Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 4, April 2016, pg.453 – 464

OUTLIER DETECTION USING ENHANCED K-MEANS CLUSTERING ALGORITHM AND WEIGHT BASED CENTER APPROACH

J. James Manoharan[1], Dr. S. Hari Ganesh[2] Ph.D., Dr. J.G.R. Sathiaseelan[3]

[1]Associate Professor in Dept. of Computer Applications, Bishop Heber College (Autonomous), Tiruchirappalli, India
[2]Assistant Professor, Dept of Computer Science, H.H.Rajah's College, Pudukkottai, India
[3]Associate Professor, Department of Computer Science, Bishop Heber College, Trichy, TAMIL NADU, India
[1] james_7676@yahoo.com; [2] hariganesh17@gmail.com; [3] jgrsathiaseelan@gmail.com

ABSTRACT-In Data mining there are lots of methods are used to detect the outlier by making the clusters of data and then detect the outlier from them. In general Clustering method plays a very important role in data mining. Clustering means grouping the similar data objects together based on the characteristic they possess. Outlier Detection is an important issue in Data mining; particularly it has been used to identify and eliminate anomalous data objects from given data set where outlier is the data item whose value falls outside the bounds in the sample data may indicate anomalous data. In this work we have suggested a clustering based outlier detection algorithm for effective data mining which uses enhanced k-means clustering algorithm to cluster the data sets and weight based center approach. In proposed approach, two techniques are combined to efficiently find the outlier from the data set. Threshold value can be calculated programmatically by taking absolute value of minimum and maximum value of a particular cluster. The experimental results demonstrate that enhanced method takes least computational time and concentrates on reducing the outlier that could improve efficiency of k-means clustering for achieving the better quality clusters.

Index Terms— Data Mining, Outlier, Density Based Outlier detection, Distance Based Outlier detection, K-means Clustering.

I. 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 [5]. Cluster analysis is the task of assigning a set of data objects into groups called clusters so that the objects in the same cluster are more similar in some sense to each other than to those in other clusters [11].

A data items whose values are different from rest of data or whose values falls outside the described range are called outlier [12]. Detecting outliers has important applications in data cleaning as well as in the mining of abnormal points for fraud detection, intrusion detection, network sensors, stock market analysis and marketing. Finding anomalous points among the data points is the basic idea to find out an outlier. Outlier detection is a significant research problem that aims to discover data objects that are considerably different, exceptional and inconsistent in the database [10]. There are various origins of outliers. With the growth of the medical dataset day by day, the process of determining outliers becomes more complex and tedious. Efficient detection of outliers reduces the possibility of making poor decisions based on erroneous data, and aids in identifying, preventing, and repairing the effects of malicious or faulty behavior [7][8].

Furthermore, a lot of data mining and machine learning algorithms and techniques for statistical analysis may not work well in the presence of outliers. Outliers may introduce skew or complexity into models of the data, making it difficult, if not impossible, to fit an accurate model to the data in a computationally feasible manner. For example, statistical measures of the data may be skewed because of erroneous values, or the noise of the outliers may obscure the truly valuable information residing in the data set. Accurate and efficient removal of outliers may greatly enhance the performance of statistical and data mining algorithms and techniques [15]. Detecting and eliminating such outliers as a pre-processing step for other techniques is known as data cleaning [6]. As can be seen, different domains have different reasons for is covering outliers: They may be noise that we want to remove.

Outliers are ordinary elements when specified as input, but will direct in inefficient outputs when processed with them [7]. They are elements which behave very differently from the norm, and are the major disadvantage of k-means clustering algorithm.

II. RELATED WORK

Outlier detection can be mostly classified into three groups. First is the distance based outlier detection, it detects the outlier from the neighborhood points. Second is the density based outlier detection, here it detects the local outlier from the neighborhood based on the density or the number of data points in the local neighborhood [9]. Third is the distribution based outlier detection, this approach is based on the finding outlier using some statistical model.

Pallavi Purohit and Ritesh Joshi et. al [1] proposed an enhanced approach for traditional K-means clustering algorithm due to its certain limitations. The poor performance of traditional K-means clustering algorithm is selection of initial centroid points randomly. The proposed algorithm deals with this problem and improves the performance and cluster quality of traditional k-means algorithm. The enhanced algorithm selects the k initial centroids in a efficient manner rather than randomly selecting. It first discover the closest data objects by calculating Euclidian distance between each data objects and then these data points are deleted from population and forms a new data set. The enhanced algorithm provides more precise results and also reduces the mean square distance. But the proposed algorithm works better for dense dataset rather than sparse data set.

