing values, transforming values in one form into another and relevant attribute/variable selection. One of the most useful data mining techniques for e-learning is classification. Raw data is useless without techniques to extract information from it. [3]

Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semi-automatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage. The

data invariably present in substantial quantities. Different types of learning techniques can be used, including classification, association rules, clustering, attribute selection, normalization, instance based measures and decision trees. Selection of a learning technique is a difficult task that depends on the database and the types of desired results. [1]

Data mining should be regarded as strategic and competitive move. So before the data mining process starts, the goal which is focus in the analysis should be clarified. Otherwise it's not possible to search for new valuable information if the necessary parameters cannot define as there are different models for the data mining process based on the task at hand. The following steps are involved in data mining [2]:

1. **Data selection**: The first step is data selection (where data relevant to the analysis task are retrieved from the database). It defines the particular KKD task. The important aspect of this step is to determine the data to be analyzed and how to obtain it.

2. **Data integration:** This step specifies the data has to be integrated (where multiple data sources may be combined).

3. **Data Preprocessing:** Data Preprocessing: In this stage the process of data cleaning and data integration is done.

4. **Data cleaning:** Data cleaning (to remove noise and inconsistent data). This means for example the filling of missing values. Thus, the missing values need to be completed and inconsistent data should be corrected or left out.

5. **Data transformation:** Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance). In data transformation stage the data are transfer into new format, if necessary

6. **Data mining:** Data mining is the application of efficient algorithms to detect the desired patterns contained within the given data. Thus, the data mining step is responsible for finding patterns according to the predefined task. Since this step is the most important within the KDD process. [4]

7. **Pattern evaluation:** The user evaluates the extracted patterns with respect to the task defined in the focusing step. An important aspect of this evaluation is the representation of the found patterns. Depending on the given task, there are several quality measures and visualizations available to describe the result. If the user is satisfied with the quality of the patterns, the process is terminated. However, in most cases the results might not be satisfying after one iteration.

8. **Knowledge presentation:** Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).The quality of the results varies, depending on the right choice of feature transformations, feature selections and data mining algorithms. A user must decide which techniques should be employed and select the step where the KDD process should be modified to improve the result after each iteration. [6]

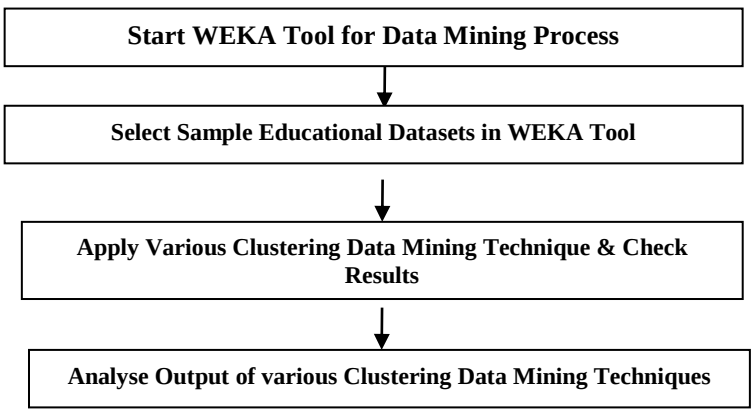

Figure 1: Framework of the proposed system

WEKA (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. WEKA is free software available under the GNU General Public License. The WEKA workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. This original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (WEKA), is now used in many different application areas, in particular for educational purposes and research. [9]

CLASSIFICATION TECHNIQUE

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. Classification is a simple process to finding a model that describes and distinguishes data classes of test. It is both types supervised learning and unsupervised. It consists of two steps: [3]

- Model construction.
- Model usage

Classification Technique used for data mining:

- ✓ BAYESIAN NETWORKS
- ✓ DECISION TABLE
- ✓ Lad Tree
- ✓ J48 Tree

Comparison of classification techniques as per the following parameters:

- ✓ Comparison based on Mean Absolute Error
- ✓ Comparison based on Root Mean Squared Error
- ✓ Comparison based on Relative Absolute Error
- ✓ Comparison based on Correctly & Incorrectly Classified Instances

