VOLUME-6 | ISSUE-2 | FEBRUARY-2017 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 78.46

Data mining algorithm used in Educational Datasets

Computer Science

KEYWORDS: Data mining algorithm, Educational Data Mining, K-means

Gauri Shanker Kushwaha — Research Scholar, MGCGV Chitrakoot, Satna(M.P.)

Bharat Mishra — Associate Professor, Department of Physical Sciences, MGCGV Chitrakoot, Satna(M.P.)

ABSTRACT

Data mining means to explore the hidden data from the large repositories of datasets with the use of data mining technique and tools. Educational Data Mining is a rising field exploring data in educational perspective by applying diverse data mining techniques/tools. It provides built-in knowledge of teaching and learning method for successful education preparation. This paper represents data mining algorithms can help out discovering appropriate knowledge enclosed in databases obtained from learning systems with comparative study of clustering techniques.

INTRODUCTION:

In recent years, the number of educational institutes that adopted an information system has been increasing quickly, one after the other the amount of data available in each educational institute's databases are also improved. Educational data mining is spontaneously useful to discover hidden in order from this data that would develop the quality of the whole educational system. Educational data mining can be applied to discover patterns in un–trusted datasets to computerize the decision making process of students, teachers, institutional leaders and quality teaching units [1].

The students are interested in understanding their needs and methods to improve the experience and performance. Teachers attempts to understand the learning process and the methods they can use to improve their teaching methods and applications of EDM to determine how to organize and structure the curriculum, the best methods to deliver course information and the tools to use to engage their students for best learning outcomes [2]. While institutional leaders focus on the development and the evaluation of data mining techniques for effectiveness and quality teaching units are responsible for allocating the resources for implementation in institutions. As institutions are increasingly held responsible for student success, the administering of educational data mining applications is becoming more common in educational settings [2]. As in higher education, many of the units are responsible for their datasets; hence it is necessary to select one appropriate algorithm technique from set of data mining algorithms. In this paper identifies different data mining algorithm and to select suitable EDM technique.

EDUCATIONAL DATA MINING

Data mining is the process of analyzing data from different forms or patterns and summarizing it into valuable information. It is also known as Knowledge Discovery in Database (KDD) useful information can be fetched from a large database, in the field of discovering the new techniques. Data mining have an ability to find existing relationship and pattern. Data mining consists itself with machine learning, statistics and visualization techniques to discover and extract knowledge [2, 6]. We know that data mining can use every sector like business, education, agriculture, marketing etc. Application of data mining in education sector is an emerging trend. The data mining terms, tasks, techniques and application can be used to developing data mining in education sector.

There is an increasing interest in the field of education, this emerging field called Educational Data Mining (EDM), which helps discovering knowledge and originates data in the education field. Educational data mining methods belong to a diversity of literatures. These literatures include data mining, machine learning, information visualization, and computational modeling. EDM applications will focus on allowing non-technical users use and engage in data mining tools and activities, making data collection and processing more accessible for all users of EDM [3].

The areas of EDM applications are analysis and visualization of data, providing feedback for supporting instructors, recommendations for students, predicting student performance, student modeling, detecting undesirable student behaviors, grouping students, social network analysis, developing concept maps, constructing courseware and planning and scheduling.

DATA MINING TECHNIQUES AND ALGORITHMS

Data mining techniques are used to find the hidden or new patterns to store the data. Clustering is used to define the data into groups or regions according to their specification or collected data and these groups are called clustered. Classification techniques are used to predefine the classification of data. Various approaches and techniques of data mining which can be applied on Educational data to build up a new environment to improve performance of existing data and help to create the new predictions on the data [4].

Data mining methods like prediction, clustering and relationship mining are mostly used in the field of marketing, agriculture and finance etc. These methods can be efficiently applied on educational data. Data mining having many types of techniques like clustering, classification and neural network etc. but in this paper we are considering only clustering techniques. These techniques can also be used with many other specific discovery techniques or algorithms. Clustering and classification both are very useful to improve the performance on education sector. Clustering can be used in Educational Data Mining (EDM) it can use the techniques like k-means, k-medoids, agglomerative, divisive. Using this technique student's performance can be predicted to improve current trends in higher education, motivating the students can be done by managing and processing the educational datasets.

Clustering Techniques

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). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings

(including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. The most common methods of clustering are partitioning, Hierarchical, density based and grid based methods [5, 7]. Characteristics of these four techniques are presented and compared in table 1.

