International Journal of Technology and Engineering System (IJTES)
Vol 7. No.3 2015 Pp. 244-249

available at : www.ijcns.com
ISSN: 0976-1345

A DECISION SUPPORT SYSTEM FOR PERFORMANCE ANALYSIS OF STUDENTS THROUGH CLUSTERING

M.SINDHUJA , S.GAYATHR , R.B.AKSHAYA, I.R.PRAVEEN JOE
Student , Assistant Professor
Information Technology
KCG College of Technology
Chennai,India
sindhuja1903@gmail.com, gayathrijuly8@gmail.com,
akshaya110694@gmail.com, praveenjoeir@yahoo.com

ABSTRACT

It is the decision support system for the performance analysis of the students. The input data consists of marks(cgpa) and the behavioral test score, which is collected by conducting Myer's test for the students. Clustering is done on the collected data. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). This paper deals with the most popular density based clustering method called DBSCAN. In contrast to many methods, it features a cluster model called "density-reachability" and it is based on connecting points within certain distance measures. However, it only connects points that satisfy a density criterion, defined as a minimum number of other objects within this radius. The cluster consists of all density-connected objects (which can form a cluster of an arbitrary shape, in contrast to many other methods) plus all objects that are within these objects' range. The clusters are obtained. The eligible students and those who are likely to be placed comes under one cluster and then the other students category ,who are not likely to be placed are obtained at the last. Finally the intra-cluster distance among the cluster is low and the inter-cluster distance between the cluster is high. It is the decision making process for the department and the college to give further more improved training for the "Eligible students".

Keywords—Clusteranalysis,Clustering,DBSCAN,density reachability

I.INTRODUCTION

This paper mainly focuses on collecting data from the students, preprocessing it and then clustering is carried on the students data based on the inputs given, which specifies the eligible and those students who are likely to be placed and those who donot..An unsupervised clustering algorithm is expected to cater the requirement than a supervised one because it ends up in an unanticipated grouping where as the later demands fixed class labels in advance before grouping The next part of the paper gives a detailed note on the study made on the nature of clustering algorithm suitable for the problem.The reasons for choosing DBSCAN algorithm are listed. The working of the algorithms are provided. A brief summary of how the algorithms are used in the reference papers are also highlighted. The summary of approaches and algorithms followed for students clustering are discussed in this section. Section 3 portrays the model of the complete problem design and its implementation. Section 4 summarizes the results and the interpretation of the same. Section 5 concludes the paper with the future enhancement.

1.1 Clustering of Students

With clustering the groups (or clusters) are based on the similarities of data instances to each other. No already defined output class is used in training and the clustering algorithm is supposed to learn the grouping. Clustering is finding groups of objects such that the objects in one group will be similar to one another and different from the objects in another group . Clustering can be considered the most important unsupervised learning technique. Classification of clustering algorithm. In educational data mining, clustering has been used to group the students

according to their behavior for e.g. clustering can be used to distinguish active student from non-active student according to their performance in activities. However a suitable choice of an unsupervised algorithm and the proper building of features sets considering both functional and non functional characteristics would improve the quality of clustering [1].

1.2 Unsupervised Clustering

The Clustering function searches the input data for characteristics that frequently occur .It groups the input data into clusters. The members of each cluster have very similar properties. There are no preconceived notions of what patterns exist within the data. Clustering is a discovery process. Data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. "Clusters" is often used to describe mutually exclusive sub-segments without a list of predefined characteristics. If we have predefined labels it is only a classification. Hence it is required to perform clustering in an unsupervised way [2].

1.3 Density Based Clustering

The clustering methods like K-means or Expectation-Maximization are suitable for finding ellipsoid-shaped clusters. However, for non-convex clusters. these methods have trouble finding the true clusters, since two points from any different clusters may be closer than two points in the same cluster. The density-based methods which we consider in this chapter are able to mine such non-convex or shape-based clusters [3].