Wang Shunye et. al [2] proposed enhanced k-means clustering algorithm basically consists of three steps. The first step talk about the construction of the dissimilarity matrix .Secondly, Huffman algorithm is used to create a Huffman tree according to dissimilarity matrix. The output of Huffman tree gives the initial centroids. Finally the k-means clustering algorithm is be appropriate to initial centroids to get k cluster as output. Wine and Iris datasets are selected from UIC machine learning repository to test the enhanced algorithm. Proposed algorithm gives better accuracy rates and results than the traditional k-means clustering algorithm.

Fahim A.M,Salem et. al [3] suggested an efficient k-means clustering algorithm to prevail over problems in traditional k-means. Traditional k-means is well-known due to its easiness, straightforward and flexibility to sparse data. Even though traditional k-means clustering algorithm has its large number of advantages, it has certain disadvantages also. The final result depends on the initial centroids. The improved algorithm initially allot datasets to its nearby centroid and then calculate distance with other centroids. Further, compare the distance between the two data objects and then the new distance is smaller than the earlier distance then the data object is moved to new cluster or else if it is smaller than it is allocated to same cluster. This procedure will reduce the time and improve the efficiency of the traditional k-means clustering algorithm. The proposed method uses two functions .Initially the distance() function that is used to calculate the distance between data object and its nearest cluster head. Next, the distance_ new() function can be used to calculate distance between data objects and other remaining clusters. The experimental results demonstrate that the proposed k-means clustering algorithm is much faster and efficient than the traditional k-means clustering algorithm.

Juntao Wang and Xiaolog [4] in his study, an improved k-means clustering algorithm to deal with the problem of outlier detection of traditional k-means clustering algorithm. The enhanced algorithm makes use of noise data filter to deal with this problem. Outliers can be detected and removed by using Density based outlier detection method. The purpose of this method is that the outliers may not be engaged in computation of initial cluster centres. The Factors used to test are clustering time and clustering accuracy. The drawback of the enhanced k-means clustering algorithm is that while dealing with large scale data sets, it takes more time to produce the results.

Md.SohrabMahmud et. al [5] proposed an algorithm uses heuristic method to calculate initial k centroids . The proposed algorithm yields accurate clusters in lesser computational time. The proposed algorithm initially calculates the average score of each data objects that has multiple attributes and weight factor. Next, the Merge sort is applied to arrange the output that was generated in first phase. The data points are then divided into k cluster .Finally the nearest possible data point of the mean is taken as initial centroid. Although the proposed algorithm still deals with the problem of assigning number of desired k-cluster as input.

2.1 Density based outlier detection

Density-based technique has been developed for predicting outliers in a spatial data. These methods can be clustered into two categories called multi-dimensional metric space-based methods and graph-based methods. In the first category, the definition of spatial neighborhood is based on Euclidean distance, while in graph-based spatial outlier detections the definition is based on graph connectivity.

2.2 Distance based outlier detection

In Distance-based methods outlier is defined as an data object that is at least d_{min} distance away from k percentage of objects in the dataset. The problem is then finding appropriate d_{min} and k such that outliers would be correctly detected with a small number of false detections. This process usually needs domain knowledge.

III. METHODOLOGY

In the proposed approach the outliers can be detected in *two phases.* In the First phase we used enhanced K-Means clustering algorithm to produce a set of k clusters. *In the second* phase for each cluster the absolute distance between each object and the cluster center is calculated using weight based center approach.