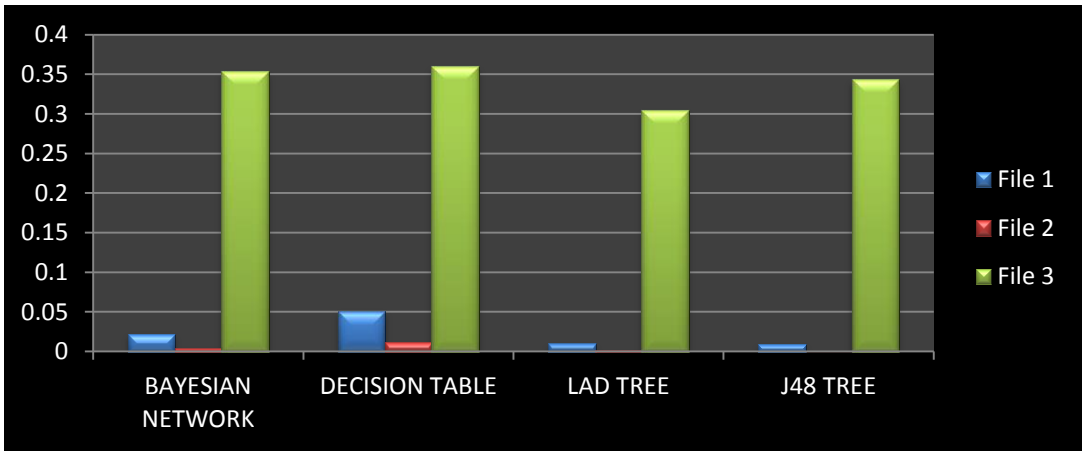

Figure 2: Comparisons between Parameters for Mean Absolute Error

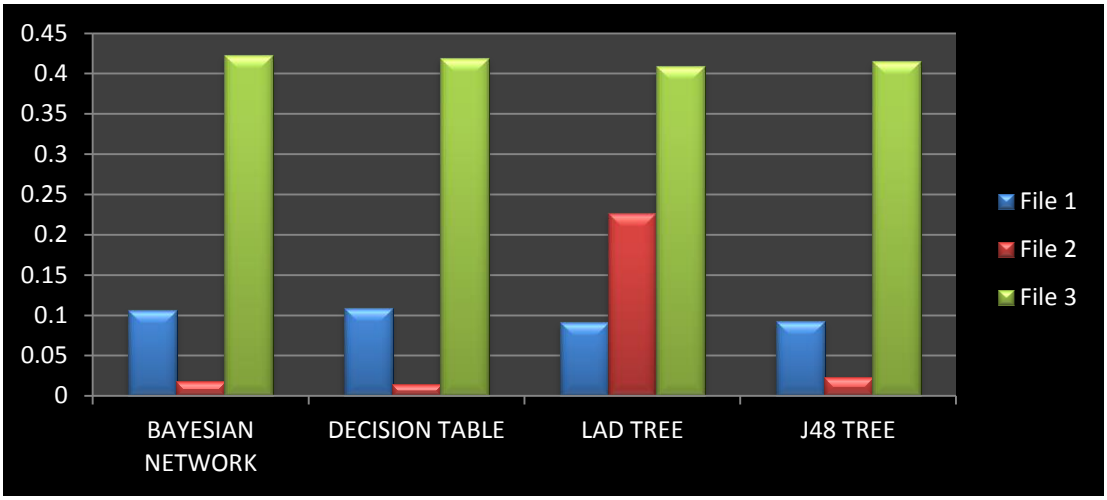

Figure 3: Comparisons between Parameters for Root Mean Squared Error

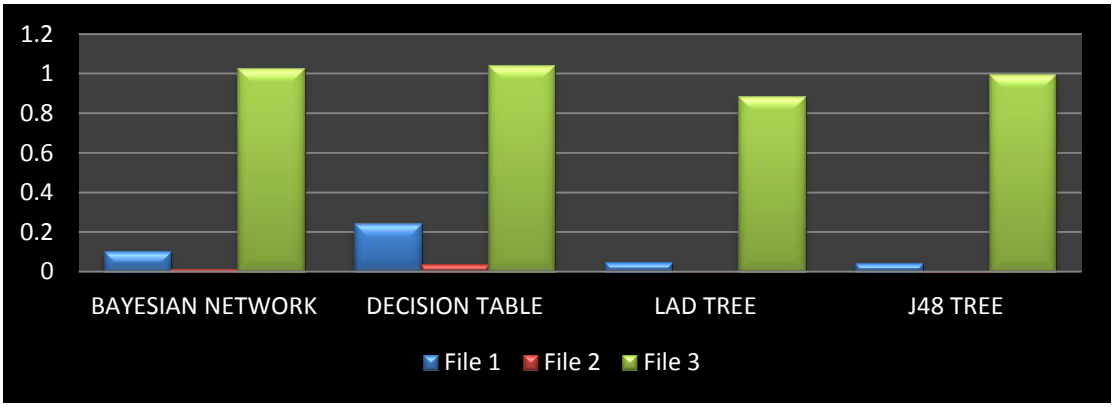

Figure 4: Comparisons between Parameters for Relative Absolute Error

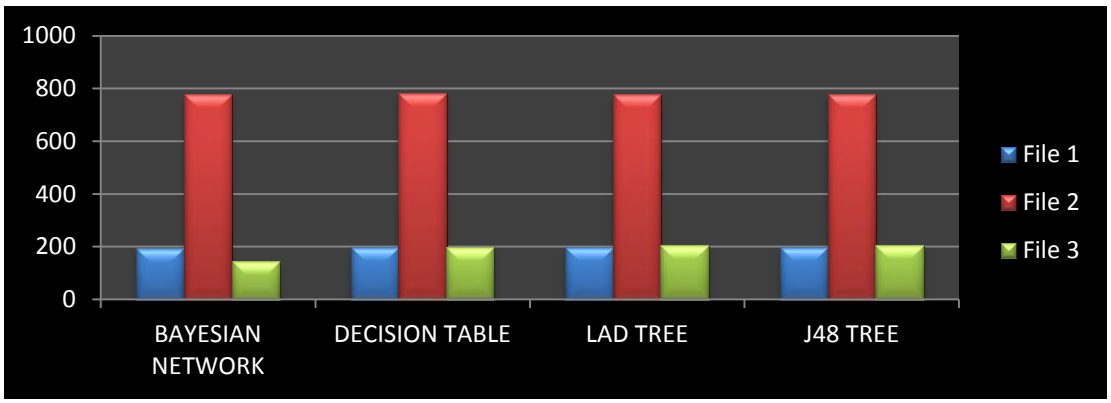

Figure 5: Comparisons of Parameters for Correctly Classified Instances

CLUSTERING TECHNIQUE

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. [8]

Types of clustering methods are:

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Farthest First method

- Grid-based methods
- Model-based methods

In educational data mining, clustering is play very important role. It is used to provide group the students according to their behavior e.g. clustering define clusters according to active student from non-active student according to their performance in activities. [7]

Clustering Technique used for data mining:

- ✓ K-MEAN PARTITIONING METHOD
- ✓ HIERARCHICAL CLUSTERING METHOD
- ✓ DENSITY BASED (DBSCAN) CLUSTERING METHOD
- ✓ FARTHEST FIRST CLUSTERING METHOD

Comparison of clustering techniques as per the following parameters:

- ✓ Comparison based on No. of Iteration
- ✓ Comparison based on Clustered Distribution
  (No. of Clusters and Instances in Cluster 1, Cluster 2 etc.)