1.4 Dbscan

DBSCAN (for density-based spatial clustering of applications with noise) is a density based clustering algorithm. It is using the concept of "density reachability" and "density connect ivity "both of which depends upon input parameter- size of epsilon neighborhood e and minimum terms of local distribution of nearest neighbors. Here parameter e controls the size of the neighborhood and size of clusters. It usually starts with an arbitrary starting point that has not been visited DBSCAN algorithm is an important part of clustering technique which is mainly used in scientific literature. Density is measured by the number of objects which are nearest the cluster [4].

2.RELATED WORKS

2.1 K-means

K-means was proposed by Macqueen ,is one of the most popular partitioning methods. It partitions the dataset into k subsets, and k is already defined. The algorithm keeps adjusting the objects to the closest current cluster until no new assignments of objects to clusters is made. One Advantage of this algorithm is its simplicity. It also has several drawbacks in it. It is very difficult to specify number of clusters in advance. Sinceit works with squared distances, it is sensitive to outliers.

Another drawback is centroids and it is not meaningful in most problems[5].

2.2Hierarchical Clustering

Hierarchical clustering algorithm generally divide or merge dataset into a series of nested partitions. The way of the nested partitions can be either bottom-up or top-down. In the bottom up method, clustering is done with each single object in a single cluster and it continues to cluster the close pairs of clusters until all the objects are together in only one cluster. Top-down hierarchical clustering, on the other hand, starts with all objects in one cluster and keeps separating larger clusters into smaller clusters until all objects are separated into single cluster. Both the hierarchical methods show the most natural way of representing the clusters, called as dendrogram. Examples of this algorithms are ROCK, BIRCH (Balance Iterative Reducing and Clustering using Hierarchies), CURE (Cluster Using Representatives).It handles only numeric records and sensitive to data records[5].

3.ALGORITHM CHARACTERISTICS

Following are the drawbacks observed in the above works which are to be eliminated at present when choosing the clustering algorithm. However no particular clustering algorithm can be quoted as 'best', it all depends on the need and application. This is a result on the survey made on mathematical, supervised, unsupervised algorithms and other swarm based approaches.

4.PROPOSED ALGORITHMS

A. DBSCAN ALGORITHM DBSCAN (Density-Based Spatial Clustering of Application with noise) is density based cluster formation algorithm for spatial and non spatial high dimensional data base in the presence of outlier. The working is based on the following definitions, for more detail refer DBSCAN :

Def.1: The -neighborhood of an object p, denoted by N (p), is defined as total number of objects lying in the radius , i.e. N (p) = {q D | dist (p, q) ≤ }.

Def.2: An object p is said to be Core object if | N (p)| ≥ μ (minimum objects).

Def.3: An object p is said to be directly density reachable from an object q with respect to and μ if p N (q) and q is a Core object.

Def.4: An object p is said to density-reachable from an object q if there is a chain of objects p1. . . pn, p1= q, pn= p such that pi+1 is direct density-reachable from pi with respect to and μ.

Def.5: An object p is said to density-connected to an object q with respect to and μ if there is an object o such that both the p and q are density reachable from o with respect to and μ.

Def.6: An object which is lying at the border is not a Core object, but it will be a part of cluster. An object which is not lying in any of the cluster is treated as a noise object.
Def.7: A cluster X is non empty subset of database with respect to and μ, for every p, q: if p X, q is density reachable from p then q X and p is density- connected to q. DBSCAN detects density connected clusters by discovering one of its core object's p and computing all objects which are density-reachable from p. The collection of density reachable objects is performed by iteratively computing directly density reachable objects. DBSCAN checks the -neighborhood of each object p in the database. If N (p) of an object p consists of at least μ objects, i.e., if p is called as the core object, a new cluster X containing all objects of N (p) is created. Then, the - neighborhood of all objects q X, which have not yet been processed, is then checked. If object q is also a core object, the neighbors of q, which are not already assigned to cluster X, are added to X and their - neighborhood is checked in the next step. This procedure is then repeated until no new object can be added to the current cluster X. The DBSCAN algorithm proposed in this work is - 1: Label all the points as core, border, or noise points. 2: Eliminate the noise points. 3: Put an edge between all the core points that are within Eps of each other. 4: Make each group of the connected core points into a single cluster. 5: Assign each of the border point to one of the clusters with its associated core points.