3.1. System architecture

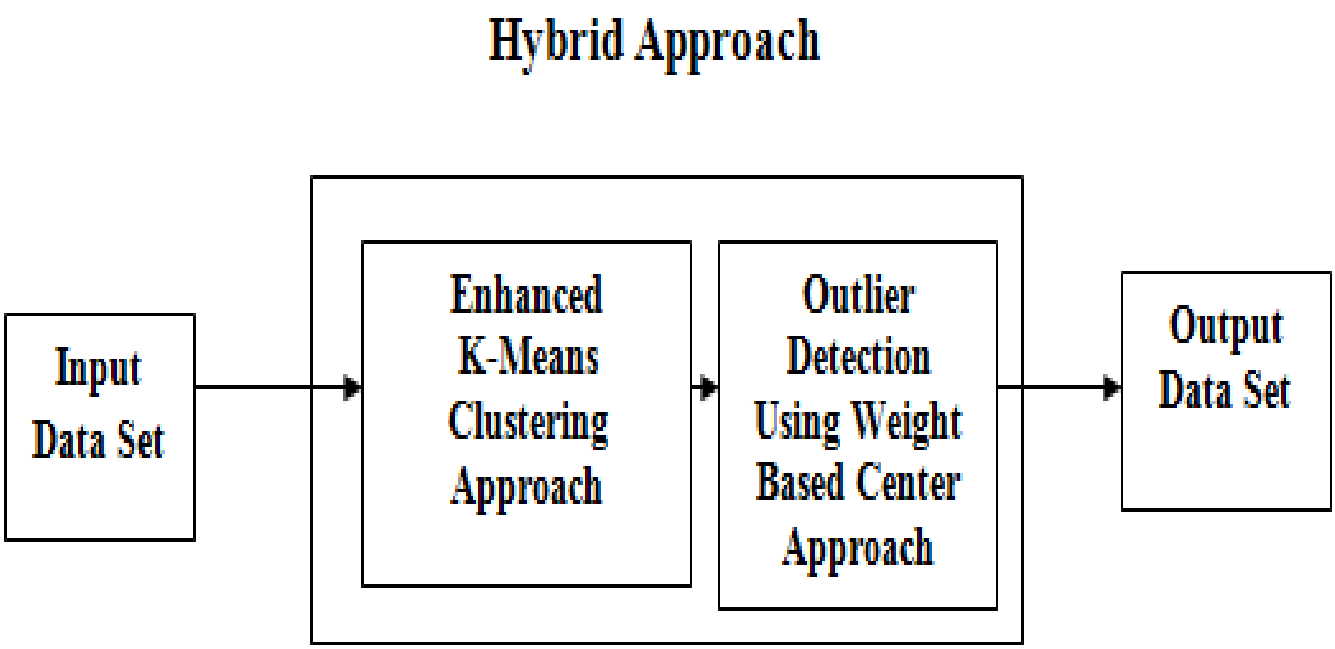


Fig. 1 System Architecture

3.1.1 Input Data Set:

The input dataset can be collected from UCI Machine learning repository.

3.1.2 Enhanced K-Means Cluster Based Approach:

Clustering is a popular technique used to group similar data points or objects in groups or clusters [6]. Clustering is an important tool for outlier analysis. The proposed framework for k-means algorithm which eliminates the problem of generation of

empty clusters and increases the efficiency of traditional k-means algorithm [17]. The framework is composed of 3 phases; choosing initial k-centroids phase, calculate the distance phase and recalculating new cluster center phase. In short, in the choosing initial k-centroids phase the initial cluster centers have obtained using divide-and-conquer method [16]. In calculate the distance phase the distance between each data items and cluster centers in each iteration could be calculated using linear data structure List [15]. Finally, in the recalculating cluster center phase to modify the center vector updating procedure of the basic *k*-means that reduce the formation of empty clusters [17].

3.1.3 Outlier Detection Using Weight Based Center Approach:

Outlier detection is an extremely important task in a wide variety of application domains. Outlier detection is a task that finds objects that are dissimilar or inconsistent with respect to the remaining data or which are far away from their cluster centroids.

First calculate the weight of each cluster using equation (1) and then store the result into a Vector W.

$$W_{(k)} = \sum_{x=1}^{n} d_x \quad \text{---------- (1)}$$

where k is the cluster number(k=1,2,….,n) and d_x is the number of data items($d_1, d_2, \ldots, d_n$) in particular cluster K.

Next, find the mean of each cluster using equation (2) and then the result stored into a Vector M

$$M_k = (\sum_{x=1}^{n} d\,x)/Tn \quad \text{---------- (2)}$$

where k is the cluster number (k=1,2,….,n) and d_x is the number of data items in particular cluster k and T_n is the total number of data items in cluster K.

