Mody University International Journal of Computing and Engineering Research
2 (2), 2018, 74-77
http://www.publishingindia.com/micer

A Survey on K-Means Clustering in Various Domains

Shivangi Singla[1*] and Pinaki Ghosh[2]

[1]Research Scholar, Department of Computer Science and Engineering, Mody Univesity of Science and Technology, Lakshmangarh, Sikar, Rajasthan, India. Email: shivangisingla306@gmail.com
[2]Assistant Professor, Department of Computer Science and Engineering, Mody Univesity of Science and Technology, Lakshmangarh, Sikar, Rajasthan, India. Email: pinakighosh.cet@modyuniversity.ac.in
*Corresponding Author

Abstract: Data mining is used to extract the hidden patterns from large datasets and extracted patterns are helpful to identify knowledge about data to users. As various approaches are there for data mining named Clustering, Classification, Association rule mining, etc. Amongst all we consider clustering, which is an unsupervised learning and grouping. This paper demonstrates clustering technique named k-means clustering and its various improvements in different domains exterminate the limitations of traditional k-means clustering. K-means clustering is the simple partitioning clustering algorithm and exhibit many limitations, so it is very important to understand various enhancements for constructing hybrid algorithms to improve accuracy of algorithms. Various areas are defined where k-means clustering is widely used nowadays such as in healthcare, improving academic performance and optimization of search engine and much more.

Keywords: Academics, Clustering, Data mining, Healthcare, K-means, Search engine.

I. Introduction

Data mining is the process of collecting hidden knowledge and different patterns from a huge amount of data. The main objective of data mining is to process the raw data and extract the valuable information from it. The outcome of the particular situation or problem in data mining is analyzed from the past data [1]. In other words, it has the potential for determining large datasets, extracting hidden relationships between different attributes and outlining the extracted knowledge more useful to data users. In the middle of the 1990's, it comes as a strong tool for extracting useful information from a large amount of data. According to some researchers Knowledge Discovery in Database (KDD) and data mining seems to be related with each other, but several other researchers believe that they both are different terms as KDD is the process and data mining is one of the stages of the KDD. KDD is a non-trivial extraction of useful information from raw data. KDD is an organized process which includes different stages like selection, preprocessing, transformation, data mining and knowledge discovery or interpretation [2]. The KDD process and their different stages are shown in Fig. 1.

The KDD process consists of the following stages:

1) Data Cleaning and Integration: In this step, firstly, the noisy data and irrelevant data are eliminated and then the data is integrated from multiple sources.

2) Selection and Transformation: The data is selected which is relevant and extracted from the data collection and then this data is transformed into data mining procedure.

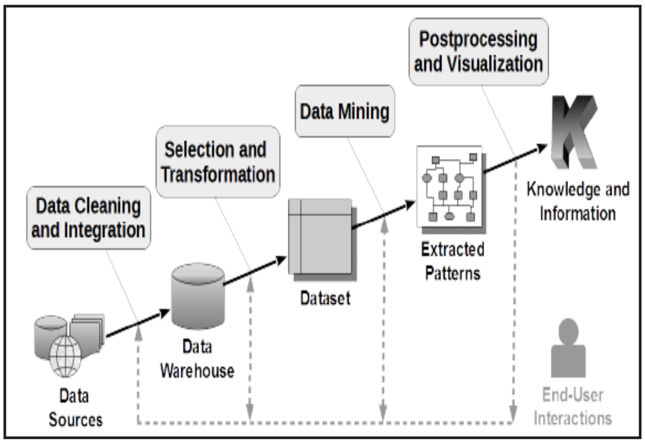
Postprocessing and Visualization
Data Mining
Selection and Transformation
Data Cleaning and Integration
Knowledge and Information
Extracted Patterns
Dataset
Data Warehouse
Data Sources
End-User Interactions

Fig. 1: Different Stages of Knowledge Discovery in Databases

3) Data Mining: It is the technique where hidden patterns are determined with the help of various methodologies.

4) Postprocessing and Visualization: Different patterns are identified which represents knowledge and it is represented to the end user to understand the results of data mining.

Data mining includes various methods, namely classification, clustering, regression, association rule mining [2]. Among all techniques and methods of data mining, clustering is a useful technique. Clustering is called unsupervised learning because there is no information about class labels, hence clustering is a type of learning by observation rather learning by examples.

