Analysis of Data using K-Means Clustering Algorithm with Min Max Function

S. Narain Sinha[1], Ram Lal Yadav[2]

[1]M.Tech. Student, Department of CSE, Kautilya Institute of Technology and Engineering,
Jaipur, Rajasthan, INDIA
[2]Associate Professor, Department of CSE, Kautilya Institute of Technology and Engineering,
Jaipur, Rajasthan, INDIA

Abstract — *The information is currently used for wide range of applications. Data mining is a logical process that is used to search through large amount of data in order to find useful data. Data mining is studied for different databases. For the proper utilization of data the data analytics techniques are applied on the data. Data analytics uses clustering, normalization, etc. Clustering is the process of organizing the objects into groups whose members are similar in some way to others. Lot of work is done in this field by different researchers. In this work the new data analytics technique is proposed. The base technique is modified by the new proposed technique. New technique uses the min max function instead of the scaling. The new technique is proposed, designed, implemented in the R language. The results obtained and analysed. The new proposed technique gives the better and compact clusters.*

Keywords — *Data Mining, K-Means Clustering, Data analytics, normalization, Min Max Function.*

I. Introduction

The demand for organizing the data and learning valuable information from data in increasing day by day, which makes the clustering techniques popular. These are widely applied in many application areas [1]. The rate of data creation at present has increased. Approximate 90% of the data in the world present today has been created in last two years alone [6]. This huge amount of data is viewed by business organizations and researchers as a great resource of knowledge that needs to be discovered. This knowledge can be accessed from the processing of data [2]. The research in databases and information technology has been increased. There is an approach to store and manipulate this valuable data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbour method etc. are used for knowledge discovery from databases.

Every day people add a large amount of information in the sea of data and store this data for further analysis and management. One of the vital means in dealing with these data is to classify or group in such a way that it takes the form of clusters [1]. Representing the data by fewer clusters loses certain details, but it achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters [2]. A cluster is an ordered list of objects, which have some common objects [3]. Basically, classification systems may be supervised or unsupervised [1].

There are many types of clustering. In general, the major clustering methods can be classified into Partitioning Method, Hierarchical Method, Grid Based Method, Model Based Method, Density Based Method, etc [4]. Large number of researchers developed different clustering techniques [5], [6], [8], [9].

Data analytics is science to examining the raw data with the purpose to extract useful information to find the conclusions [7]. Data Analytics is used by many industries and organizations to get the better business decision. Data analytics focuses on inference, the process of deriving a conclusion based solely on what is already known by the researcher. Classification models predict the categorical class labels; and prediction model predict continuous valued functions.

The paper is organised in to five sections. In the first section the basic introduction of the data mining, clustering, data analytics, etc. are given. In second section, the previous work has been discussed. In third section, the proposed technique discussed. In fourth section results of the work are given. In fifth section the conclusion of the work is given along with the future work.

II. Literature Survey

In the data analytics various techniques are proposed to analyse the data. Large amount of work has been performed by different researchers. The techniques were proposed by different authors.

Arpit Bansal et al. in [7] proposed a technique for a modification in K-Means Clustering Algorithm. By this proposed modification, the K-Means clustering will removed the two major drawbacks of

K-Means clustering: - the accuracy level and the calculation time consumed in clustering the dataset. Akhilesh Kumar Yadav et al. in [10] explained that huge amount of data is available in the health field. To extract valuable information from large data sets analytic tools can be used. In this work a real data set has been taken from SGPGI. Real time data sets are interlinked with some kind of challenges like missing values, high dimensional values, noise etc. which is not efficient for classification. Sanjay Chakraborty et al. in [11] discussed that clustering is one of the powerful tool which used in various forecasting tools. The generic methodology of incremental K-mean clustering is proposed in this work for weather forecasting. Chew Li Sa et al. in [12] proposed a system Student Performance Analysis System (SPAS) that keep track of student's result in a particular university. The proposed work offers a system that predicts the performance of students of the university on the basis of their result. Abdelghani Bellaachia et al. in [13] presented an analysis for the prediction of survivability rate of breast cancer patients. Qasem a. Al-Radaideh et al. in [14] explained that the forecasting stock return is an important subject that is to be learning for the prediction for data analysis. Jian Di et al. in [15] proposed an improved bisecting K-means algorithm which is based on the automatically determining the K value and the optimization of the cluster centre. Many authors performed lot of work to improve the clustering and data analytics.

