HCTL Open International Journal of Technology Innovations and Research (IJTIR)
http://ijtir.hctl.org
Volume 14, April 2015
e-ISSN: 2321-1814, ISBN (Print): 978-1-62951-946-3

Educational Data Mining: Performance Evaluation of Decision Tree and Clustering Techniques using WEKA Platform

Ritika Saxena

ritikasin25@gmail.com

Abstract

Data Mining plays a vital role in information management technology. It is a computational process of finding patterns from large databases. It mainly focuses on extracting knowledge from the given or the available data. Different knowledge extracting tools are used. This tool is most common among every sector be it educational, organizational etc. Educational Sector can take advantage out of these tools in order to increase the quality of education. But the sad part is still in present educational systems are not using it. Higher education Institutions needs to know which student will enroll in which course, which student needs more assistance. In data mining users are facing the problem when database consists of large number of features and instances. These kinds of problem[s] could not be handled using decision trees alone or clustering technique alone. Because, decision trees depend upon the dataset used and the configuration of the trees. Similarly, clustering alone doesn't work for all kind of patterns. So in order to find that which technique is most suitable, in this paper we have evaluated the performance of both the algorithms. Educational data is mined and the algorithms are applied to it so as to predict the results.

Keywords

Weka, EDM, Decision Trees, Clustering, KDD.

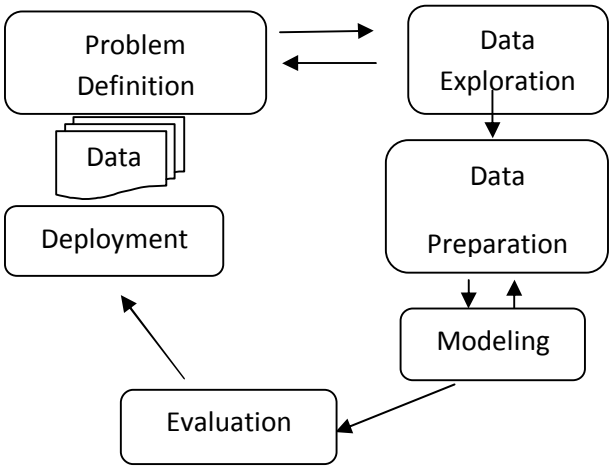
Introduction

Data mining is widely used in diverse areas. Data Mining is the process through which we can analyse the different type of data and further extract the useful information. Data Mining is sometimes known as KD i.e. Knowledge Data. Data Mining is one of the tools which help in analyzing the data. It helps in analyzing the data from different angles, categorising them and hence summarize the relationships identified. Therefore, data mining is the process through which we can find different relations and patterns generated among different fields of databases. Educational Data Mining is a different research field

with the application of data mining, machine leaning and statistics to information generated from educational settings. The main goals of educational data mining are predicting student's future learning behaviour, advancing scientific knowledge, effects of educational support. In this research paper we are using data mining methodologies in order to understand student's performance using the two commonly used techniques of data mining. Data mining helps in studying the performance of students based on their past records. In this paper, we are using two techniques of data mining that is **clustering** and **decision trees** so as to get the results of the inputs provided by us and compare those results as to which of the techniques is more preferable and gives the better results.

Background Study

Data Mining is the analysis step of KDD (Knowledge Discovery Database). Data mining tasks are semi-automatic or automatic based on the quantity of the data available. Data collection, preparation, interpretation and reporting are not the part of data mining whereas they are the part of KDD. Data Mining could be performed using different tools some of the tools are listed below. Different Phases in Data Mining include;



Data Mining Phases

The data in the educational institutions that is stored in electronic form has seen a dramatic increase. Historical as well as organisational data is stored in the databases. It is really cumbersome to manage such data manually. Different relations have to be produced out of the stored data in the database. The categorization of the students according to their academic result is the important task for the institutions in order to increase the credibility of their institution. There is no assurance whether there are predictors that can determine or predict whether the student is academic is below average, average or a genius student. In present day of the educational system, a student's performance is determined by the combination of primary results, secondary results, internal assessments, tests, assignments and attendance etc. Therefore, in this research paper we are analysing the results as to which algorithm is more efficient and predictable.