Article can be accessed online at http://www.publishingindia.com

Clustering is the procedure for grouping data objects into disjoint groups and these groups are called clusters. The main property of clustering the data object is that the data objects within a cluster have the highest similarity measure, but the high dissimilarity measure with the data objects in other clusters. These similarity and dissimilarity measure is defined by the attributes of the data objects and distance measures. In data mining, the focus is on how to make the effective and efficient cluster analysis in large databases. So the typical requirements for clustering are scalability, ability to deal with noisy data, the discovery of clusters with arbitrary shape, ability to deal with different types of attributes, etc. Clustering techniques are divided into four methods, namely, partitioning methods, Grid-based methods, Hierarchical methods and Density-based methods [3].

The most simple and effective method of clustering is Partitioning method. It partitions the data objects into groups or clusters on the basis of distance measure. It includes two methods named K-means and K-medoids methods. K-means clustering is the simple, fast and efficient algorithm [4]. In early 1967, the K-means algorithm was proposed by James MacQueen. According to MacQueen, “K” describes a number of clustered datasets and “means” belongs to mean value, an average of data objects, which is center of clusters. This algorithm is widely used for large data objects and used in various areas like information retrieval, healthcare, optimization of the search engine, pattern recognition as for many applications, it includes multiple datasets and attributes [5]. It is also known as Centroid-Based technique as it accounts for the centroid of a cluster for symbolizing the cluster. There are various ways to determine the centroid as by mean or by mediod of data objects. The K-means algorithm is an iterative process and continues until there is no change in the cluster values. The algorithm is described as follows:

Inputs: k- Number of clusters, D- Data points of objects.

Method:

1. Randomly choose number of k from D.
2. To divide data points into k clusters, associate each data point in D with nearest centroid.
3. Recalculate the position of centroids. Repeat steps 2 and 3 until no change in their cluster values.

Outputs: Different clusters with data points.

Properties of K-means clustering are terminates at local optima, numeric values are used, shape of the cluster is convex and large datasets are processed in an efficient manner [6]. Their are two main limitations of K-means clustering. First, the number of clusters is not defined as initially we have to assume the number of clusters and second, initialization of centroids. Other limitations of k-means clustering: fails for categorical data, unable to handle noisy data, fails for the non-linear dataset. All these limitations are to be considered while approaching to k-means clustering algorithm.

Applications of K- means clustering-

1) Optimization of Search Engine: Search engine optimization is the process to enhance the perceptibility of the searched results on a search engine for a web page or a website. Different types of items such as articles, videos, images, documents are being targeted [7]. The major issue is to get back the favourable website from large number of websites [8]. The number of available Web pages is growing day by day, so it is very crucial for users to find documents appropriate for them. When a query is made by the user, search engines return millions of documents or pages in answer to it. Clustering of the search result is the appropriate way, to sum up, the huge amount of documents or pages in the form of clusters [9]. With the help of clustering, there will be an improvement in the quality of websites. Clustering includes partitioning a set of objects into a described number of groups.

2) Improving Academic Performance: The academic community of higher learning is facing a critical issue of monitoring the progress of students’ academic performance [10]. The complete description of the student performance is necessary to define the various ways of learning and teaching [11]. Major tasks are in student placement, admission and in the curriculum. Admission and placement are considered to be the very significant process where the data are collected and analyzed. The university ranking completely depends on students’ performance and placement [12]. Evaluation of performance on the basis of marks is a support to guide the improvement in student performance. It is a very tough job to categorize the students into different groups on the basis of their performance. Data mining is thus used for improving the evaluation system. Clustering is the best technique for the evaluation of students’ performance to detect the key attributes of students’ performance and it may be used for future prediction [10]. Data clustering technique is useful in building the knowledge gap in higher education system [13]. It will be useful for the teachers to eliminate the drop out ratio and improve the students’ performance.

3) In Healthcare: In healthcare, data mining is becoming very popular as unknown and valuable information in health data is to be detected. Data mining techniques can be helpful to find out the causes of various diseases, unknown diseases, and to identify medical treatment methods. Earlier, the large amount of patient data has been collected which does not produce any useful and hidden information, and thus for building effective decisions data mining come into subjection. It supports in making efficient policies on health care, formulating proper drug recommendations and establish health profiles of individuals [14]. The study of health data enhances the patient management performance. Data mining techniques include clustering, which is used to provide an advantage to healthcare management by grouping the patients containing the same type of health issues for effective treatments. By applying data mining techniques, the hidden patterns among patients can be extracted. It helps in building an effective plan for the efficient information system and predicting the patients stay in the hospital for medical diagnosis [15].

In section 2, we talk about the different research papers on optimization of search engine using K-means clustering. In section 3, we have done a literature survey on improving academic performance using K-means clustering. In section 4, we explain survey on examining healthcare using K-means clustering and finally, we summarize our paper.

