Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm

Vaishali R. Patel[1] and Rupa G. Mehta[2]

**[1] Department of Computer Engineering, Shri S'ad Vidhya Mandal Institute of Technology
Bharuch-392 001, Gujarat, India**

**[2] Department of Computer Engineering, Sardar Vallabhbhai National Institute of Technology
Surat-395 007, Gujarat, India**

Abstract

Clustering technique is mainly focus on pattern recognition for further organizational design analysis which finds groups of data objects such that objects in a group are similar to one another and dissimilar from the objects in the other group. It is important to preprocess data due to noisy data, errors, inconsistencies, outliers and lack of variable values. Different data preprocessing techniques like cleaning method, outlier detection, data integration and transformation can be carried out before clustering process to achieve successful analysis. Normalization is an important preprocessing step in Data Mining to standardize the values of all variables from dynamic range into specific range. Outliers can significantly affect data mining performance, so outlier detection and removal is an important task in wide variety of data mining applications. k-Means is one of the most well known clustering algorithms yet it suffers major shortcomings like initialize number of clusters and seed values preliminary and converges to local minima. This paper analyzed the performance of modified k-Means clustering algorithm with data preprocessing technique includes cleaning method, normalization approach and outlier detection with automatic initialization of seed values on datasets from UCI dataset repository.

Keywords: *Clustering, k-Means, Normalization Approach, Outlier Removal, Preprocessing.*

1. Introduction

Data mining techniques automate the process to extract hidden patterns from the heterogeneous data sources and to analysis the results which is helpful to the organization for decision making with the development of number of technologies. Data mining is one of the most important research fields that are due to the expansion of both computer hardware and software technologies, which has imposed organizations to depend heavily on these technologies [1]. According to [2], the obtained clusters should react some mechanism at work in the domain from which instances or data points are drawn, a mechanism that causes some instances to bear a stronger resemblance to one another than they do to the remaining instances." Clustering is one solution to the case of unsupervised learning, where class labeling information of the data is not available. It is a method where data is divided into groups (clusters) which 'seem' to make sense. Clustering algorithms are usually fast and quite simple. They need no beforehand knowledge of the used data and form a solution by comparing the given samples to each other and to the *clustering criterion* [3]. Various clustering algorithms according to different techniques have been designed and applied to various data mining problems successfully. Accomplishment of clustering algorithms in many areas, it causes many precincts to the researchers when no or little information are available. There is also no universal clustering algorithm developed; that's why it is very crucial job to choose appropriate clustering technique and algorithm considering above precincts. A good survey on clustering techniques and algorithms found in [4]. A simple and commonly used algorithm for producing clusters by optimizing a criterion function, defined either globally (over all patterns) or locally (on a subset of the patterns), is the k-means algorithm [5]. The k-Means algorithm [5] is effective in producing clusters for many practical applications. This algorithm results in different types of clusters depending on the random choice of initial centroids. Several attempts were made by researchers to improve the performance of the k-Means clustering algorithm. k-Means use Euclidean distance measure centroids of the clusters and distortion among the data objects. These distances are not computed from standardized data. In method of Euclidean distance, the measured distance between data objects is not affected by addition of new objects to the analysis [6]. Missing value is a common problem in almost every real world data. The presence of missing values in data results in datasets that [7] refers to as "incomplete" datasets since some

information will not be available. Data pre-processing techniques are applied on raw data to make the data clean, noise free and consistent. Data Normalization standardize the raw data by converting them into specific range using linear transformation which can generate good quality clusters and improve the accuracy of clustering algorithms. The outlier detection is searching for objects in the database that do not obey laws valid for the major part of the data [8]. In clustering, outliers are considered as observations that should be removed in order to make clustering more reliable. Different approaches have been proposed to detect outliers and some of these are discussed in literature survey explained in section 2. Section 3 explains the traditional k-Means with pros and cons. Section 4 proposes modify k-Means clustering algorithm which detect outlier using 5-95% method, apply different normalization techniques like Min-Max, Z-Score and Decimal Scaling to improve the performance and accuracy of the k-Means algorithm. The proposed method first checks to ensure that the data apply to the algorithm are clean and standardized then apply 5-95% method which discard the data and consider it as outlier of the given dataset. Section 5 discusses the implementation of modify k-means and result analysis is done on dataset from the UCI dataset repository which shows that outlier detection and removal with normalization approach improve the effectiveness and performance of the modified k-Means clustering algorithm.