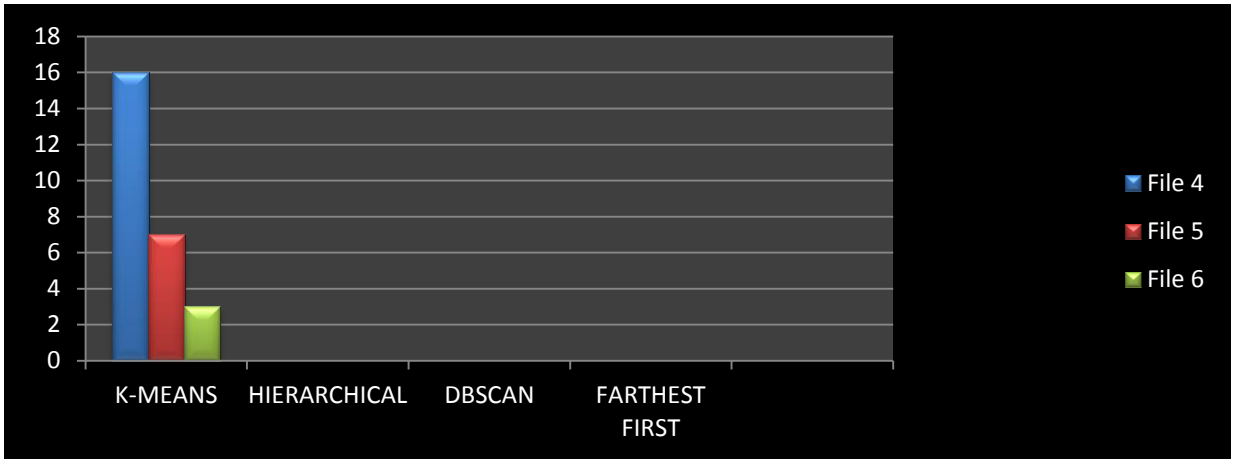

Figure 6: Compared according to No. of Iteration

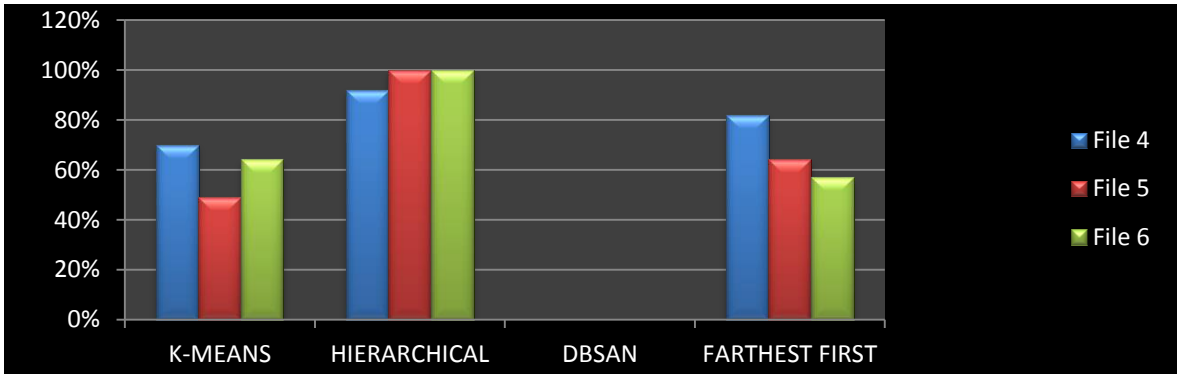

Figure 7: Compared According To Instances in Cluster 1

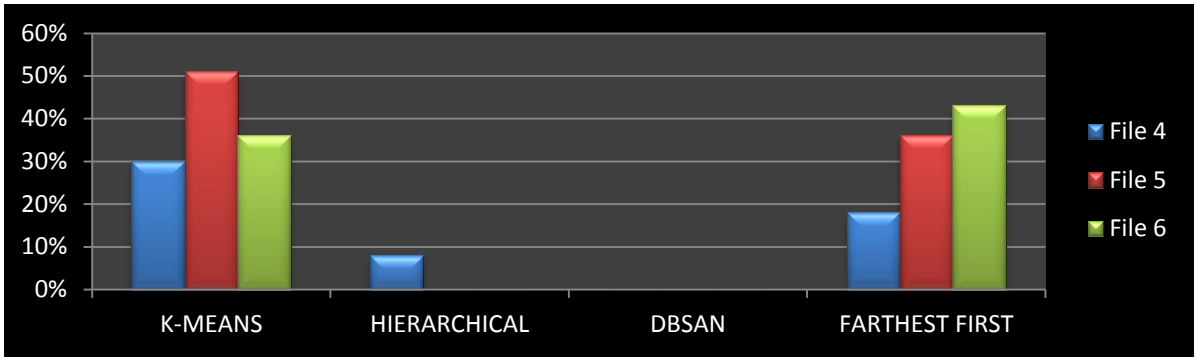

Figure 8: Compared According To Instances in Cluster 2

II. RESULTS & DISCUSSIONS

The results here have been formulated in the form of tables & graphs. Six datasets files in the CSV format have been selected. In WEKA, all data is considered as instances and features in the data are known as attributes. Accuracy, mean absolute error, Root mean squared error, Relative absolute error and other parameter are found from the different selected classification methods. Clustered Instances and No. of Iteration are the parameters found from the different selected clustering methods.

