Journal of Basic and Applied Engineering Research
p-ISSN: 2350-0077; e-ISSN: 2350-0255; Volume 3, Issue 6; April-June, 2016, pp. 528-531

http://www.krishisanskriti.org/Publication.html

A Review Paper On: Predictive Analysis of Student Performance and Outlier Detection

Sapna Kataria[1] and Meghna Sharma[2]

[1]M.tech, North Cap University
[2]North Cap University
E-mail: [1]sapnakataria20@gmail.com, [2]meghnasharma@ncuindia.edu

Abstract—*Educational institutes are spread all over the world; it is a very big wide spread domain. So the huge amount of data is available in educational research field. The data is associated with evaluation of the candidates among all of the institutes. Educational datamining is the prominent system that can resolve the problem of statistical performance assessment of the candidates; on the other hand it also helps in predictive analysis. The main objective of this paper is to review the previous work done by the various data miners in the last decade and it also explain various data mining methods useful for performance evaluation and prediction of the students.*

Keywords*: performance evaluation, prediction, classification, clustering, open source tools etc.*

1. INTRODUCTION

At present, in the time of so much competition where every student, college and university, wants to be ahead in the race, every year the stored database is mounting speedily, every college wants its student to get placed in the best companies. In this scenario colleges need to refine its student database accurately, so that the concealed information about student's performance can be used. Predictive analysis of the student's performance is become the necessity for the college authorities who desire to know student's performance in advance, so that they can prepare accordingly and the placement of the college and the future of the students can be improved. For the furtherance of students, colleges and universities, the Predictive analysis of student performance is done by the data miners.

Data mining techniques are useful for various areas like hospitals, banks, airline or railway passenger's record, buyer's record at the store, businesses, and bioinformatics etc. data mining techniques used in educational area is known as educational datamining. Some well-known methods such as clustering, classification and regression are used to discover the unknown facts from the immense volume of data stored.

In this survey paper, we discussed about the work of the various data miners, how the predictive analysis is done on the student database, the data mining techniques and the tools used for the mining process. The study is always performed on the real time college data, the results are provided back to the training and placement office of the college. This paper is organized in 5 sections, section I consist introduction about the paper, whereas the section II describe the previous work done by the various data miners, the popular data mining trends are described in section III, a comparative study of the open source tools available is presented in the section IV and at last the final section V concluded the final result.

2. RELATED WORK

There has been a lot of work done in this field earlier; some of the boundless work is discussed here, we consider all the respectable work and there significance as well. Ali Buldu, Kerem Üçgün [1] studies the student's database and find out the relation between the student behavior, performance and courses studied by them. They concluded the result in the form of the generated rule that shows if a student is failed in the course studied by him in ninth class then there are enough chances for those candidates to be unproductive in tenth class also. These outcomes are produced for various courses. By this research students can be assisted to pick the subjects wisely according to the results about the relation between their selected courses.

Baha Sen, Emine Ucar[2] estimate the performance of Engg. Students of distance learning by using the various datamining techniques. The outcome exposes that with the increase in the student's age the chances of success declines, further candidate's achievement level is abundant in distance learning comparative to regular education, and the students that passed from professional high schools are extra fruitful in artistic learning than occupational lessons.

Hijazi etal [3] performed a research on the performance of the students via picking a model of 300 candidates from various colleges allied to Punjab University (Pakistan). The proposition was specified as the attitude of the Students to being present in class, total time used up in study every day after attending college, family income of the students, age and education of their mothers is pointedly linked with the candidate's performance. Simple linear regression technique

was used and it was concluded that the aspects like parent's education and annual household earnings were extremely interrelated with the learner's educational achievements

A study was conducted by Ventura and Romero [4] on educational datamining from 1995 to 2005. The result was concluded that educational datamining is a favorable area of exploration and it had definite requirement that is not accessible in other fields. Therefore, analysis should be focused on educational area of datamining.

Er. Rimmy Chuchra [5] conducted the analysis using clustering, decision tree, and neural network techniques to estimate performance of the candidates by choosing a group of students from Sri Sai Engineering University, Phagwara, India. Via using these methods it was concluded that lecturers can certainly estimate the candidate's performance.