2. Literature Survey

Clustering algorithms generate clusters having similarity between data objects based on characteristics belongs to same cluster. Clustering is extensively used in many areas such as pattern recognition, computer science, medical, machine learning. Result of clustering is dependent on the type of data and application area. An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism [9]. In outlier detection methods based on clustering, outlier is defined to be an observation that does not fit to the overall clustering pattern [10]. This section discusses the various approaches proposed by many researchers to detect outliers in k-Means and other clustering algorithms to handle noise and generate successful results.

Authors [11], proposed a new clustering based approach, which divides the stream in chunks and clusters each chunk using k-median into variable number of clusters. Instead of storing complete data stream chunk in memory, they replace it with the weighted medians found after mining a data stream chunk and pass that information along with the newly arrived data chunk to the next phase. The weighted medians found in each phase are tested for outlierness and after a given number of phases, it is either declared as a real outlier or an inlier. This technique is theoretically better than the k-means as it does not fix the number of clusters to k rather gives a range to it and provides a more stable and better solution which runs in poly-logarithmic space. This approach works only for numeric dataset.

Ville Hautam aki et al. [12] proposed an Outlier Removal Clustering (ORC) algorithm which detects outlier and data clustering simultaneously. The method employs both clustering and outlier discovery to improve estimation of the centroids of the generative distribution. During the first stage of this algorithm, basic k-Means algorithm executes, while during the second stage it iteratively removes the vectors which are far from their cluster centroids. This approach outperforms, particularly in the case of heavily overlapping clusters. In this approach, setting of correct parameter depends on the type of dataset.

Authors [13] have proposed shortest distance method for detecting outliers in k-Means and k-Medians clustering algorithm. In this algorithm, outliers are detected by computing its distance which is far away from the rest of the data objects in the dataset.

Moh'd Belal and Al- Zoubi [14] have proposed an algorithm based on clustering approaches to detect outliers. This algorithm first performs the PAM clustering algorithm. Small clusters are then determined and considered as outlier clusters. The rest of outliers are then detected in the remaining clusters based on calculating the absolute distances between the medoid of the current cluster and each one of the points in the same cluster. This algorithm can be easily implemented on other clustering algorithms that are based on PAM.

During the first phase of proposed two phase clustering algorithm [15], traditional k-Means algorithm is modified using a heuristic "if one new pattern is far enough away from all clusters' centers, then assign it as a new cluster center". As a result, data objects in the same cluster may be most likely all outliers or all non-outliers. During second phase of this proposed algorithm, Minimum Spanning Tree (MST) is constructed and removes the longest edge and small clusters are considered as outliers.

3. Naive k-Means Algorithm

This section discusses the working of traditional k-Means clustering algorithm. k-Means algorithm is one of the most popular clustering algorithms due to its efficiency and simplicity in clustering large data sets. In traditional k-

Means algorithm, a set of data set D is classified using a certain number of clusters (k clusters) which are initialized apriori. It define k centroids, one for each cluster and then consider data object belonging to the given data set and associate this data objects to the closest centroid. Euclidean distance generally considered to determine the distance between data objects and the centroids [16]. First step is completed when there is no data object is remaining and early group is done. Here, there is need to re-calculate new centroids. After obtaining new centroids same data objects are binded with the closest centroid and generate a loop. At the end of loop, k-centroids change their point step by step until centroids do not move any more. This algorithm works on basis of minimizing squared error function. k-Means algorithm suffers from the problems of giving number of clusters and initial seed values preliminary. The k-Means algorithm always converges to a local minimum and it depends on the initial cluster seed values. We have make modifications in traditional k-Means algorithm to initialize the seed values with data preprocessing and data normalization techniques like Min-Max, Z-score and decimal scaling to improve the accuracy and efficiency of traditional algorithm. Section 4 discusses the modifications in k-Means clustering algorithm with normalization approach.