5. DETAILED EXPLANATION

The basic key idea is that data objects in dense regions are clustered together. The algorithm uses a fixed value called as threshold value to decide the dense regions. It discovers the high density regions in space i.e. separated by very low region density. The disadvantage of the algorithm is that it captures only certain types of noise when clusters of different densities exist. Unlike other clustering techniques, it does not require the pre-specification of number of clusters. It also discovers clusters of arbitrary shape in spatial databases with noise .Variants of DBSCAN are: Incremental DBSCAN acts as the core algorithm of query clustering tool. SDBDC (Scalable Density-Based Distributed Clustering method, first it generally works on each local site and then clusters distributed objects on global site. Basic Definitions: a) Ɛ-neighborhood: The neighborhood is distance between two points in a cluster. The neighborhood in a cluster is less than the threshold input value, Ɛ. The neighborhood within a radius Ɛ of a given object is Ɛ-neighborhood. b) MinPts: It presents the minimum number of data objects in any cluster. c) Core Object: It refers Ɛ-neighborhood of an object contains at least MinPts of objects. d) Directly density-reachable: A data object p is directly density-reachable from the data object q if p is within the Ɛ-neighborhood of q and q is a core object. e) Density-reachable: An object p is density-reachable from the object q with respect to Ɛ and MinPts if there is a chain of objects p1,p2, . . . ,pn, where p1 = q and pn = q such that pi+1 is directly density-reachable from pi with respect to Ɛ and MinPts.[9] f) Density Connected: An object p is density connected to object q with respect to Ɛ and MinPts if there exists an object o D. g) Density Based Cluster: It is a set of density connected objects i.e .the maximal with respect to density-reachability. h) Border point: An object p is a border point if it is not a core object but density reachable from another core object. i) Noise: The objects not assigned in any cluster act as noise. The algorithm works as follows: It first checks Ɛ-neighborhood of each point in the space. If the Ɛ-neighborhood of the point p contains more than MinPts, a new cluster created in which p acts as the core object. The algorithm iterates and gathers all the objects within Ɛ distance from the core objects. The process then terminates when there is no new point to add to any cluster.

OUTLIER DETECTION DBSCAN deals with outliers (data objects which are different with the remaining set of the data). The algorithm avoids the noise or outlier to insert into the clusters. It's only capable to capture some types of outliers when different densities of clusters are present. This leads to a huge loss of important hidden information as sometimes the outlier are of particular interest. Examples are fraud detection, intrusion discovery.
Input: D = {t1, t2, t3 … tn} // Set of elements MinPts //Number of points in cluster Ɛ // Maximum distance for density measure Output: K = {K1, K2, K3 …, Kk } // Set of clusters Method: k=0; // initially there are no cluster for i = 1 to n do if ti is not in a cluster, then X = {tj | tj is density-reachable from t if X is a valid cluster, then k = k+1; Kk = X
DBSCAN deals with outliers (data objects which are different with the remaining set of the data). The algorithm avoids the noise or outlier to insert into the clusters. It's only capable of capturing some types of outliers when different densities of clusters are present. This leads to a huge loss of important hidden information as sometimes the outlier are of particular interest. Examples are fraud detection, intrusion discovery.
6.STEPS:
1.The input is obtained from the students(Marks, behavioral test)
2.The clustering algorithm is applied and the students clusters are formed.
3The clusters specifies those students who are likely to be placed and those who do not.
4.Finally the students data is represented in a interpretation tool and the analysis is done.
5.The intra cluster distance among the students in one cluster is less then the inter cluster distance among the clusters are high. The eligible students for placements is also obtained.

7.SYSTEM ARCHITECTURE:

Figure 1

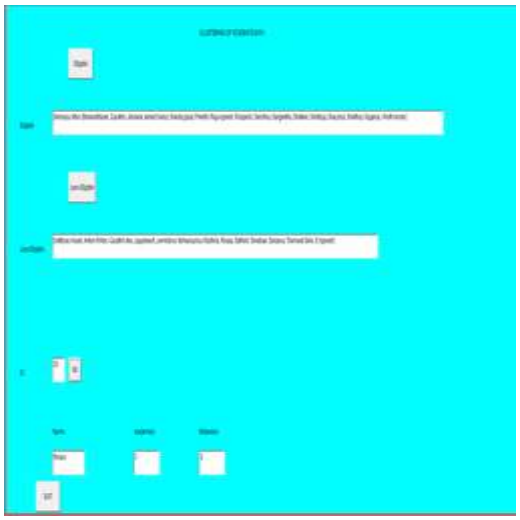
8.PERFORMANCE MEASURES:

1.The intra cluster similarity is low.