C.Marquez, et al [6] conducted the study on detecting the aspects that has effect on the student's performance at various educational stages. 670 intermediate school student's dataset was taken from Zacatecas, Mexico. 10 classification algorithms were used by them and 15 different attributes were selected, it was concluded that sociological, educational and economic attributes are more significant for the estimation of the low performance of the school students.

3. POPULAR DATA MINING TRENDS

There are some popular data mining techniques such as clustering, classification, prediction, and regression etc. here we discuss some of these techniques and how are they used for the mining of the huge datasets.

Clustering: clustering is one of the famous data mining methods, It is a type of unsupervised classification technique; there are no predefined classes. it is the process of grouping of the similar type of data objects in one cluster, and the cluster could be of any shape spherical, polygon, arbitrary etc. For visualization the clusters are represented graphically because it may expose the hidden information. There could be more than one cluster; each of different shape within a single dataset. The data items within a cluster are more similar while dissimilar with other cluster's data items. Those data item which are not included any of the clusters are known as outlier, hence clustering also helps us to find out the outlier data items in our database. It is a familiar method for the analysis of statistical data; it is used in numerous fields like machine learning, image processing, pattern recognition etc. many type of various algorithms are used that are significantly different in their concept of what organize a cluster.

There are some requirements of clustering in data mining such as scalability, capability to handle different types of aspects, detection of clusters with random shape, capable to handle noise and outliers, eminent dimensionality, accountability and utility, integration of user defined constraints, minimum requirements for constraint information etc. Clustering could be performed on different type of data such as categorical data items, interval scaled, binary variables, variables of mixed type, ordinal data values and ratio scaled etc. Some familiar algorithms that are used for clustering analysis purpose are like K-means algorithm, hierarchical clustering algorithm, distribution algorithm and density based algorithms, grid based algorithms, partitioning algorithms etc.

The worth of the clustering outcomes is measured by means of its capability to determine the unknown patterns. A worthy clustering produces high quality clusters with high intraclass similarity and low interclass similarity.

Fuzzy c-means clustering: The data we are going to use in student performance analysis is not certain; infact the data is uncertain; this means that there is a possibility associated with the data items. In uncertain dataset items every transaction has some non-zero probability associated with each transaction which tells about the likelihood of presence of the item in that transaction. We see almost every real world data relationships are uncertain or fuzzy in nature. For this type of data we use fuzzy clustering or soft clustering like fuzzy c-means clustering. Let's understand the concept of the uncertain data with the help of an example.

For example an ice-cream parlor sold

(i) 75% vanilla ice-cream (I_1).
(ii) 28% butterscotch (I_2).
(iii) 68% strawberry flavor (I_3).
(iv) 88% chocolate flavor (I_4).

These are represented in the table as follow:

Table 1: Example of uncertain dataset.

	I_1	I_2	I_3	I_4
T1	0.75	0.28	0.68	0.88
T2	0.24	0.56	0.49	0.76
T3	0.19	0.72	0.89	0.62

The fuzzy association rules are calculated as follow:

The fuzzy itemset F={T1, T2, T3}; where T1, T2, T3 are fuzzy sets. Support for F is

$Sup(F)=\sum \mu_F (X_n)/n$

Where *sup*(F) is support for F, X represents the transaction, $\mu_F (X_n)$ represents the relationship or belonging derived from product operation of each item, and *n* symbolized the no. of transactions. Further the confidence for the fuzzy association rules is as follow:

$(A \rightarrow B)=sup(AUB)/sup(A)$.

In fuzzy c-means clustering *N* no. of clusters are made in which every item belongs to every cluster to some extent; the item which is near to the center of the cluster has high degree of membership value and the item which is far has the low degree of membership value. In this technique the starting

estimate of the center of the cluster is most possibly wrong; FCM iteratively moves the center to the right position.

k-means clustering: Another popular method used for the clustering is k-means clustering. K-means clustering is popularly used with certain or precise data. This is the simplest and most frequently used method, it was invented in 1965. To minimize the inter cluster distance it use the Euclidean distance. In the k-means algorithm we have to predefine the number of cluster, denoted by K. Initially, the centers of the clusters are chosen randomly and then the centers are changed iteratively. K-means algorithm stops when the center of the cluster stops changing their position. i.e. the position of the center at $(i-1)^{th}$ iteration is equals to the center position at i^{th} iteration. Here the inter-cluster distance is minimum and intra-cluster distance is maximum in k-means clustering algorithm.[12]