4. Modified k-Means Algorithm with Outlier Detection and Removal

k-Means algorithm can generate better result after modification on the datasets. We apply the modified k-Means algorithm with automatic initialization of number of seed values on river dataset for number of iterations and clusters and compute MSE. Next, we preprocess and normalize river dataset before apply on modify k-Means algorithm. This proposes method works in two stages. During the first stage, we preprocess the dataset then compute and discard 5-95% data from the dataset. Store and normalize these discarded data separately which we consider outliers and use it as new cluster with modified k-Means algorithm. During the second stage, apply modify k-Means algorithm to remaining data to generate clusters. Block diagram and flow chart of the proposed modified k-Means algorithm is shown in Figure 1 and Figure 2 respectively:

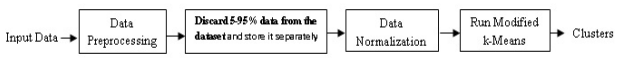

Fig. 1. Block Diagram of modified k-Means clustering algorithm with Outlier Detection

Platform used: VB 6.0 and MS SQL Server 5.0

Input: RIVER Datasets from UCI Machine Leaning

Transform Module: This module accept the dataset in text format and convert it in database file.

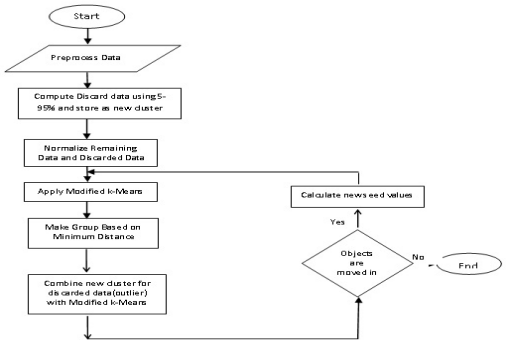

Fig. 2: Flow Chart of Modified k-Means with Outlier Detection and Removal

Data Preparation Module: This module works in two parts. First, data preprocessing technique apply on the dataset received by transform module. The clean data is then passed to second part; Data Normalization which transform the clean raw data into specific range using different techniques.

Data Pre-processing: This is a very important step since it can affect the result of a clustering algorithm. This module calculates tuples with missing values using different options like maximum, minimum, constant, average and standard deviation for the treatment of missing values tuples before we apply normalization approach on the dataset. This process gives the treatment of missing value data and then it applies to the second part (data normalization) of data preparation.

Normalization Approach: Data Mining can generate effective result if proper and effective data mining technique can apply to the dataset. According to authors [17], normalization is used to standardize all the features of the dataset into a specified predefined criterion so that redundant or noisy objects can be eliminated and use made of valid and reliable data which can effect and improve accuracy of the result. k-Means clustering algorithm uses Euclidean distance measures which are highly susceptible to inconsistencies in the size of the features. k-Means algorithm which is using Euclidean distance measure, data normalization is an essential step to prevent larger features from randomized value to the specific range. The importance of normalization is that it enhances the accuracy of the results that are obtained during clustering [18]. Better results are generated when data preparation with data preprocessing and normalization is carried out with different techniques. Data normalizations techniques include min-max normalization, Z-Score normalization,

and Decimal Scaling normalization. There is no universally defined rule for normalizing datasets and thus the choice of a particular normalization rule is largely left to the discretion of the user [17]. The proposed Mk-Means algorithm has utilized the three normalization techniques and compares analysis of achieved results. The Min-Max normalization technique involves the linear transformation on raw data. MinA and MaxA are minimum and maximum value for the attribute A. This technique maps the value of attribute A into range of [0, 1]. Equation (1) shows the computation of Min-Max normalization technique.

$$v' = \frac{v - MinA}{MaxA - MinA} \qquad (1)$$

Z-Score normalization technique is useful when the actual minimum and maximum value of attribute A are unknown the value of an attribute using standard deviation and mean of the attribute A. Equation (2) shows the computation of Z-Score normalization technique. In equation (2), $\bar{A}$, σA and v are mean, standard deviation and value of attribute A.

$$v' = \frac{v - \bar{A}}{\sigma A} \qquad (2)$$

Decimal scaling normalization technique normalize the data by moving the decimal point of values of attribute A. Number of decimal points moved depends on the maximum absolute value of A. Equation (3) shows the computation of decimal scaling normalization technique. In equation (3), v is the value of attribute A and j is the smallest integer where Max (|v'|) <1.