Calculate the Maximum and Minimum value of each cluster and then store the maximum value in to an Array $D_{k\text{-}max}$ and store the minimum value in $D_{k\text{-}min}$ Array.

$$D_{k\text{-}max} = \text{Max}\ (d_1, d_2, \ldots\ldots, d_n)$$
$$D_{k\text{-}min} = \text{Min}\ (d_1, d2, \ldots\ldots, d_n)$$

Where k=1,2…,n is the cluster number .

The threshold value for each cluster K(1,2,…,n) can be calculated and then the results stored into an array Th using following equation (3)

$$Th_{k\text{-}critical} = \text{ABS}\ (M_k - ((Dkmax + Dkmin)/2)) \quad \text{---------- (3)}$$

where k is the cluster number.

Compare each data item d_x(x=1, 2,….., n) in a particular cluster K (1,2,…..,n) with the threshold value $Th_{k\text{-}critical}$. If the value is found less than $Th_{k\text{-}critical}$ then the data item is the outlier.

IV. PROPOSED ALGORITHM

Algorithm: Outlier Detection using Enhanced K-Means Clustering Algorithm and Weight Based Center Approach

Input: Data set D= $\{d_1,d_2.....d_n\}$,where n is the number of data points.
Cluster centre C= $\{c_1,c_2,.....c_k)$, where c_i is the cluster centre and k is the number of cluster centres.

Output: A set of K-clusters without outliers.

Step 1: Select k observations from data set using Divide-and-Conquer method.

Step 2: Calculate distance with each data instances using List Data structure.

Step 3: Assign each instance to the cluster with the nearest seed.

Step 4: *R*ecalculate the cluster center using Vector data structure.

Step 5: Repeat the process step 2 to 4 until the convergence criteria is achieved.

Step 6: Showing the clustering results.

Step7: Calculate the weight based centre W_k as given in the equation (1).

Step 8: Calculate the Mean of each cluster centre, let it be M_k as given in the equation (2).

Step 9: Calculate the maximum and minimum value of each cluster K, where $D_{k\text{-}max}$ = Max ($d_1,\ldots, d_n$) and $D_{k\text{-}min}$= Min($d_1,d_2,\ldots d_n$).

Step 10: Calculate $Th_{k\text{-}critical}$=ABS (M_k – (($D_{k\text{-}max}$ + $D_{k\text{-}min}$)/2) using equation (3).

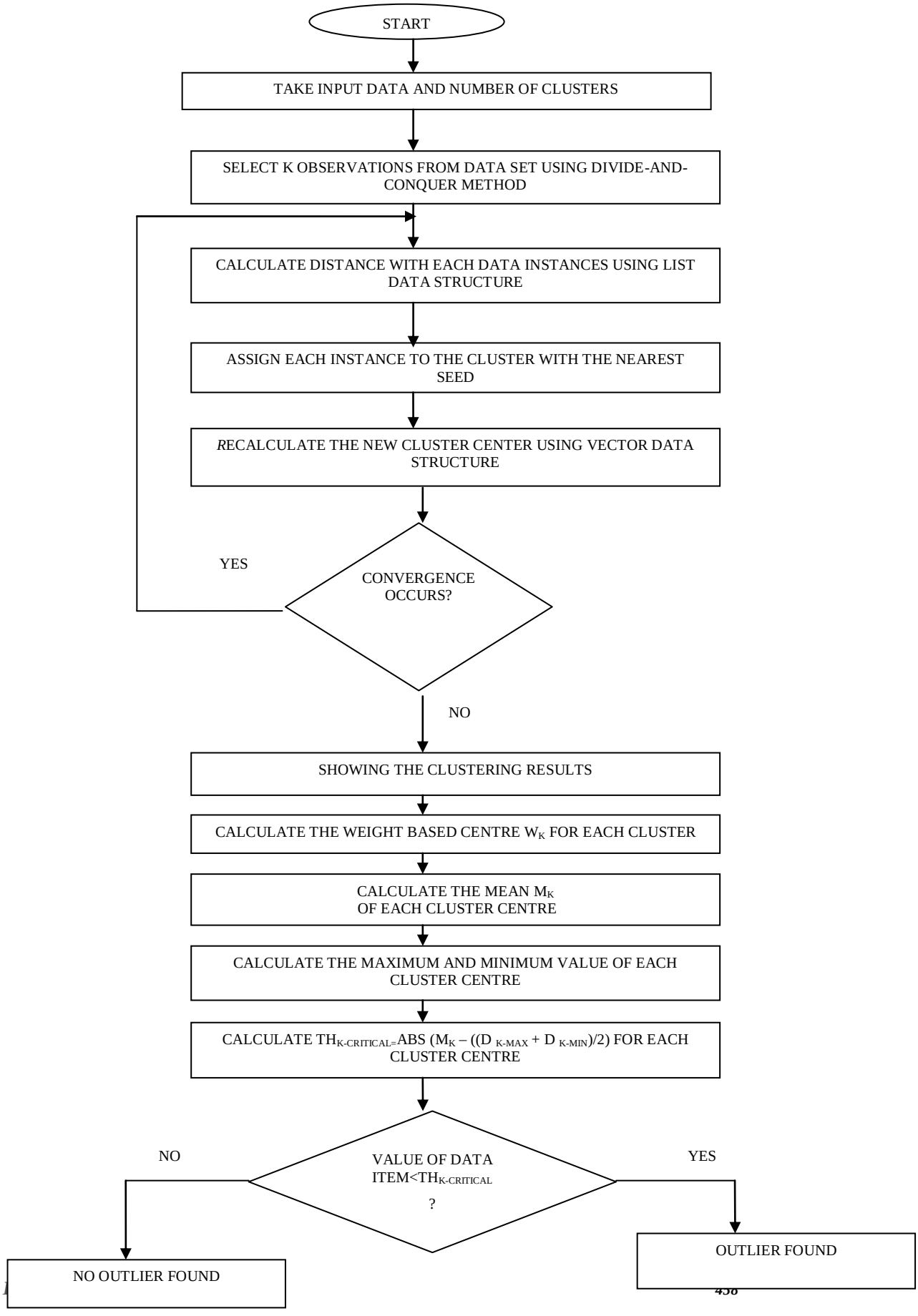
Step 11: Compare each data item d_x in a cluster with the threshold value $Th_{k\text{-}critical}$ where k is the cluster number.