Let M, denotes the function that minimize the inter- cluster distance and maximize the intra-cluster distance; the function M is defined as

$$M = \sum_{i=1}^{n} Zi|Di - Ci|$$

In this equation, Zi denoted the membership function; Di denoted the data items and Ci denotes the centers with in K no. of clusters. The initialization of the center of clusters strongly influenced the results of the k-means clustering. To overcome this problem a no. of iterations could be done; choose the best result. Besides this k-means algorithm takes all the data set items equally important but the real time objects or things do not have equal value. In k-means we have only round shape clusters.

Hierarchical clustering: hierarchical clustering method is different than other clustering methods; it don't cluster straight in one go infact it use 2 division approaches

(i) ***Agglomerative/ bottom up approach*:** in this approach all the item in the data set are clustered individually and then iteratively the closet cluster will be merged into one giving the final cluster output.

(ii) ***Divisive/ top down approach*:** in this approach all the items in the dataset are clustered in one single cluster and the iteratively the items having different properties are divided into smaller cluster giving out the final cluster output.

Hierarchical clustering could not produce better results for large data sets. There are many clustering algorithms available for hierarchical clustering like BIRCH, ROCK, CURE etc. For Boolean data & categorical data we use ROCK clustering; it is a strong clustering method which can easily handle Boolean type data.[13] It processes relationship (similarity/ dissimilarity) between 2 data items based on the links to its neighbors according to a new concept introduced by this clustering method. It helps in detecting outliers as well; like k-means algorithm we have to define no. of clusters initially. ROCK clustering produces arbitrary shape clusters. The clustering criterion for ROCK is as follow [10]

$$E = \sum_{i=1}^{k} n_i \left[\frac{link(P_q, P_r)}{n^{1+2f(\theta)}} \right]$$

In this equation n_i represents the cluster size, and link (P_q, P_r) represents the entire sum of links with the neighbors of that data points with in the single cluster whereas f(θ) represents dependency of the dataset and the cluster it belongs to.

4. COMPARISON OF THE CLUSTERING TECHNIQUES

We have studied some of the main clustering techniques; here in this section we are summarizing the various features of these clustering techniques in the table below.

Table: Comparison of different clustering

Sr. no.	Features	K-means clustering	Fuzzy c-means clustering	Hierarchical clustering
1	developed	1965	1973	1978
2	Clustering type	Hard clustering	Soft clustering	Agglomerative and divisive
3.	Data sets	Precise	Fuzzy	Text/Numerical/DNA sequence/categorical
4.	Membership function	Membership vector	Membership matrix	Membership matrix
5.	Uses	Simple, virtual & linear problem	Real, complex & non-linear problem	Real, complex & non-linear problem
6.	Computational time	Less	More	More
7.	Cluster shape	Round	Arbitrary	Cluster tree/dendograms
8.	Speed	High	Low	
9.	Distance measure	Euclidean distance	Euclidean distance	Any valid distance measure
10.	Outliers	No	No	Yes
11.	Complexity	O(n)	O(n)	$O(N^2)$

5. OUTLIER DETECTION REQUIREMENT AND RESPONSIBILITY

An outlier is an anomaly in the data set i.e. any item that is not showing normal behavior or having different qualities than the dataset they are known as outliers. Generally, all real time data contains outliers for example medical, banks or college data. These outliers can drastically change the results of any

statistical computation or may lead to wrong reports of any patients. Because of this reason outlier detection process is required in almost every field. In this section we have defined some techniques of outlier detection different methods and applications.

What is outlier detection?

Outlier detection is a process in which we find out the anomalies of the dataset; those items which can have adverse effect on results of computation are detected; they are used to access the hidden information from the dataset. For example, in the college data if any student who hadn't score good marks and got placed is an outlier or any student having good marks may not be selected is also an outlier; although they could have different reasons. Similarly if any customer does any money transaction or shopping which is an unusual then earlier record, there may be a fraud which needed to be detected at time and take the required action against it to stop online theft and to make online transactions safe.