Related Work

Lakshmi et al[1] describes the student's performance . They used ID3 algorithm in order to classify the students performance and according to which they will be allotted the area for their master. ID3 algorithm is the classification algorithm through which we can construct decision trees using top down, bottom up and greedy search methodologies. In order to select the attribute which is most useful for the classification of the datasets, metric-information gain is used.

Ali [2], also describes that how data mining could be used in educational sector. As the information is collected by the students at the time of admission and saved in the computer this provides benefit for business point of view. He used data mining to classify and cluster the information based on psychographic, behavioural and demographic variables. Therefore, it helps in describing about the student's profile whether they are successful or unsuccessful based on their percentage or GPA secured during secondary examination and semesters.

Sembiring [6], created a model based on psychometric analysis of the students using data mining techniques. He created a rule model of the student's performance based on their psychometric behaviour. The predictor variables used are- Interest, Believe, Family Time, Study Behaviour. The model developed here uses the two main methodologies i.e. kernel k-means clustering and support vector machine (SVM) classification. As it could be used on large as well as high dimensional data that are non–linearly separable.

Bhullar [7], describes a data mining tool that helps to find out the student that are weak in academics and need assistance. He used Weka classification algorithms that provide stability between precision, speed and interpretability of results. J48 algorithm is used which helps in classifying similarly as the decision trees.

Baradwaj [8], collected the information of the 200 students from VBS Purvanchal University, Jaunpur(U.P) such as their previous semester marks (PSM), class test grade (CTG), Seminar performance(SEM), General Proficiency(GP),Attendence(ATT) and Lab Work(LW) . Using this record he used the classification technique in order to cluster them based upon the percentage and good, average, poor. He also measured entropy so as to check the impurity.

Romero [9], in his survey described the different studies carried out in this field. He described different types of used techniques and educational environments and also the similar work that is done related to educational data mining. Data mining tools are normally designed in their flexibility rather than their simplicity. He explains both the aspects where on the one hand the user has to select the algorithm to carry out with the given data and on the other hand the algorithm has to be configured before its execution. XML, PMML, OWL, RDF, SCORM are some of the current data mining tools.

Calders [11], presents their works in which there are four EDM different paper that represent a crosscut of different applications areas of data mining.

M. Abu Tair [12], in their research work presented a case study to improve the students' performance mining the data. They extracted the useful knowledge from the database and after preprocessing the data applied mining techniques such as association, outlier detection rules and classification.

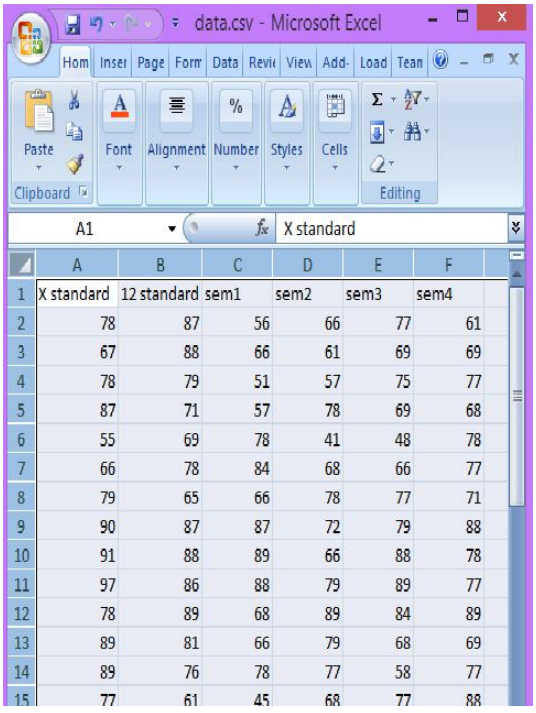
Methodology

In this paper we have used Weka 3.7 as a comparison tool. Weka contains the collection of machine learning algorithms. We can perform classification, clustering, association rules, pre-processing and visualization of the data. In our work we have created a dataset

containing 7 attributes in excel file using CSV file format. The process that is carried out in this paper is described below.

