ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.6

Predicting Instructors Performance using Data Mining Techniques

Amanpreet Kaur [1], Amanpreet Kaur Batth [2]
Research Scholar of M.Tech[1], Assistant Professor[2]
Department of Computer Science and Engineering
Global Institute of Management & Emerging Technologies, Amritsar, Punjab, India

Abstract:
Data mining is very important in the field of Education especially when examining the students learning behaviour. It is very useful technique to analyze and uncover the hidden information about data sets which itself is very hard and time consuming process and it is not possible to done it manually. Most of the times research in educational mining focus on students' performance only but here we calculate the instructors' performance using various clustering techniques over classification for better accuracy.

Keywords: Data mining, Educational Data Mining, Clustering, Classification, SVM, KNN algorithm

1. INTRODUCTION

1.1 DATA MINING:

Data Mining is also known as Knowledge discovery. It is the practice of examining large pre-existing databases in order to generate new information. It is used to analyze and uncover hidden patterns of datasets. These data sets are currently used for wide range of applications like customer retention, education system, production control, healthcare, market basket analysis, manufacturing engineering, scientific discovery and decision making etc [1]. Data mining is playing a vibrant role in many applications like financial banking, manufacturing engineering etc. Frequent item sets have significant role in data mining which is used to find out the correlations between the fields of database [13]. Another name of the data mining is KDD (Knowledge discovery from the database).discovery of frequent item set is done by association rule. Retail store also used the concept of association rule for managing marketing, advertising, and errors that are presented in the telecommunication network.

Data Mining consists of following elements:

1) It takes out, convert and put transaction data onto the data ware house system.

2) Reserve and organize the data in multi-dimensional database system.

3) Business professionals can access the information.

4) Forecasting the data by various softwares.

5) Shows the data in a useful format, such as a graph or table.

There are special kinds of functionalities within the data mining. These are utilized for specifying certain kind of patterns which can help in identifying various tasks of the data mining process. There are two categories in which the data mining tasks can be classified. They are the descriptive as well as the predictive. The tasks which characterize the general purpose properties of the data within the database are known as the descriptive mining tasks.

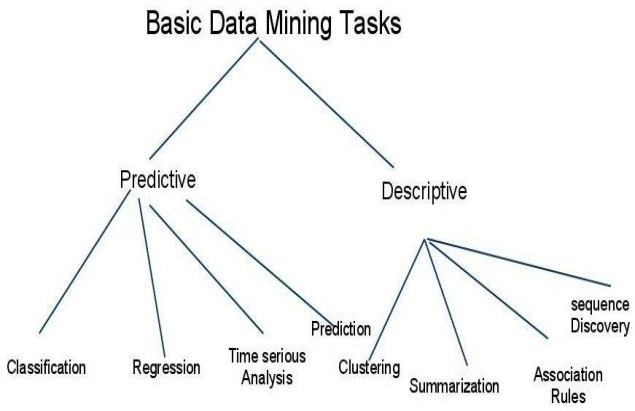

Figure.1. Data Mining Tasks

The various data mining functionalities are given below [2]:

1. Characterization and Discrimination: Data characterization is a representation of data of the class under study and data discrimination is a comparison of the target class with one or a set of comparative classes.

2. The Mining of Frequent Patterns, Associations and Correlations:-Frequent patterns are the patterns that occur frequently in data. Association rule mining is the process of finding interesting correlations, relationships among sets of items in various kinds of databases [3].

3. Classification and Regression:-Classification is a data mining (machine learning) technique used to predict group membership for data instances. Regression analysis is a methodology that is mainly used for numeric prediction.

4.Cluster Analysis:-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 to each other than to those in other groups (clusters).

5. Outlier Analysis:-Some objects in a data set do not comply with the general behaviour or model of the data. These data objects are outliers and analysis of outlier data is known as outlier analysis [4].

1.2 COMPONENTS OF DATA MINING

- Database and Data warehouse
- Database and Data warehouse Server
- Knowledge Base(KDD)
- Data mining Engine
- Pattern Evaluation Method
- Graphical User Interface

KDD process

The Discovery of knowledge in Databases process includes steps to gain unique knowledge.