Different methods:

For outlier detection there are various methods available which can be summed up into two main categories "*unsupervised, semi- supervised and supervised"* outlier detection methods. In supervised method data is classified into some classes, normal and abnormal, classification trained data and use it for outlier detection. There are different approaches for classification based outlier detection methods like "*global Vs local", "labeling Vs scoring"* etc.

In semi-supervised method half of the data is trained and rest half is classified by the designed framework; the item with abnormalities are uncovered in this method by using the half trained data.in unsupervised method or clustering based detection the data set having normal trend are clustered into clusters

C_i {info, N, Ct};

Where info gives the information about the type of cluster i.e. cluster or outlier, N gives the no. of items in the cluster, and Ct gives the center of the cluster.[11] Hierarchical clustering is a type of unsupervised abnormalities detection method; it clusters the data items as well as locates the outliers.

Other popular techniques are like density-based outlier detection technique, unconventionality from "frequent item sets", fuzzy outlier discovery techniques. There are some open source tools that have many in-built algorithms to find outliers like ELKI.

Applications:

Outlier detection is used in many fields like in college and universities to find abnormal student performance, in banks to detect frauds or online thefts, in medical checkups to detect the abnormal functioning of any body part, any intrusion in any restricted area or any event detection using sensors etc. generally it is used in preprocessing of data and in removing abnormalities or noise from the data set. Eliminating this noise lead to substantial increase in accurateness of the results. Besides this outlier detection has many more application in different areas like detecting eco-system disorders; in sports, performance of players may show abnormal values, different parameters are recorded for player's performance evaluation. In the process of detecting the errors of measurements in data originated from sensors.

6. CONCLUSION

Because of the speedy evolution in the educational data mining techniques there are several kinds of open source tools and techniques available at present to make the assessment easy. They are having diverse compatibility with many types of dissimilar applications [7]. The survey, in this research paper is done on the performance evaluation techniques used in data mining and outlier detection methods and techniques as well with Overall description of requirement and functions. It is expected that this review paper provides an understanding of the methods useful for student performance prediction, evaluation and anomaly discovery methods.

REFERENCES

1. Kerem Üçgün ,Ali Buldua,(2010) Procedia Social and Behavioral Sciences 2, "Data mining application on students' data".
2. Emine Ucar, Baha Sen.(2012) Procedia Technology 1, "Evaluating the achievements of computer Engg. deptt. of distance education students with data mining methods."
3. Hijazi, RSMM and Naqvi(2006). Bangladesh e-Journal of Sociology, "Factors Affecting Student's Performance: A Case of Private Colleges".
4. Romero and Ventura (2007), "Educational data mining: A Survey from 1995 to 2005', Expert Systems with Applications."
5. Er. Rimmy Choura (October 2012.), "Use of Data Mining Techniques for the evaluation of student performance: A Case Study".
6. C.Marquez-Vera, Ventura and Romero (2011). "Predicting School Failure Using Data Mining".
7. Nikitaben shelke,shriniwas gadage(volume5, issue 4, 2015). "International journal of advanced research in computer science and software engineering."
8. Wikipedia, "https://en.wikipedia.org/wiki/R_(programming_language)".
9. Wikipedia, "https://en.wikipedia.org/wiki/MATLAB".
10. Sudipto Guha, Rajeev Rastogi, Kyuseok Shim,(volume 25, issue 5) "ROCK: A Robust Clustering Algorithm for Categorical Attributes", information system archive, July, 2000.
11. Sheng-Yi Jiang, Ai-Min Yang, "Sixth International Conference on Fuzzy Systems and Knowledge Discovery" IEEE, 2009.
12. Abdelkarim Ben Ayed, Mohamed Ben Halima, Adel M. Alimi, "Survey on clustering methods : Towards fuzzy clustering for big data**",** international conference of soft computing and pattern recognition, 2014 IEEE.
13. M. Halkidi, Y. Batistakis, M. Vazirgiannis,(volume 2, issue 5) "Clustering algorithms and validity measures", 2001, IJARCSSE- IEEE.