Data Mining Process

The data sets used in this study is obtained from the database of the student's from one of the educational institutions. Initially the marks of the students are stored such as their 10th percentage, 12th percentage, semester marks. First of all the marks are collected of different students in table and then the processing is carried on. Using Weka 3.7 we can classify and cluster the data available in our dataset. Weka 3.7 is one of the most efficient comparison tool when it's related to data mining techniques. The results can be tabulated very fast and accurate through this tool. When we perform such tasks manually without any involvement of tools then it becomes cumbersome to reach to the desired results as well as the tasks becomes time taking.
In the first phase, clustering technique → k-means clustering is used through the Weka 3.7. From the given options in the Weka 3.7 tool we choose the simple k-means algorithm and then through parameterise option we allot the number of clusters to be formed.
In the second phase, classification technique → decision tree is used through Weka 3.7. It helps in visualizing the tree structure of the input dataset.
Then the results of both the algorithms are analysed and the results are predicted as to which algorithm is more efficient and preferable.



	A	B	C	D	E	F
1	X standard	12 standard	sem1	sem2	sem3	sem4
2	78	87	56	66	77	61
3	67	88	66	61	69	69
4	78	79	51	57	75	77
5	87	71	57	78	69	68
6	55	69	78	41	48	78
7	66	78	84	68	66	77
8	79	65	66	78	77	71
9	90	87	87	72	79	88
10	91	88	89	66	88	78
11	97	86	88	79	89	77
12	78	89	68	89	84	89
13	89	81	66	79	68	69
14	89	76	78	77	58	77
15	77	61	45	68	77	88

Figure1. Data Collected through Students Database

The above shown is the dataset collected from the university's database. The excel file is saved using csv extension; csv denotes comma separated values which will be used as an input file. Benefit of using this extension is that data is automatically tabulated as the separated values while the processing in the Weka 3.7 tool.

Clustering

A clustering is a method in which clusters or groups are formed. It helps in grouping or collecting the elements of the same kind in one class or group. These elements are of same type and pattern and are different to those that belong to different groupings. This can be said as one of the main tasks of data mining and also a common technique for statistical data analysis. It is used in many different fields such as machine learning, pattern recognition, bioinformatics, image analysis and information retrieval. There are many types of clustering algorithms such as hierarchical clustering, k-means algorithm, Expectation maximization algorithm (EM), Density-Based Spatial Clustering Of Applications With Noise (DBSCAN) , Biclustering algorithm, Fuzzy clustering.

K – Means Clustering:

The term 'k-means' was first coined by James MacQueen in 1967. It uses an iterative refinement technique. In simple words , it is the algorithm that is used to classify or collect the group of objects that is based or classified according to K number of attributes, here there are 7 attributes according to which clustering is done. This grouping is done by calculating the sum of square of distances between the given data and cluster centroid and the elements having the minimum distance are grouped together.

The basic steps of k-means are following:

- Find the centroid.
- Calculate the distance of each objects to the determined centroid.
- Group the objects based on the minimum distance, i.e. the closest centroid.