III. PROPOSED TECHNIQUE

In [7], the authors used K- Means in cluster for similar type of data for prediction analysis. To improve the limitations, there is requirement to improve the base technique.

The data analytics technique is modified to improve the clustering and data analysis. The new technique uses the Min Max function instead of scaling in normalization. The steps of proposed technique are as following –

1. Start
2. Taken data N number of Data set.
3. After this apply the K-means clustering algorithm on the datasets and record the results.
4. After this, apply the Modified Normalization technique i.e. Min Max Function.
5. Now apply the hierarchical clustering algorithm. Record the results.
6. Calculate the timing of execution and accuracy level on the basis of different factors.
7. End.

The proposed work improves the accuracy and execution time in comparison to the previous data analytic technique. The different comparison factors are used.

IV. RESULT ANALYSIS

The base data analytics technique and the proposed Min Max Function data analytics technique are implemented in the R Studio. The base technique and proposed technique is applied on the random samples of 50, 100, 150, 200, 250, 500, 1000 and 2000 datasets. The results are as following on the basis of different comparison factors –

A. Based on Corrected Rand Index

The comparison results for base technique and proposed techniques are shown in fig. 1. The proposed technique is less dependent on external factors for validation than the base technique after applying the min max function. As this index shows the dependency on external factors for validation, so the proposed technique clearly shows that it is less dependent on external factors than the base technique.

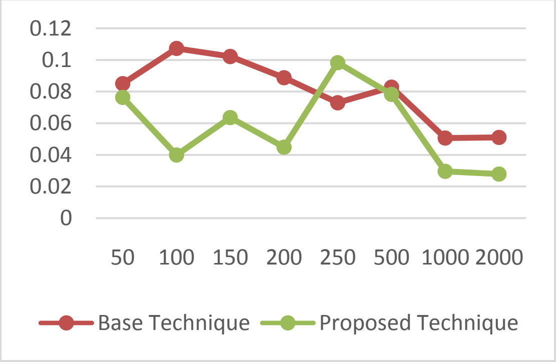

Fig. 1 Comparison Graph between Base and Proposed Technique on Corrected Rand Index

B. Variation of Information

The comparison results for base technique and proposed techniques are shown in fig. 2. The proposed technique has almost equal VI value in comparison to the base techniques. In both cases, there is less variation in the Variation of Information value. So there is no major difference between the two techniques.

C. Dunn Index

The comparison results for base technique and proposed techniques are shown in fig. 3. The proposed technique has better Dunn index value in comparison to the base technique. The proposed technique gives the compact and well separated clusters while the base technique gives the less Dunn Index clusters.

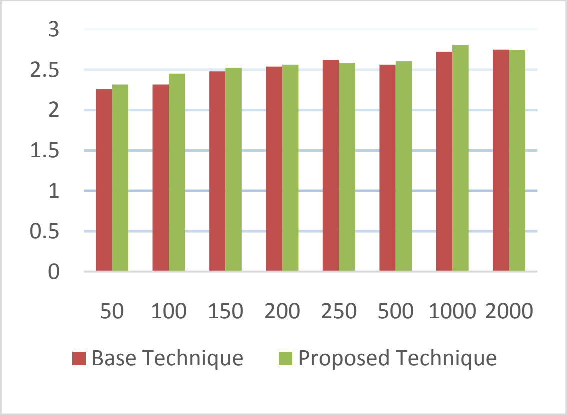

Fig. 2 Comparison Graph between Base and Proposed Technique on Variation of Information

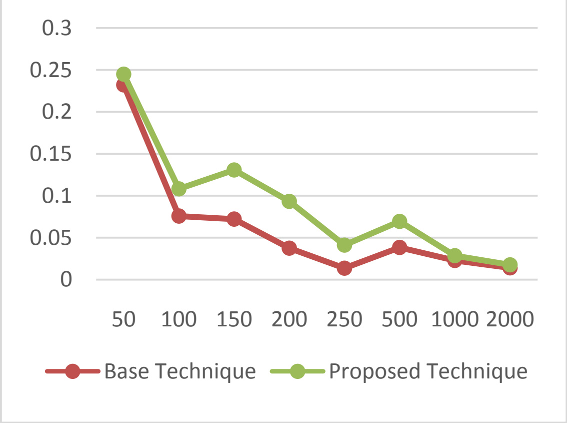

Fig. 3 Comparison Graph between Base and Proposed Technique on Dunn Index

D. Time Taken

The comparison results for base technique and proposed techniques are shown in fig. 4. The proposed technique has better Dunn index value in comparison to the base technique. This affects value of time taken in the proposed technique a little bit in fig. 4. The clusters are more compact and precise in proposed technique, which require time. So time taken factor value is higher in proposed technique than base technique.