Steps are:

1. Data which is not accurate and contains noise is cleaned. This step contains clearing the data.

2. At data integration step, heterogeneous data is combined with the different data sources.

3. In the selection step, applicability of analyzed data is taken into consideration.

4. Under the transform step, changes in the data are occurs with respect to various mining techniques.

5. Data Mining is used to find the required and unique patterns using many available techniques.

6. Within severely unique patterns, which includes acquaintance are recognized. This step involves, evaluation of required patterns.

7. In this last step the exposed results, which includes knowledge are represented.

1.3 EDUCATIONAL DATA MINING:

It is research technique designed for automatically extracting meaning from large repositories of data generated by people's learning activities in educational settings.

The goals of EDM (Educational Data Mining) are:

1. It predicts how students grasp the things.

2. Discovering or improving domain models to engage learners.

3. It studies about the educational support.

4. Calculating the instructor performances on the basis of students' grades.

1.4 CLUSTERING IN DATA MINING

Cluster analysis [11] has been widely used in numerous applications, including text mining, and another real world fields. Clustering can help marketers discover interests of their customers based on purchasing patterns and characterize groups of the customers. It can also be used to derive plant and animal taxonomies, categorize genes with similar functionality. In geology, specialist can employ clustering to identify areas of similar houses in a particular area etc. Data clustering can also be helpful in classifying documents on the Web to discover some new information.

Data clustering [14] is an unsupervised classification method aims at making groups of objects, or clusters, in this manner that objects in the same cluster are very similar and objects in different clusters are quite not associated. Now objects within a class have high resemblance to each other in the meantime objects in separate classes are more unlike.

Clustering is a method used to group similar documents, but it differs from categorization of documents are clustered on the fly instead of through the use of predefined topics. A basic clustering algorithm forms a vector of topics for each document and measures the weights of how healthy the document fits into each cluster [5]. Clustering comes under unsupervised classification. Unsupervised clustering is different from pattern reorganization in the area of statistics known as discriminate analysis and decision analysis which classify the objects from a given set of objects [8]. There are many clustering algorithms used for clustering. The major fundamental clustering methods can be classified into following categories [12]:

1. Partitioning Methods:-The general criterion for partitioning is a combination of high similarity of the samples inside of clusters with high dissimilarity between distinct clusters. Most partitioning methods are distance-based. Given k, the number of partitions to construct, this mehod forms some initial partitioning and then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. In this construct a partition of a data set containing n objects into a set of k clusters, so to minimize a criterion θ. The goal is, given a k, find a partition of k clusters that optimizes the chosen partitioning criterion. Here k is a input parameters. E.g. K-mean and K-centroid [1].

2. Hierarchical Methods:-In this method hierarchical decomposition of the given set of data objects is created. Agglomerative approach is the bottom up approach starts with each object forming a separate group. It then merges groups close to one another until all the groups are merged into one. In this type of clustering it is possible to view partitions at different level of granularities using different types of K. E.g. Flat Clustering [2].

3. Density Based Methods:-Most partitioning methods cluster objects based on distance between objects. Spherical shaped clusters can be discovered by these methods and encounter difficulty in discovering clusters of arbitrary shapes. It helps to discover arbitrary shape clusters. It also handles noise in the data. It is one time scan. It requires density parameters also [6].

4. Grid Based Methods:-Grid based methods quantize the object space into a finite number of cells that form a grid

structure. It is a fast method and is independent of the number of data objects and depends only on the number of cells in each dimension in the quantized space. In this objects together to form grid. Grid-based algorithms quantize the space into a finite number of grids and perform all operations on this quantized space. These approaches are dependent only on the number of segments in each dimension in the quantized space [7].

1.5 **Data Analytics**: Data Analytics is the way to examine the data sets and then draw some conclusions from it. Data Analytics can be used by many industries and organizations to get better business decision. Data analytics focuses on reasoning, the process of deriving a conclusion based solely on what is already known by the researcher. There are two types of data analytics:

These are:
1. Classification
2. Prediction

1. Classification: Classification models predict categorical class labels; and prediction models predict continuous valued functions.