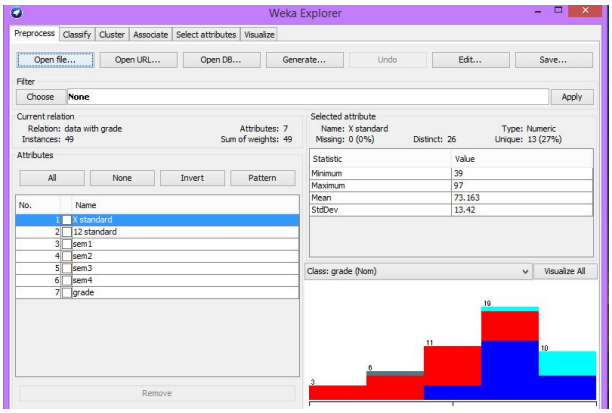


Figure 2. Pre-processing the input .csv file

Firstly, we pre-process the data before being clustered using Weka 3.7 tool. It helps in to transform the unmanaged data to the understandable format. It also helps in preparing data for further processing. The steps that data undergo during pre-processing are:
Data Cleaning, Integration, Transformation, Reduction and Discretion.
After this we apply K-means Clustering through Weka in which the input is the dataset collected.

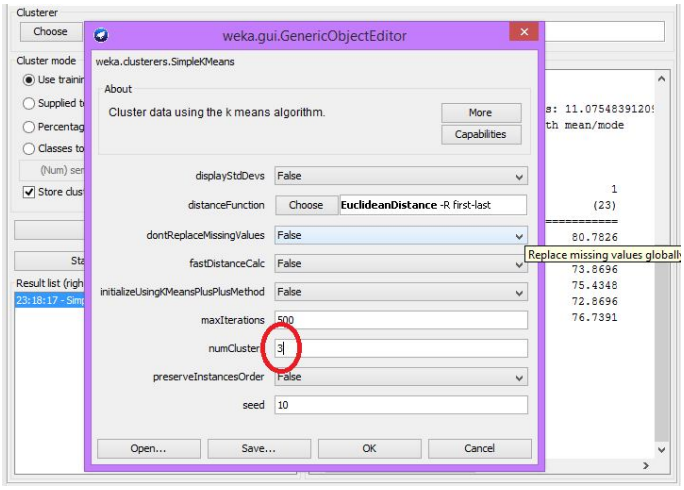


Figure 3: K-means applied using 3 clusters

We parameterise number of clusters equal to 3. And then the results are tabulated. In Weka 3.7 the default clusters are 2 but in this processing 3 clusters are formed that could be related to Grade A, B, C. That is, here we parameterise the number of clusters into which we want our data to be clustered.

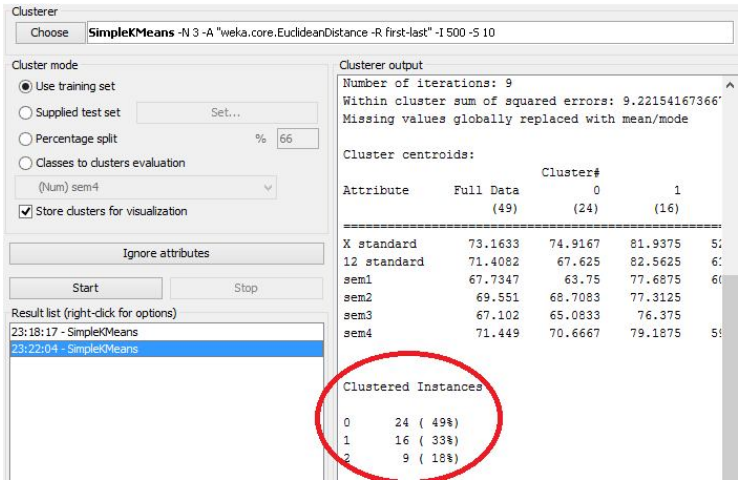


Figure 4: Clusters 0, 1, 2 are formed after applying algorithm

Therefore we can see that 3 clusters are formed shown within the circle. (0,1,2) [Figure 4]. Cluster 0 contains 24 elements, Cluster 1 contains 26 elements and lastly Cluster 2 contains 9 elements. From the following figure we can see the cluster visualization for our input dataset provided . We can also parametrise the data using x axis and y axis values. Accordingly to which the data can be visualized.

Figure 5: Cluster Visualization

From the tool we can access that in this the percentage of errors are 17.792%.