CLASSIFICATION TECHNIQUE RESULTS

Figure 9: Average errors for different classification techniques

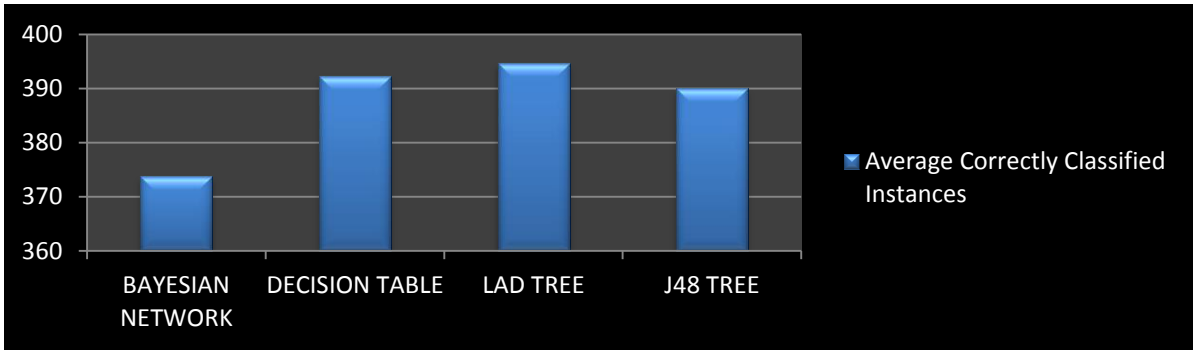

Figure 10: Average correctly classified instances for different classification techniques

So, finally we can say that **"LAD Tree Classifier"** is best suitable for educational datasets.

