JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

A Study of Clustering Algorithm for Student Analysis

Bhawna Janghel[1*], Asha Ambhaikar [2]

[1,2] Department of Computer Science, Kalinga University, Raipur,India

DOI: https://doi.org/10.26438/ijcse/v7i6.937940 | Available online at: www.ijcseonline.org

Accepted: 10/Jun/2019, Published: 30/Jun/2019

Abstract**—**In this paper using k-mean clustering method use for students school academic performance are measured by quarterly exam, half yearly exam, and final exams result. So, by taking the marks of three of exams, we can compare the final result of govt. school data and private school data. By using data clustering technique we can predict which school is best.And try to identify the weak student of particular school and will identify the result of best school.This will lead to the identification of best between private & government school in town.Strategies and techniques of best school will be followed which will help in making the education system better.

Keywords**—** Data clustering , k-mean, academic performance etc

I. Introduction

Data Mining

Data mining is the process of extracting the useful information, which is stored in the large database.It is a powerful tool, which is useful for organizations to retrieve the useful information from available data warehouses.Data mining can be applied to relational databases, object-oriented databases, data warehouses, structured-unstructured databases, etc.Data mining is used in numerous areas like banking, insurance companies, pharmaceutical companies etc.[1]

Patterns in Data Mining

1.Association

The items or objects in relational databases, transactional databases or any other information repositories are considered, while finding associations or correlations.

2. Classification

The goal of classification is to construct a model with the help of historical data that can accurately predict the value.It maps the data into the predefined groups or classes and searches for the new patterns.

3. Regression

Regression creates predictive models. Regression analysis is used to make predictions based on existing data by applying formulas.Regression is very useful for finding (or predicting) the information on the basis of previously known information.

4. Cluster analysis

It is a process of portioning a set of data into a set of meaningful subclass, called as cluster.It is used to place the data elements into the related groups without advanced knowledge of the group definitions.

5.Forecasting

Forecasting is concerned with the discovery of knowledge or information patterns in data that can lead to reasonable predictions about the future.

Technologies used in data mining

Lots of techniques used in the development of data mining methods. Some of them are mentioned below:

a) 1. Statistics:

It uses the mathematical analysis to express representations, model and summarize empirical data or real world observations.Statistical analysis involves the collection of methods, applicable to large amount of data to conclude and report the trend.

b) 2. Machine learning:

Arthur Samuel defined machine learning as a field of study that gives computers the ability to learn without being programmed.When the new data is entered in the computer, algorithms help the data to grow or change due to machine learning.In machine learning, an algorithm is constructed to predict the data from the available database (Predictive analysis).It is related to computational statistics.

The four types of machine learning are:
1. Supervised learning : It is based on the classification.It is also called as inductive learning. In this method, the desired outputs are included in the training dataset.

2.Unsupervisedlearning: Unsupervised learning is based on clustering. Clusters are formed on the basis of similarity measures and desired outputs are not included in the training dataset.

3.Semi-supervisedlearning: Semi-supervised learning includes some desired outputs to the training dataset to generate the appropriate functions. This method generally avoids the large number of labeled examples (i.e. desired outputs) .

4. Active learning: Active learning is a powerful approach in analyzing the data efficiently.The algorithm is designed in such a way that, the desired output should be decided by the algorithm itself (the user plays important role in this type).

c) 3. Information retrieval :

Information deals with uncertain representations of the semantics of object (text, images).

4. Database systems and data warehouse : Databases are used for the purpose of recording the data as well as data warehousing.Online Transactional Processing (OLTP) uses databases for day to day transaction purpose.To remove the redundant data and save the storage space, data is normalized and stored in the form of tables.Entity-Relational modeling techniques are used for relational database management system design.Data warehouses are used to store historical data which helps to take strategical decision for business.It is used for online analytical processing (OALP), which helps to analyze the data.

d) 5. Decision support system

Decision support system is a category of information system. It is very useful indecision making for organizations.It is an interactive software based system which helps decision makers to extract useful information from the data, documents to make the decision.

KDD Data mining

The process of discovering knowledge in data and application of data mining techniques are referred to as knowledge Discovery in Database (KDD).KDD consists of various application domains such as artificial intelligence,pattern recognition,machine learning and data visualization.The main goal of KDD is to extract knowledge from large database with the help of data mining methods.

The different steps of KDD are as given below:

1.Data cleaning:

In this step,noise and irrelevant data are removed from the database.

2.Data integration :

In this step,the heterogeneous data sources are merged into a single data source.

3.Data selection:

In this step,the data which is relevant to the analysis process gets retrieved from the database.

4.Data transformation :

In this step,the selected data is transformed in such forms which are suitable for data mining.

5.Data mining:

In this step,the various techniques are applied to extract the data patterns.

6.Pattern evalution:

In this step,the different data patterns are evaluated.

7.Knowledge representation:

This is the final step of KDD,which represents the knowledge.