Outlier Detection: Outliers detection is a task that finds objects that are dissimilar or inconsistent with respect to the remaining data. It has many uses in applications like fraud detection, network intrusion detection and clinical diagnosis of diseases. Clustering algorithms are frequently used for outlier detection. The clustering algorithms consider outlier detection only to the point they do not interfere with the clustering process. In this proposed approach, outliers are detected using 5-95% method in which 5% of data from minimum side and 5% data from maximum side are detected and removed from the dataset. The experimental results prove that Modified k-Means clustering algorithm with outlier detection and removal improves the accuracy and increases the time efficiency of Mk-Means algorithm.

$$v' = \frac{v}{10^j} \qquad (3)$$

Run modified k-Means Algorithm: The modified k-Means algorithm is implemented and applies it on River dataset which contains 133 tuples and 9 attributes. We have applied this algorithm using two dimensional on RIVER dataset using 4 iterations and different number of clusters. We have applied the same algorithm with data preparation (cleaning method) and data normalization techniques and compute the MSE for the same.

5. Result Analysis

To analyze the accuracy of proposed Mk-Means clustering algorithm, result is taken on River dataset from UCI machine learning data repository. Comparison of computed MSE of Mk-Means with computed MSE of Mk-Means algorithm with cleaning method and various normalization techniques like Min-Max, Z-Score and Decimal Scaling is done on River dataset. This Analysis shows the best result for Mk-Means with normalization approach. Computed MSE and graph of proposed Mk-Means algorithm with outlier removal and cleaning method is shown in Table 1 and Figure 3 respectively. Computed MSE and graph of Mk-Means with different normalization techniques like Min-max, Z-score and Decimal scaling is shown in Table 2 and Figure 4 respectively. Computed MSE and graph of Mk-Means with different normalization techniques like Min-max, Z-score and Decimal scaling and outlier removal is shown in Table 3 and Figure 5. If there are N tuples in the dataset, then, the similarity matrix can be computed in O(KNT). Let N is the number of tuples in the dataset. K is the number of clusters or centroids and T is the time to calculate the distance between to data objects. Time complexity of each iteration is O(KNT). I is the number of iterations in k-Means algorithm. So, during I number of iteration the time complexity of this algorithm is O(IKNT). The performance analysis of modified k-Means clustering algorithm shows that decimal scaling normalization technique gives the best results for the modified k-means clustering algorithm and secondly min-max data normalization generates the best results for modified k-means clustering algorithm. The analysis shows that outlier detection and removal with generates the best and most effective and accurate results than other techniques used in this paper. Comparison of MSE of proposed Mk-Means algorithm with MSE of Data Normalization techniques with Mk-Means algorithm. MSE of proposed modified k-means clustering algorithm with min-max, z-score and decimal scaling. Analysis

shows that by applying data preparation techniques like cleaning method for missing value treatment, different normalization approaches and outlier detection and removal improve the performance of modified k-Means clustering algorithm.

Table 1: MSE of proposed Mk-Means Clustering Algorithm with Outlier Removal

DataSet	No. of Samples	Data Preprocessing	Number of clusers	MSE of proposed Mk-Means Algorithm	MSE of proposed Mk-Means Algorithm with Outlier Removal
RIVER	133	Average	2	4.967	4.332
			3	114.597	103.703
			4	55.284	32.095
			5	30.124	18.414
			6	32.531	28.879

Table 2: MSE of proposed modified k-means clustering algorithm with min-max, z-score and decimal scaling normalization techniques

DataSet	No. of Samples	Data Preprocessing	Number of clusers	MSE of proposed Mk-Means with Normalization techniques		
				Min Max	Z-Score	Decimal Scaling
RIVER	133	Average	2	0.769	0	0.88
			3	25.701	83.701	4.377
			4	3.567	64.744	0.761
			5	4.619	22.833	1.19
			6	4.0	18.384	1.308