Table 1: Characteristics of different clustering methods

SN	Clustering Techniques	Characteristics
1.	Partitioning Method	Mutually exclusive clusters of spherical shape, Distance based May use mean or mediods to represent cluster, Effective for small to medium size datasets
2.	Hierarchical Method	Clustering is multiple level decomposition, Can't correct erroneous merges or splits, May incorporate other techniques like micro clustering
3.	Density based method	Mutually exclusive clusters of spherical shape Distance based May use mean or mediods to represent cluster Effective for small to medium size datasets Clustering is multiple level decomposition Can't correct erroneous merges or splits May incorporate other techniques like micro clustering Can find arbitrarily shaped cluster Clusters are dense regions of objects, Each point have a minimum number of point, May filters out outliers
4.	Grid based method	Use a multi resolution grid data structure, Fast processing time

From the above discussion of different methods of clustering techniques, the most suitable and easiest method is partitioning method in the context of educational datasets. In the partitioning method two techniques are available for data mining namely k-means and k-modoids. The general properties of these techniques are given in brief.

k-means clustering is an algorithm to classify the objects based on attributes/features into K number of group. K is positive integer number. By minimizing sum of squares of distances between data and the corresponding cluster centroid grouping is done and the intention is to classify the data. The basic step of k-means clustering is simple and easier to use. In the beginning a number of K cluster determined and assumed that the centroid or center of these clusters. We can take any random objects as the initial centroid or the first K objects in sequence can also serve as the initial centroid and k-means algorithm will do the below given steps until convergence Iterate until constant:

1. Determine the centroid coordinate
2. Determine the distance of each object to the centroid
3. Group the object based on minimum distance

A clustering algorithm related to the k-means is k-medoids. The algorithms are attempt to minimize squared error and based on partitioning method, the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm k-medoids chooses data points as centers. k-medoids is a clustering technique which clusters is the data set of n objects into k clusters of classical partitioning technique. It is more robust to noise and outliers as compared to k-means. A k-medoids can be object of a cluster that average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the given data set, whose k-medoids clustering algorithm is as follows:

1. The algorithm begins with arbitrary selection of the k objects as medoids points out of n data points (n>k).
2. After selection of the k medoids points, associate each data object in the given dataset to most similar medoids.
3. Randomly select non-medoids object O.
4. Compute total cost, S of swapping initial medoids object to O.
5. If S<0, then swap initial medoids with the new one (if S<0 then there will be new set of medoids)
6. Repeat steps 2 to 5 until there is no change in the medoids.

Result & Discussion

For finding spherical-shaped clusters in small to medium-sized data points partition based algorithms work well. Execution time of k-Means algorithm is more efficient than k-medoids. It is observed that k-Means algorithm is efficient for smaller data sets and k-medoids algorithm seems to perform better for large data sets. If the number of data points is less, then the k-Means algorithm takes lesser execution time. But when the data points are increased to maximum the k-Means algorithm takes maximum time and the k-Medoids algorithm performs reasonably better than the k-Means algorithm. The characteristic feature of this algorithm is that it requires the distance between every pair of objects only once and uses this distance at every stage of iteration.

CONCLUSION:

Data mining having various tools and techniques available like as clustering, classification and neural network etc. for different compatibilities with different types of applications. On the basis of above study it is concluded that for educational data mining, the partitioning based method is suitable and under this k-means technique is most suitable and easy to use for EDM purpose.

REFERENCES:

1. Ananthi Sheshasaayee and C.Kabila, A Comparative Analysis of K-Means and K-Medoids Algorithm for Educational Data, International Journal for Scientific Research & Development, Vol. 4(05), 2016.
2. Suman and Mrs.Pooja Mittal, A Comparative Study on Role of Data Mining Techniques in Education: A Review, International Journal of Emerging Trends & Technology in Computer Science, Volume 3, Issue 3, 2014.
3. Shiwani Rana and Roopali Garg, Evaluation of Student's Performance of an Institute Using Clustering Algorithms, International Journal of App.Engineering Research, Volume11(5), pp 3605-3609,2016.
4. Garima Sehgal and Dr. Kanwal Garg, Comparison of Various Clustering Algorithms, International Journal of Computer Science and Information Technologies, Vol. 5 (3), 3074-3076, 2014.
5. Jiawei Han, Micheline Kamber and Jian Pei, Data Mining Concepts and Techniques, Morgan Kaufmann Publisher: An Imprint of Elsevier, 443-490, 2012.
6. Tagaram Soni Madhulatha, Comparison between K-Means and K-Medoid Clustering Algorithms, Communications in Computer and Information Science, Springer-Verlag Berlin Heidelberg 472-481, 2011.
7. Ling Liu and M. Tamer Özsu, Clustering on Streams, Encyclopedia of Database Systems-Springer, 379-400, 2009.