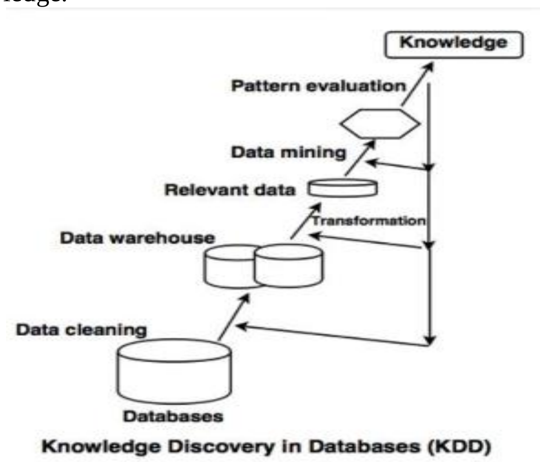

Fig.1

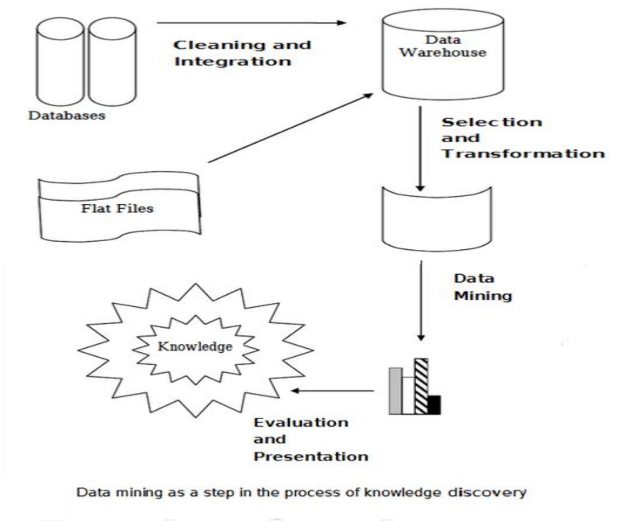
Databases
Cleaning and Integration
Data Warehouse
Flat Files
Selection and Transformation
Data Mining
Knowledge
Evaluation and Presentation
Data mining as a step in the process of knowledge discovery

Fig.2

II. RELATED WORK

On the basis of review literature the problem is identified to build data mining model by using clustering method use of clustering method for students performance and mostly papers prediction is the based on students results, but sometimes prediction may be wrong because the prediction is based on their previous result, more dataset instance will be collected and will be compared and analyzed with other data mining techniques such as association and clustering.

III. Methodology

K-MEANS CLUSTERING

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

where,

'||xi - vj||' is the Euclidean distance between xi and vj.

'ci' is the number of data points in ith cluster.

'c' is the number of cluster centers.

Algorithmic steps for k-means clustering

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, 'ci' represents the number of data points in ith cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

Advantages

1) Fast, robust and easier to understand.

2) Relatively efficient: O(tknd), where n is # objects, k is # clusters, d is # dimension of each object, and t is # iterations. Normally, k, t, d << n.

3) Gives best result when data set are distinct or well separated from each other.

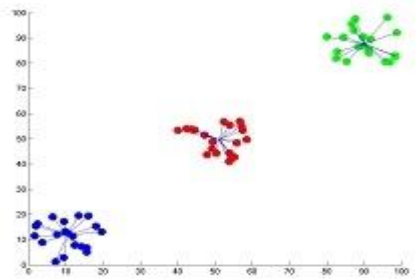

Fig.3

The result of k-means for 'N' = 60 and 'c' = 3

Note: For more detailed figure for k-means algorithm please refer to k-means figure sub page.

Disadvantages

1) The learning algorithm requires apriori specification of the number of cluster centers.

2) The use of Exclusive Assignment - If there are two highly overlapping data then k-means will not be able to resolve that there are two clusters.

3) The learning algorithm is not invariant to non-linear transformations i.e. with different representation of data we get different results (data represented in form of cartesian co-ordinates and polar co-ordinates will give different results).

4) Euclidean distance measures can unequally weight underlying factors.

5) The learning algorithm provides the local optima of the squared error function.

6) Randomly choosing of the cluster center cannot lead us to the fruitful result. Pl. refer Fig.

7) Applicable only when mean is defined i.e. fails for categorical data.

8) Unable to handle noisy data and outliers.

9) Algorithm fails for non-linear data set.

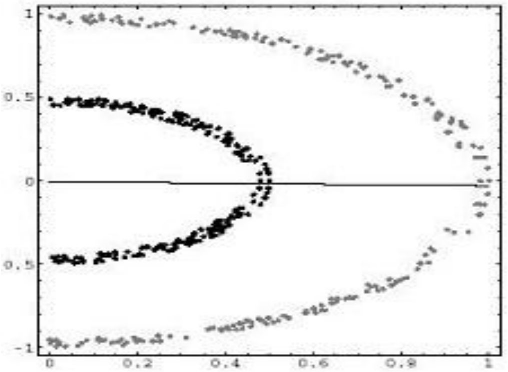

Fig.4
The non-linear data set where k-means algorithm fails

IV. Conclusion and Future Scope

The processed data will be analyzed using different data mining techniques like, classification, clustering, association rule mining etc. In this process ,we will compare the result of both schools(govt. school or private school) and will find out which school has better long term results and follow the techniques on remaining school for quality education.