Table 3: MSE of Proposed Mk-Means algorithm with Outlier Removal and Normalization Techniques

DataSet	No. of Samples	Data Preprocessing	Number of clusers	MSE of Mk-Means with Outlier Removal and Normalization Techniques		
				Min Max	Z-Score	Decimal Scaling
RIVER	133	Average	2	0.76	0	0.839
			3	24.566	62.225	4.197
			4	3.203	43.957	0.655
			5	3.554	15.475	0.979
			6	5.111	10.616	1.096

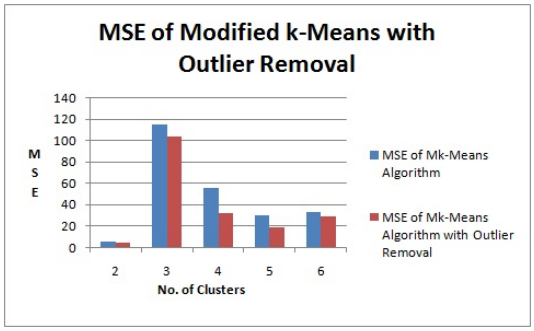

Fig 3: MSE of Mk-Means with Outlier Removal

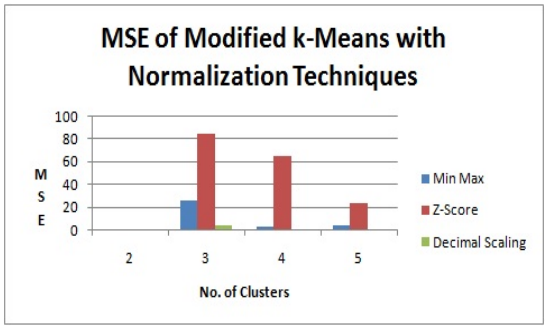

Fig 4: MSE of Mk-Means with Normalization Techniques

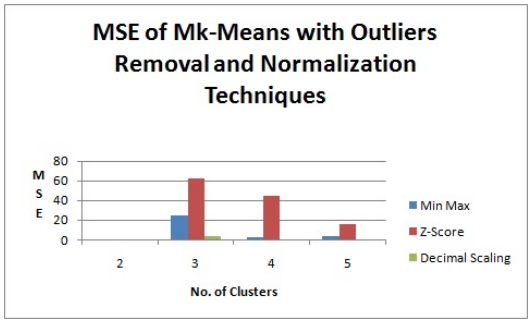

Fig 5: MSE of Mk-Means with Outlier Removal and Normalization Techniques

6. Conclusion

Data Mining is the step in KDD to extract useful pattern. Clustering organize the data in group having similarity. k-Means clustering is the well known partition based clustering algorithm. K-Means suffers from one of the problem of initializing seed values. This paper give explore of data mining and literature survey of methods proposed by many researchers to remove initialization seed values in k-Means. Real world data may be noisy, with missing values or inconsistent. There are a number of data preprocessing techniques to clean the data. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data transformations, such as normalization, may be applied to improve the accuracy and efficiency of mining algorithms. These data processing techniques, when applied prior to mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining. This paper also propose modifications in naive k-Means to auto initialize seed values with data preparation which preprocess the data with cleaning method and transform the data into specific range using min-max, z-score and decimal scaling data normalization techniques. Outlier is a noise in the clustering algorithm which is detected using 5-95% method and removes from the dataset. Performance analysis of computed MSE for Mk-Means and Mk-Means

with three normalization techniques with outlier removal shows best and effective result for Mk-Means which generate minimum MSE and improve the efficiency and quality of result generated by this algorithm.

Acknowledgments

The authors wish to thank Computer Engineering department of Sardar Vallabhbhai National Institute of Technology for the support and providing an environment for this research work.

References

[1] R. Agrawal, T. Imielinksi and A. Swami, “Mining association rules between sets of items in large database, “The ACM SIGMOD Conference, Washington DC, USA, 1993, pp. 207-216.

[2] Babu S. and Widom J. (2001), “Continuous Queries over Data Streams”, Stanford University, SIGMOD Record, 30, pp. 109–120.