Decision Trees

After clustering all the datasets given, we will perform the decision tree algorithm on the same dataset. In decision trees the data is categorised in the form of the tree where at the end i.e. the leaf nodes, depict the classes which contain the datasets. It is one of the most simple and precise technique so as to mine the data and get the result efficiently. Thus, this study will also help the students in order to improve their performance for the future grades.

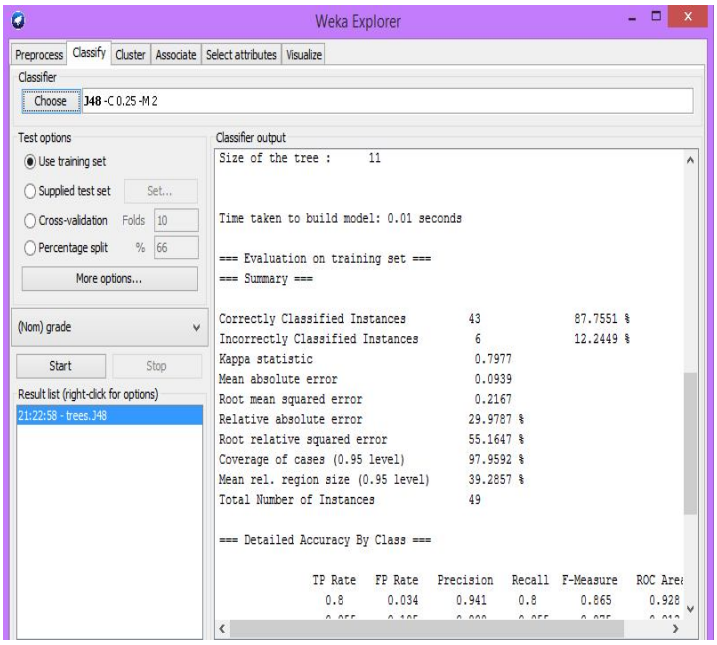


Figure 6: Decision Tree Implementation

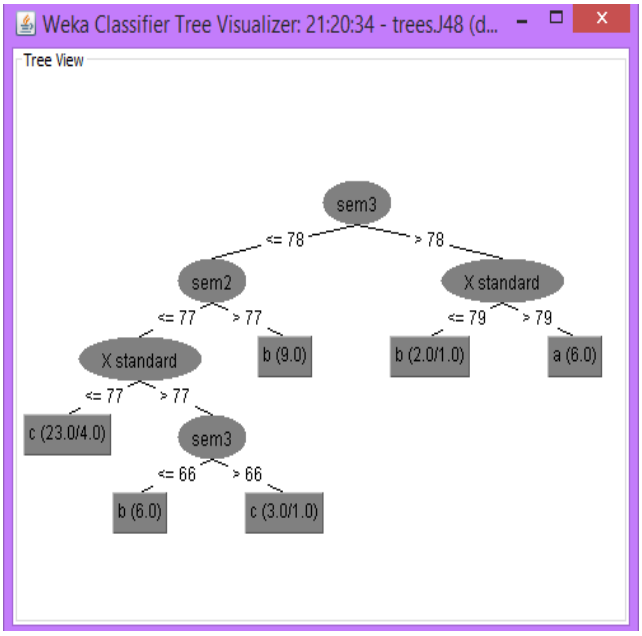


Figure 7: Decision Tree Visualization

As in this case all the datasets were initially classified and then the results are passed through the decision trees algorithm and hence the results are tabulated through visualizing the tree.

As in decision tree we can see that an accuracy of 87.751% is achieved.