CLUSTERING TECHNIQUE RESULTS

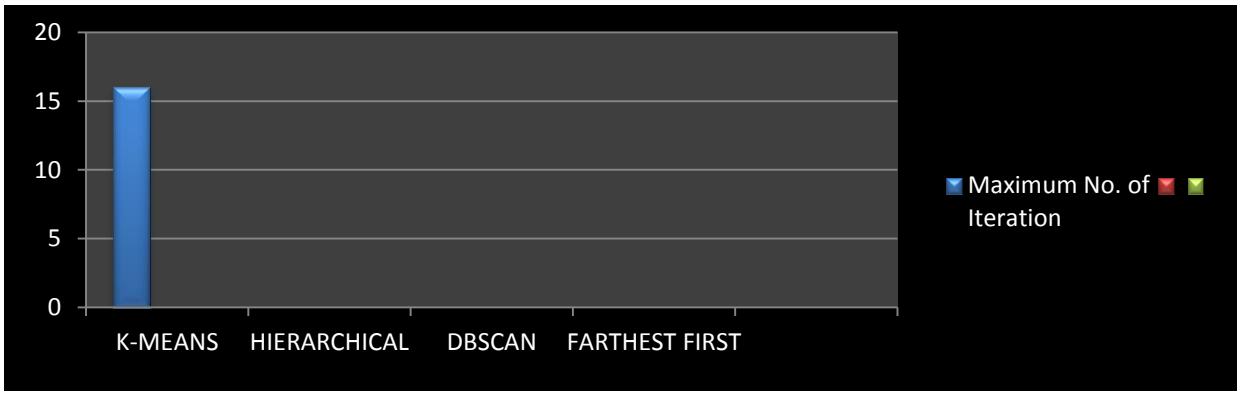

Figure 11: No. of Iteration for different clustering techniques

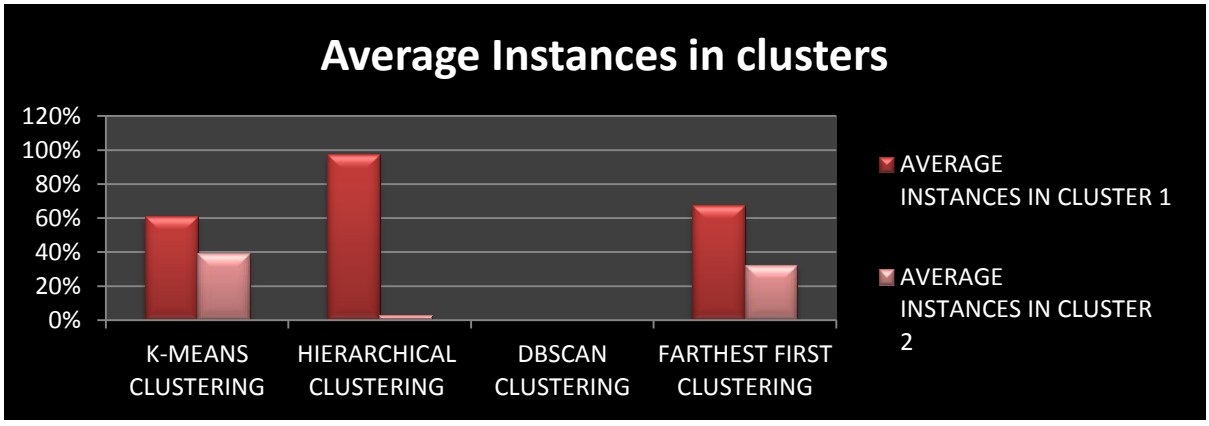

Figure 12: Average correctly classified instances for different clustering techniques

So, finally we can say that **"Hierarchical Clustering"** is best suitable for educational datasets

III. REFERENCES

1. JVasuki, S.Priyadarshini, "A STUDY ON BASICS OF DATA MINING, MACHINE LEARNING AND BIG DATA" International Journal of Innovative Research in Computer and Communication Engineering, ISSN: 2320-9801, Vol. 5, Issue 1, January 2017, pp. 199-206.
2. NV. Ramana Murty and Prof. M.S. Prasad Babu, "A CRITICAL STUDY OF CLASSIFICATION ALGORITHMS FOR LUNGCANCER DISEASE DETECTION AND DIAGNOSIS" International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, Number 5 (2017), pp. 1041-1048.
3. K Sumathi, S. Kannan, K. Nagarajan "DATA MINING: ANALYSIS OF STUDENT DATABASE USING CLASSIFICATION TECHNIQUES" International Journal of Computer Applications, ISSN: 0975 – 8887, Volume 141 – No.8, May 2016, pp. 22-27.
4. PKeerthana, P.Thamilselvan, J.G.R. Sathiaseelan, "PERFORMANCE ANALYSIS OF DATA MINING ALGORITHMS FOR MEDICAL IMAGE CLASSIFICATION", International Journal of Computer Science and Mobile Computing, ISSN 2320–088X, Vol. 5, Issue. 3, March 2016, pg.604 – 609.
5. Alpa Shah, Ravi Gulati "PRIVACY PRESERVING DATA MINING: TECHNIQUES, CLASSIFICATION AND IMPLICATIONS - A SURVEY" International Journal of Computer Applications, ISSN: 0975 – 8887, Volume 137 – No.12, March 2016, pp. 40-46.
6. Dr. P. Nithya, B. Umamaheswari, A. Umadevi, "A SURVEY ON EDUCATIONAL DATA MINING IN FIELD OF EDUCATION" International Journal of Advanced Research in Computer Engineering & Technology

(IJARCET), ISSN: 2278 – 1323, Volume 5 Issue 1, January 2016, pp. 69-78.

7. Sukhvir Kaur, "SURVEY OF DIFFERENT DATA CLUSTERING ALGORITHMS", International Journal of Computer Science and Mobile Computing, ISSN 2320–088X, Vol. 5, Issue. 5, May 2016, pg.584 – 588.
8. Mythili S, Madhiya E, "AN ANALYSIS ON CLUSTERING ALGORITHMS IN DATA MINING", International Journal of Computer Science and Mobile Computing, ISSN 2320–088X, Vol. 3, Issue. 1, January 2014, pg.334 – 340.
9. Sonam Narwal, Kamaldeep Mintwal "COMPARISON THE VARIOUS CLUSTERING AND CLASSIFICATION ALGORITHMS OF WEKA TOOLS" International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 3, Issue 12, December 2013, pp. 866-878