II. Optimization of Search Engine Using K-Means Clustering

Ezaz Ahmed [8], proposed to cluster the top 5 websites in one cluster and another in the second cluster using the k-means clustering algorithm. The author describes that by clustering, the quality of websites is improved by grouping similar websites in groups. The data are collected on the basis of various websites source code, namely, the length of the title, number of keywords in the title, domain length, number of backlinks and top rank websites. The tool here used by the author is Weka tool. According to these attributes, the accuracy is discovered and clusteres are build.

Hasitha Indika Arumawadu [9], illustrates the improvement in the segmentation of web search results using the k-means clustering algorithm. It uses a vector space model for the representation of documents and for measuring the similarity between the user query and search results, cosine similarity is used. There is a major drawback of k-means clustering is to decide the initial cluster points and to overcome this distortion curve method is called. For better results and better speed now search engine designers are accounting big data concept.

K. O. Khorsheed [7], aimed the clustering of the database to improve the search time of search engines using the K-Means algorithm. As experiments are carried out, search time increases due to increment in the size of data and the search process become slower. To overcome this, the author proposed a technique called MPAPI (Message Passing Application Programming Interface) with K-Means clustering. The use of MPAPI is established to make the parallel processing of different clusters at a time to reduce the time taken by the clusters to build or processa cluster at a time. The process is verified and proves that the accuracy of k-means using MPAPI is solemnly increased.

III. Improving Academic Performance Using K-Means Clustering

Oyelade, O. J. [10], presented a significant and simple technique for analyzing students' result data. The author described the K-means clustering algorithm for this purpose and for measurement of similarity distance Euclidean distance is used. The technique was joined by a deterministic model to determine the students' results for making an adequate decision by academic planners. The performance matrix is defined and by changing the values of k, the performance of the student is evaluated.

Md. Hedayetul Islam Shovon [13], proposed that in university or institution students' performance is evaluated by external and internal assessment. External assessment includes previous semester grade point and final semester grade point while internal assessment includes performance in lab, quiz, assignments, attendance. The author used the K-means clustering technique to group the student data using internal assessment and previous exam grade point and with the help of this data, we can predict the final grade point of a student. The clusters in which the students are classified are High, Low, Medium.

Dr. T. Meyyappan [16], aimed to use the K-means clustering algorithm on the student database and clusters are built on their mean value. It helps to conclude the performance evaluation of student on the basis of their marks, year, subject and series wise. The student database contains 180 records and 65 attributes and among all them, CGPA and Arrear are considered and on the basis of these two, the data is clustered. The clusters which are made are eminently available to end users after all it is first partitioned and then clustered.

Ishwank Singh [12], proposed that it uses the K-means clustering algorithm and considered one of its limitation to choose the value of 'k', so the author uses the silhouette measured to choose the value of k. The value of k is chosen by highest average silhouette width for the dataset. The test is achieved using the Rapid Miner studio and it considers the attributes not only the marks, skills, projects, internships and backlogs but also Xth, XIIth and graduation marks.

IV. Examining Healthcare Using K-Means Clustering

Aigerim Altayeva [17], developed a Heart Disease Prediction System (HDPS) with the help of data mining techniques. The hybrid technique is made by combining the Naive Bayes and K-means clustering algorithm. To improve the teaching method of K-means, the Naive Bayes algorithm is combined with it. It uses attributes like sex, age blood pressure, blood sugar levels elctrodiogram and chest pain and determines the different study of patients. It also finds out the different methods for initial centroid selection for the diagnosis of heart patients and experiment proves that with this hybrid technique the accuracy in the diagnosis of patients is improved.

Ahmed Alsayat [18], proposed two healthcare datasets namely, liver dataset and heart dataset. The author introduces a new hybrid algorithm by combining Self Organizing Map (SOM), genetic algorithm and K-means clustering algorithm. These three are combined to remove the limitation of K-means clustering as SOM is used to find out the number of clusters to build and a genetic algorithm is used for initialization of centroids. The comparison is thus made between the proposed algorithm, the traditional k-means, and DBSCAN algorithm and evaluated that the proposed algorithm gives the better classification accuracy.

Ms. A. Malarvizhi [19], presented a comparison between three algorithms named K-means, K-means++ and Fuzzy C-Means (FCM) algorithm on two datasets as Breast tissue and diabetes. K-means++ is an enhanced version of K-means in which the problem of poor initialization of the centroid is removed while FCM considers that grouping is mad on the membership values. As per the experiments are carried out, it is examined that the K-means++ algorithm takes less time to complete and gives better accuracy results amongst three.