References

[1]. .Datamining Tutorial,Home>BigData&Analytics>Datamining,http://www.tutorialride.com>datamining.

[2]. https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm

[3]. 3.Oyelade, O. J ,Oladipupo,O.O,Obagbuwa.I.C (IJCSIS), Application of k-Means Clustering algorithm for prediction of Students' Academic Performance, Vol. 7, _o. 1, 2010,

[4]. Sunita B Aher, Mr. LOBO L.M.R.J.,Data Mining in Educational System using WEKA, (ICETT) 2011

[5]. Bindiya M Varghese, Jose Tomy J, Unnikrishnan A,Poulose Jacob K,Clustering Student Data to Characterize Performance Patterns, (IJACSA)

[6]. Mohammed M. Abu Tair, Alaa M. El-Halees, Mining Educational Data to Improve Students' Performance: Volume 2 No. 2, February 2012

[7]. Dorina Kabakchieva, Student Performance Prediction by Using Data Mining Classification Algorithms, Vol 1 Issue 4 November 2012

[8]. Surjeet Kumar Yadav,Saurabh Pal,Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification, Vol. 2, No. 2, 51-56, 2012

[9]. Dr. Sudhir B. Jagtap, Dr. Kodge B. G.,Census Data Mining and Data Analysis using WEKA, (ICETSTM – 2013)

[10]. P.Veeramuthu#1, Dr.R.Periyasamy#2, V.Sugasini, Analysis of Student Result Using Clustering Techniques, (IJCSIT) Vol. 5 (4) , 2014

[11]. M. Durairaj , C. Vijitha , Educational Data mining for Prediction of Student Performance Using Clustering Algorithms, (IJCSIT) Vol. 5 (4) , 2014

[12]. Kashish Kohli,Shiivong Birla,Data Mining on Student Database to Improve Future Performance, 15, July 2016

[13]. Mr. Shashikant Pradip Borgavakar, Mr. Amit Shrivastava, Evaluating Student's Performance using K-Means Clustering, Vol. 6 Issue 05, May – 2017

[14]. Dr. K. Karthikeyan, P. Kavipriya, On Improving Student Performance Prediction in Education Systems using Enhanced Data Mining Techniques, Volume 7, Issue 5, May 2017

[15]. Hilal Almarabeh, Analysis of Students' Performance by Using Different Data Mining Classifiers, 2017.08.02

[16]. Hafez Mousa1, Ashraf Maghari2, School Students' performance Predication Using Data Mining Classification, Vol. 6, Issue 8, August 2017

[17]. K. Govindasamya and T. Velmuruganb, A Study on Classification and Clustering Data Mining Algorithms based on Students Academic Performance Prediction, Volume 10 • Number 23 • 2017,

[18]. Abdelbaset Al-Masri, Experiences in Mining Educational Data to Analyze Teacher's Performance: A Case Study with High Educational Teachers, 2017.10.12.01

Authors Profile

Mrs. Bhawna Janghel pursed Master of Computer Application from CSVTU University of Chhattisgarh in 2010 .She is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Sciences.She has 9 years of teaching experience.

Dr. Prof. Asha Ambhaikar , Professor and Dean Students Welfare, Kalinga University, Naya Raipur. She has also worked as a Principal in G. H. Raisoni college of Engineering and Management, Amravati (Maharashtra). She has 25 years of Academic experience. She has Guided 3 Ph.D Scholars and 8 undergoing. She has published more than 75 research papers in reputed National and international Journals. She was a chairman Board of studies and Member of Academic Council of Information Technology in Chhattisgarh Swami Vivekananda Technical University, Bhilai(C.G.). She is a member of Editorial Board and Reviewer of various Reputed international journal's and conferences. She is also the member of various professional societies like Life member of IAENG (International Association of Engineers, Hong Kong, IEEE, Indian Society of Technical Education (ISTE), Computer Society of India (CSI), IET, ASDF Computer Science Teachers Association (CSTA), Association for Computing Machinery (ACM), New York, USA, IACSIT (International Association of Computer Science and Information Technology, Singapore. Member of SDIWC (The Society of Digital Information and Wireless Communication, USA. She has also chaired various National and International Conferences around various countries as a keynote speaker. She has also published two books by Lambert Publication, Germany. She has also received a various Awards like: 1. Best Personality of India 2015 at New Delhi, India. 2. Bharat Excellence Award 2015 at New Delhi, India. 3. Outstanding Teacher's Award 2014, 2015 on 5th September, at RCET Bhilai, India. 4. ASDF Global Award for Best Dean (Academics) of the Year 2014 at Bangkok- Thailand on 30th December 2014. 5. ASDF Global Award for Best Professor of the Year 2013 at Pondicherry, India. 6. Best Research paper Award in the year 2009. 7. SPARC Europe Award 2009 for the research paper "Exploring the Behavior of Mobile adhoc Network Routing Protocols with reference to Speed and Terrain Range". She is also guiding Ph.D. Scholars in various universities, Her area of research includes Computer Networking, Mobile Adhoc Networking, Sensor Networks, Data Mining, Distributed system, information systems and security and Cloud Computing etc.