2. Prediction: Prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

1.6 CLASSIFICATION IN DATA MINING

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks. In classification, there are many different methods and algorithms possible to use for building a classifier model. Some of the most popular ones can be counted as decision tree algorithms, support vector machines (SVM), artificial neural networks (ANN), discriminant analysis (DA), logistic regression, Bayesian belief networks, and rule based systems. In this study, the first four of these are used. A decision tree algorithm aims to recursively split the observations into mutually exclusive subgroups until there is no further split that makes a difference in terms of statistical or impurity measures.

1.6.1 CLASSIFICATION TECHNIQUES:

1. ID3 Algorithm

It is an algorithm used to form a decision tree from a given data set. Mostly it is used in machine learning and natural language processing. ID3 is abbreviated as Iterative Dichotomiser 3. It was invented by Ross Quinlan.

DECISION TREE:
- It classifies the data using some attributes.
- Tree consists of decision nodes and leafs.
- Nodes represent the value for attribute that is being tested.
- Leaf node produces the result.

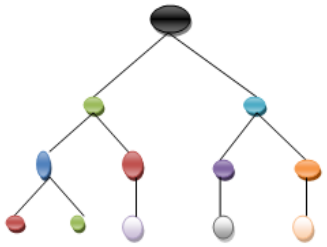

Figure.2. Decision Tree

2. SVM Algorithm

Support Vector Machines (SVMs) are a set of supervised learning methods used for classification, regression and outlier detection. It is a classification algorithm. It means we can use this method to predict if something belongs to a particular group. Support vectors are just the simple co-ordinates of individual observation. SVM has been used successfully in many real-worlds problems like text categorization, image classification, bioinformatics etc. However it is sensitive to noise and can consider only two classes.

3. KNN algorithm

KNN stands for K-nearest neighbor algorithm. This method is used for classification and regression predictive models. It is the simplest algorithm among all other machine learning algorithms. It is commonly used for its east of interpretation an low calculation time. It gives highly competitive results. The main idea behind nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point, and predict the label from these.The distance can be any metric measure: standard Euclidean distance is the most common choice. This method is successful in classification situations where the decision boundary is very irregular because it is a non-parametric method.

4. ANN Algorithm

ANN is a computational model based on biological neural networks. It is a non linear statistical data modeling tool where the complex relationships between inputs and outputs are modeled and patterns are found. It actually learns from observing data sets. ANN takes data samples rather than entire data sets to arrive at solutions, which saves both time and money. ANNs have three layers that are interconnected to each other. The first layer consists of input neurons. Those neurons send data onto the second layer and then this layer sends output neurons to the third layer. Second layer is also known as Hidden layer. The many advantages of neural networks include Adaptive learning, Self-Organization etc. It processes the information in a similar way the human brain does.