V. Conclusion

Data mining is the powerful tool nowadays to extract information from the large data. There are various approaches in data mining such as classification, clustering, association rule mining, etc. In this paper, we have discussed the advantages and limitations of K-means clustering, a type of clustering algorithm which is widely used. In various areas the K-means clustering is used, namely, improving the academic performance, optimization of search engine, in a healthcare, etc. Many authors proposed various improvements in traditional K-means clustering by eliminating their limitations according to their datasets and improves the performance of K-means clustering. In various domains, different variations are made to improve the accuracy of k-means clustering and many hybrid algorithms are also build for this purpose. The main point which is to be considered is how efficiently and effectively the grouping is made to exterminate the problems which arises in day to day life. There are various algorithms for clustering, but the best algortihm is chosen after determining its advantages and limitations. In future, consideration is also given to particular field where many variations of k-means is classified.

References

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 2012.

[2] C. C. Aggarwal, *Data Mining: The Text Book*, 1st ed., New York, Springer, 2015.

[3] M. Zaki, and W. Meira, *Data Mining and Analysis*, New York, Cambridge University Press, 2014.

[4] S. Shukla, and S. Naganna, "A review on k-means data clustering approach," *International Journal of Information & Computer Technology*, vol. 4, no. 17, pp. 1847-1860, 2014.

[5] J. Qiao, and Y. Zhang, "Study on k-means method based on data-mining," in *2015 Chinese Automation Congress (CAC)*, pp. 51-54, 2015.

[6] A. Yadav, and S. Dhingra, "A review on k-means clustering technique," *International Journal of Latest Research in Science and Technology*, vol. 5, no. 4, pp. 13-16, 2016.

[7] K. O. Khorsheed, M. M. Madbouly, and S. K. Guirguis, "Search engine optimization using data mining approach," *International Journal of Computer Engineering and Applications*, vol. 9, no. 6, part 1, pp. 184-200, 2015.

[8] M. E. Ahmed, and P. Bansal, "Clustering technique on search engine dataset using data mining tool," *2013 Third International Conference on Advanced Computing and Communication Technologies (ACCT)*, pp. 86-89, 2013.

[9] H. I. Arumawadu, R. M. K. T. Rathnayaka, and S. K. Illangarathne, "K-means clustering for segment web search results," *International Journal of Engineering Works*, vol. 2, no. 8, pp. 79-83, 2015.

[10] O. J. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa, "Application of k means clustering algorithm for prediction of students academic performance," *International Journal of Computer Science and Information Security*, vol. 7, no. 1, pp. 292-295, 2010.

[11] A. M. de Morais, J. M. F. R. Araujo, and E. B. Costa, "Monitoring student performance using data clustering and predictive modelling," *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*, pp. 1-8, 2014.

[12] I. Singh, A. S. Sabitha, and A. Bansal, "Student performance analysis using clustering algorithm," *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence),* pp. 294-299, 2016.

[13] Md. H. I. Shovon, and M. Haque, "Prediction of student academic performance by an application of data mining techniques," *International Journal of Advanced Research in Computer Scienece and Software Engineering*, vol. 2, no. 7, pp. 353-355, July 2012.

[14] R. A. Haraty, M. Dimishkieh, and M. Masud, "An enhanced k-means clustering algorithm for pattern discovery in healthcare data," *International Journal of Distributed Sensor Networks*, vol. 5, 2015.

[15] V. Rogeith, and S. Magesh, "A survey on health care data using data mining techniques," *International Journal of Pure and Applied Mathematics*, vol. 117, no. 16, pp. 665-672, 2017.

[16] T. Meyyappan, and S. Ganga, "Performance of students evaluation in education sector using clustering k-means algorithms," *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 7, pp. 579-584, 2014.

[17] A. Altayeva, S. Zharas, and Y. I. Cho, "Medical decision making diagnosis system integrating k-means and Naïve Bayes algorithms," *2016 16th International Conference on Control, Automation and Systems (ICCAS),* pp. 1087-1092, 2016.

[18] A. Alsayat, and H. El-Sayed, "Efficient genetic k-means clustering for health care knowledge discovery," *2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA),* IEEE, 2016.

[19] A. Malarvizhi, and S. Ravichandran, "Data mining's role in mining medical datasets for disease assessments - A case study," *International Journal of Pure and Applied Mathematics*, vol. 119, no. 12, pp. 16255-16260, 2018.