Results

When we perform clustering after the datasets are clustered accordingly to their relevant classes the numbers of iteration to form 3 clusters were 5. The root mean squared error value is 17.793. As lower the number of mean squared value will be the more efficient the algorithm will be.

```
kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 17.793336777151243
Missing values globally replaced with mean/mode

Cluster centroids:
                           Cluster#
Attribute      Full Data          0          1          2
```

Figure 8: Results generated after clustering algorithm analysis

Thereafter, performing the second technique of data mining i.e. classification through decision trees algorithm. Here we have used J48 algorithm, J48 algorithm is the implementation of ID3 algorithm. It helps in creating univariate decision trees. The results retrieved after the algorithm implementation is shown below,

```
=== Summary ===

Correctly Classified Instances          43               87.7551 %
Incorrectly Classified Instances         6               12.2449 %
Kappa statistic                          0.7977
Mean absolute error                      0.0939
Root mean squared error                  0.2167
Relative absolute error                 29.9787 %
Root relative squared error             55.1647 %
Coverage of cases (0.95 level)          97.9592 %
Mean rel. region size (0.95 level)      39.2857 %
Total Number of Instances               49
```

Figure 9: Results generated after Decision trees algorithm implementation

From the above decision tree algorithm results we can see that the root mean squared error in this is equal to 0.216, which is much less than that of clustering algorithm.

From the above results we can deduce that decision tree algorithm i.e. classification technique is more preferable over clustering.

Conclusion

In this paper, two algorithms are used so as to predict the performance of both the algorithms. They are applied on the marks of the student retrieved from the database of the university so as to grade the students based on their up to date performances. Here we have used the technique of clustering, decision trees in order to mine the data as the huge amount of data is available in the university containing the students record so it is required to refine the data so that the results could be used for the future evaluation. First of all we evaluated the performance of the clustering algorithm and then secondly we evaluated the performance of decision trees algorithm and then the judgement is made as to which algorithm performance is suitable. And after performing both the techniques it is concluded that decision tree using J48 algorithm is more efficient than clustering k-means technique.

The accuracy achieved through decision tree is much more than that achieved through clustering. So it is concluded that classification techniques are preferable than the clustering techniques.

References

[1] LAKSHMI ,D.BHU., . ARUNDATHI , S. and DR.JAGADEESH, “Data Mining: A prediction for Student's Performance Using Decision Tree ID3 Method”, July 2014.

[2] Ali ,Mohd. Maqsood., “Role Of Data Mining In Education Sector”, April 2013.

[3] DeLaFayette Winters, Titus.,” Educational Data Mining: Collection and Analysis of Score Matrices for Outcomes-Based Assessment”, June 2006.

[4] Torre ,Javier., Rodriguez ,Alejandro., Colomo, Ricardo., Jimenez ,Enrique. and Alor, Giner.," Improving Accuracy of Decision Trees Using Clustering Techniques", February 2013.

[5] Ranganathan ,Sindhuja., "Improvements To K-Means Clustering", August 2013.

[6] Sembiring ,Sajadin. ," An Application Of Predicting Student Performance Using Kernel K-Means And Smooth Support Vector Machine", August 2012.

[7] Singh Bhullar , Manpreet., Member IAENG, " Use of Data Mining in Education Sector", October 2012.

[8] Kumar Baradwaj , Brijesh. , "Mining Educational Data to Analyze Students Performance",2011.

[9] Romero ,Cristobel. ,, "Educational Data Mining: A Review of the State-of-the-Art",Member, IEEE, Sebastian Ventura, Senior Member, IEEE2010.

[10]Patel ,Ketul B. , Chauhan , Jignesh A. and Patel, Jigar D. , "Web Mining in E-Commerce: Pattern Discovery, Issues and Applications", 2011.

[11]Calders ,Toon . and Pechenziky ,Mysoka.,"Introduction to Special Section On Educational Data Mining", Volume 13 Issue 2.2013.

[12]Tair , Mohammad M.Abu . and El-Halees, Alaa M. , "Mining Educational Data to Improve Students' Performance: A Case Study", International Journal of Information and Communication Technology Research, Volume 2 No. 2, February 2012.

[13]Huebner, Richard A., Norwich University, "A survey of educational data mining research", Research in Higher Education Journal.

[14]Baker , Ryan SJD. and Yacef, Kalina., "The State of Educational Data Mining in 2009: A Review and Future Visions", Journal of Educational Data Mining, Article 1, Vol 1, No 1, Fall 2009.