2.The inter cluster similarity is high.

9.RESULTS:

Figure 2. Clustering of students



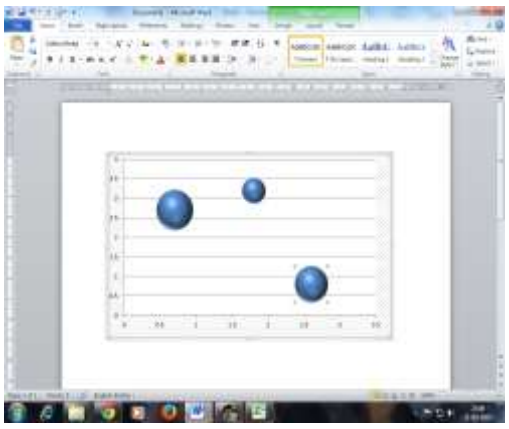


Figure 3 CLUSTERS

10.CONCLUSION & FUTURE WORK

In this paper, we have studied how data mining can be applied to educational systems. It shows how useful data mining can be in higher education, particularly to improve students' performance. We used students' data from the database of final year students' for Information Technology UG course. We collected all available data including their performance at university examination in various subjects. We applied data mining techniques to discover knowledge. Also we clustered the student into group using DBSCAN-clustering algorithm. Finally, noisy data were also detected. Each one of this knowledge can be used to improve the performance of student. For future work, a way to generalize the study to more diverse courses to get more accurate results needs to be developed. Also, experiments could be done using more data mining techniques such as neural nets, genetic algorithms, k-nearest Neighbour, Naive Bayes, support vector machines and others. Finally, the used preprocessed and data mining algorithms could be embedded into elearning system so that one using the system can be benefited from the data mining techniques.

11.REFERENCES

[1]Shannaq,B.,Rafael,Y.and Alexandro,V. "Student Relationship in higher education using data mining Techniques",Gobal Journel of computer science and technology,vol 10,no.11,pp.54-59.

[2]N.V.Anand Kumar and G.V.Uma,"Improving Academic performance of students by applying Data mining Technique,"European journal of scientific research,vol.34(4),2009.

[3]Edin Osmanbegović,Mirza Suljić,"DATA MINING APPROACH FOR PREDICTING STUDENT PERFORMANCE", Economic Review – Journal of Economics and Business, Vol. X, Issue 1, May 2012.

[4] K.Shanmuga Priya" Improving the student's performance using Educational data mining", International Journal of Advanced Networking and Application, Vol.4,pp-1680- 1685 (2013)

[5]Fhim A.M.Salem A.M.Torrkey F.A. and Ramadan M.A.,"An efficient enchanced K-means Clustering Algorithm,"pp 1626-1633,2006.

[6]. Md. Hedayetul Islam Shovon , Mahfuza Haque , "Prediction of Student Academic Performance by an Application of K-Means Clustering Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering Research Volume 2, Issue 7, July 2012 ISSN: 2277 128X.

[7].Mohammed M.Abu Tair,Alaa M.El-Hales,"Mining ,Eductaion Data to improve Student's Performance:A Case

Study, International Journal of Information and Communication Technology Research", Volume 2,No 2, February 2012

[8].Ayesha.S,Mustafa.T, Sattar.A and Khan.I, "Data Mining Model For Higher Education System", European Journal of Scientific Research", vol.43,no1,pp24-29.

[9]. Baradwaj.B, Pal.S, "Mining Educational Data to Analyze Student's Performance", International Journal of Advanced Computer Science and Applications, vol.2, no.6,pp.63-69

[10]. Han.J and Kamber.M, "Data Moining: Concepts and Techniques", The Morgan Kauffman Series in Data Management Systems,2nd edition