[3] Jukka Kainulainen, ”Clustering Algorithms: Basics and Visualization”, HELSINKI UNIVERSITY OF TECHNOLOGY, Laboratory of Computer and Information Science,T-61.195 Special Assignment 1, 2002.

[4] Vaishali R. Patel, Rupa G. Mehta, “Clustering Algorithms: A Comprehensive Survey”, International Conference on Electronics, Information and Communication Systems Engineering, MBM Engineering College, JNV University, Jodhpur, 2011.

[5] McQueen J, “Some methods for classification and analysis of ultivariate observations,” Proc. 5th Berkeley Symp. Math. Statist. Prob., Vol. 1, 1967, pp. 281–297.

[6] Grzymala-Busse and J. W., “Three Approaches to Missing Attribute Values – A Rough Set Perspective”, http://lightning.eecs.ku.edu/c97-brighton.pdf.

[7] N. Karthikeyani Visalakshi and K. Thangavel, ”Impact of Normalization in Distributed K-Means Clustering”, International Journal of Soft Computing 4, Vol. 4, 2009, pp. 168-172.

[8] Su, C.M., Tseng, S.S., Jiang, M.F., Chen and J.C.S., ”A Fast Clustering Process for Outliers and Remainder Clusters”, Lecture Notes in Artificial Intell, 1999, pp. 360-364.

[9] D.Hawkins: “Identification of outliers”. Chapman and Hall, London, 1980.

[10] Zhang T., Ramakrishnan, R. and Livny M., “BIRCH: A new data clustering algorithm and its applications”, Data Mining and Knowledge Discovery, 1997, pp. 141–182.

[11] Parneeta Dhaliwal, MPS Bhatia and Priti Bansal, ”A Cluster-based Approach for Outlier Detection in Dynamic Data Streams (KORM: k-median OutlieR Miner)”, Journal of Computing, Vol. 2, No. 2, 2010, ISSN 2151-9617, pp. 74-80.

[12] Ville Hautam aki, Svetlana Cherednichenko, Ismo Karkk ainen,Tomi Kinnunen, and Pasi Fr anti, ”Improving k-Means by Outlier Removal”, SCIA, LNCS 3540, 2005, pp. 978–987.

[13] Sairam, Manikandan and Sowndarya, ”Performance Analysis of Clustering Algorithms in Detecting Outliers”, International Journal of Computer Science and Information Technologies, Vol. 2, No. 1, 2011, ISSN: 0975-9646, pp. 486-488.

[14] Moh’d Belal Al- Zoubi, ”An Effective Clustering-Based Approach for Outlier Detection”, European Journal of Scientific Research, Vol. 28, No. 2, 2009, ISSN 1450-216X, pp.310-316.

[15] M.F. Jiang and S.S. Tseng, ”Two Phase Clustering Process for Outlier Detection”, Pattern Recognition Letters 22, Elsevier Science B.V., 2001, pp. 691-700.

[16] K. A. Abdul Nazeer and M. P. Sebastian, ”Improving the Accuracy and Efficiency of the k-means Clustering Algorithm”, Proceedings of the World Congress on Engineering, Vol. 1, 2009.

[17] N. Karthikeyani Visalakshi and K. Thangavel, ”Distributed Data Clustering: A Comparative Analysis. Foundations of Computational Intelligence (6)”, 2009, pp. 371-397.

[18] M arcilio CP de Souto, Ivan G Costa, Daniel SA de Araujo, Teresa B Ludermir and Alexander Schliep, ”Clustering cancer gene expression data: a comparative study”, BMC Bioinformatics”, 2008.

Vaishali R. Patel is M.Tech (Research) scholar and currently working as an Assistant Professor in Department of Computer Engineering and Information Technology at Shri S’ad Vidya Mandal Institute of Technology, Bharuch, Gujarat, India. Her research areas of interest include Data mining, Clustering, Database Management System` and Software Testing and Quality. She is Microsoft Certified Solution Developer.

Rupa G. Mehta is Ph. D. Scholar and currently working as an associate professor in Department of Computer Engineering at Sardar Vallabhbhai National Institute of Technology, Surat, Gujarat, India. Her research areas of interest include data mining, classification techniques, database management systems, data structures and formal language.