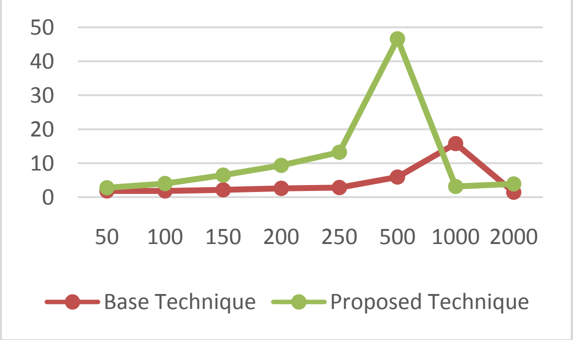

Fig. 4 Comparison Graph between Base and Proposed Technique on Time Taken

V. Conclusions and Future Work

Mining data is discovery of new knowledge in databases. Clustering is technique which is used to analyse the data in efficient manner and generate required information. In the proposed research work the data analytics technique is modified. The Min Max Function is used in the technique instead of the scaling in normalization. This has improved the technique. In future the number of data set and comparison factors can be increased. The various other clustering algorithms can be used for clustering.

References

[1] Rui Xu, Donald Wunsch, “Survey of Clustering Algorithms”, *IEEE Transactions on Neural Networks*, VOL. 16, NO. 3, MAY 2005.

[2] Osama Mahmoud Abu Abbas, “Comparisons between Data Clustering Algorithms”, *IAJIT*, Vol. 5, No. 3, 2008, p.p: 320-325.

[3] Nimrat Kaur Sidhu, Rajneet Kaur, “Clustering In Data Mining”, *International Journal of Computer Trends and Technology (IJCTT)*, Volume 4, Issue 4, April 2013

[4] Dhara Patel, Ruchi Modi, Ketan Sarvakar, “A Comparative Study of Clustering Data Mining: Techniques and Research Challenges”, *IJLTEMAS*, Volume III, Issue IX, September, 2014.

[5] Mythili S, Madhiya E, “An Analysis on Clustering Algorithms in Data Mining”, *International Journal of Computer Science and Mobile Computing, IJCSMC*, Vol. 3, Issue. 1, January 2014, pg.334 – 340.

[6] Mugdha Jain, Chakradhar Verma, “ Adapting k-means for Clustering in Big Data”, *International Journal of Computer Applications*, (0975 – 8887), Volume 101, No.1, September 2014

[7] Arpit Bansal, Shalini Goel, “Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining”, *International Journal of Computer Applications*, Volume 157, No 6, January 2017

[8] Ohidujjaman, Md. Mizanur Rahman, Ms. Raihana Zannat, “Clustering Algorithm with Asynchronous Programming”, *American Journal of Engineering Research (AJER)*, Volume-6, Issue-8, 2017, pp-286-294.

[9] Abhilash C B, Sharana Basavanagowda, “A Comparative study on clustering of data using Improved K-means Algorithms”, *International Journal of Computer Trends and Technology (IJCTT)*, Volume 4, Issue 4, April 2013

[10] 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

[11] Sanjay Chakraborty, Prof. N.K Nigwani and Lop Dey “Weather Forecasting using Incremental K-means Clustering”, 2014

[12] 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, vol., no., pp.1-6, 17-18 Nov. 2014

[13] Abdelghani Bellaachia, Erhan Guven, “Predicting Breast Cancer Survivability Using Data Mining Techniques”, Washington DC, 20052, 2010

[14] Qasem a. Al-Radaideh, Adel Abu Assaf eman Alnagi, “ Predictiong Stock Prices Using Data Mining Techniques”, *The International Arab Conference on Information Technology* (ACIT’2013), 2013.

[15] Jian Di, Xinyue Gou, “Bisecting K-means Algorithm Based on K-valued Self determining and Clustering Center Optimization”, *Journal of Computers*, Volume 13, Number 6,June2018