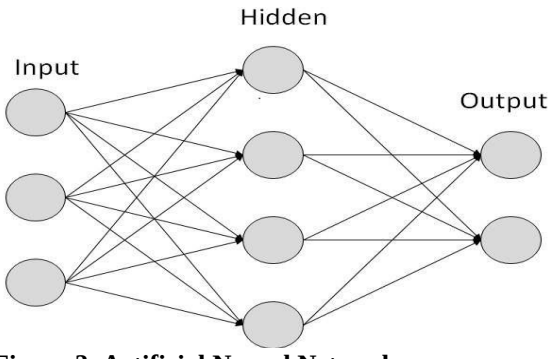

Figure.3. Artificial Neural Networks

5. LITERATURE REVIEW

AUTHOR & TITLE	YEAR	DESCRIPTION	OUTCOMES
Akhilesh Kumar Yadav, Divya Tomar, Sonali Agarwal, "Clustering of Lung Cancer Data Using Foggy K-Means"	2013	In this paper the data set has been taken from SGPGI. The main focus of this paper is to develop a novel technique based upon foggy k-mean clustering[1].	The result of the experiment depicts that foggy k-means clustering algorithm has excellent result on datasets which are real as compared to simple k-means clustering algorithm and provides a enhanced result to the real world problem.
Chew Li Sa et.al, "Student Performance Analysis System"	2013	In this paper [3] they proposed a system named Student Performance Analysis System (SPAS) to keep track of student's result in a particular university.	The proposed system offers student performance prediction through the rules generated via data mining technique. The data mining technique used in this project is classification, which classifies the students based on their grades.
Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques"	2010	In this paper they presented an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques [4]. After that they have investigated three data mining techniques: Naïve Bayes, back propagated neural network and C4.5 decision tree algorithm.	It has been concluded that the C4.5 algorithm has better results than other two techniques.
Qasem A. et.al, "Predicting Stock Prices using data mining techniques"	2013	In this paper they try to help investors in stock market better timing for the buying and selling stocks on the basis of knowledge of past historical experiments [5].	In this, they define decision tree classifier which is one of the best data mining techniques.
K.Rajalakshmi et.al, "Comparative Analysis of K-Means Algorithm in Disease Prediction"	2015	In this paper they represented an extremely fast growing field of medical. A huge amount data has been generated by this field every day [6]. To handle this data is very difficult, so there is a need of a technology to handle this data.	This paper analyzes various disease predictions techniques using K-means algorithm. This data mining based prediction system are reduces the human effects and cost effective one.
Daljit Kaur and Kiran Jyot, "Enhancement in the Performance of K-means Algorithm"	2013	This paper explained that clustering is a division of data into groups of similar objects. Each group consists of objects that are similar among them and dissimilar compared to objects of other groups. K-means algorithm is widely used for clustering data [9].	The proposed method decreases the complexity and effort of numerical calculations. It also solves the problem of dead unit.

2. CONCLUSION AND FUTURE SCOPE:

In this paper, it is been concluded that Educational Data Mining is the useful field for mining the data about students and instructors. In this paper we concentrate on the instructor's performance. Various classification and clustering techniques are explained in this paper. The future scope provides the enhancement and efficiency of data in the system.

3. REFERENCES

[1]. Akhilesh Kumar Yadav, Divya Tomar, Sonali Agarwal, "Clustering of Lung Cancer Data Using Foggy K-Means", International Conference on Recent Trends in Information Technology (ICRTIT) 2013

[2]. Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey "Weather Forecasting using Incremental K-means Clustering", 2014

[3]. Chew Li Sa; Bt Abang Ibrahim, D.H.; Dahliana Hossain, E.; bin Hossin, M., "Student performance analysis system (SPAS)," in *Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on* , vol., no., pp.1-6, 17-18 Nov. 2014

[4]. Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques", Washington DC 20052, 2010

[5]. QASEM A. AL-RADAIDEH, ADEL ABU ASSAF 3EMAN ALNAGI, " Predictiong Stock Prices Using Data Mining Techniques", The International Arab Conference on Information Technology (ACIT'2013)

[6]. K.Rajalakshmi, Dr.S.S. Dhenakaran, N.Roobin "Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 7, July 2015

[7]. Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security, Vol. 7, o. 1, 2010

[8]. Bala Sundar V,T Devi, N Saravan, "Development of a Data Clustering Algorithm for Predicting Heart", International Journal of Computer Applications (0975 – 888) Volume 48– No.7, June 2012

[9]. Daljit Kaur and Kiran Jyot, "Enhancement in the Performance of K-means Algorithm", International Journal of Computer Science and Communication Engineering, Volume 2 Issue 1, 2013

[10].Siddheswar Ray and Rose H. Turi, "Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation", School of Computer Science and Software Engineering Monash University, Wellington Road, Clayton, Victoria, 3168, Australia, 1999.

[11]. Azhar Rauf ,Mahfooz, Shah Khusro and Huma Javed "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", *Middle-East Journal of Scientific Research* 12 (7): 959-963, 2012 ISSN 1990-92332012

[12]. Madhu Yedla, T M Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center*",* *International Journal of Computer Science and Information Technologies*, Vol. 1 (2) 2010, page 121-125

[13]. K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, *Proceedings of the World Congress on Engineering* , Vol IWCE 2009, July 1 - 3, 2009, London, U.K

[14]. Osamor VC, Adebiyi EF, Oyelade JO and Doumbia S "Reducing the Time Requirement of K-Means Algorithm" *PLoS ONE,* Volume 7, Issue 12, pp-56-62, 2012.

[15]. Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed, "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity," *Middle-East Journal of Scientific Research,* pages 959-963, 2012.

[16]. Kajal C. Agrawal and Meghana Nagori, "Clusters of Ayurvedic Medicines Using Improved K-means Algorithm," *International Conf. on Advances in Computer Science and Electronics Engineering*, 2013.