Step 12: If the data value is d_x found less than $Th_{k\text{-}critical}$ then the given data item in cluster k is the outlier, where x=1,2,…n.

Step 13: Remove the outlier data items from the cluster.

Step 14: Repeat step 8-13 for each resultant cluster.

V. FLOW CHART OF THE PROPOSED ALGORITHM



VI. EXPERIMENTAL RESULTS

In this section, the Data is collected from UCI machine learning repository that provided various types of datasets. This dataset can be used for clustering, classification and regression. Dataset has various attribute and instances. A storage area of databases, domain theories and data generators are used by the machine learning community for the empirical analysis of machine. Data File Format is in .data and .xls excel file or .txt or .csv file format. This data file will be taken to find the outlier.

This section shows the performance comparison of the enhanced *k*-means clustering algorithm using Weight based center approach and distance based approach to find out the outliers. Result of Table I shows the Distance based approach and enhanced k-means algorithm using weight based approach to detect an outliers.

TABLE I
COMPARISION ANALYSIS OF DISTANCE BASED APPROACH AND ENHANCED K-MEANS CLUSTERING ALGORITHM USING WEIGHT BASED CENTER APPROACH TO DETECT AN OUTLIERS

Data Set	Number of Cluster	Number of Data items in Each Cluster	Number of Outliers	
			Distance Based Approach	Enhanced K-Means Clustering Algorithm using Weight Based Center Approach
Cancer Dataset (428)	C1	70	7	12
	C2	120	13	17
	C3	80	9	14
	C4	155	15	20
Fisher's Iris Dataset (150)	C1	48	3	6
	C2	39	2	5
	C3	63	5	10
Medical diabetes Dataset (578)	C1	146	15	17
	C2	125	10	14
	C3	133	8	9
	C4	90	4	6
	C5	84	3	6
Liver Disorder Dataset (670)	C1	182	17	24
	C2	148	15	20
	C3	126	10	15
	C4	114	8	12

Table-II shows the results of CPU elapsed time taken by Distance Based Approach and Enhanced K-means Clustering Algorithm using Weight Based Center Approach to detect an outlier over various Data Sets.

TABLE II
PERFORMANCE ANALYSIS OF ENHANCED K-MEANS CLUSTERING ALGORITHM AND DISTANCE BASED APPROACH TO DETECT AN OUTLIERS ON THE BASIS OF ELAPSED TIME

Elapsed Time in sec		
Data Set	**Distance Based Approach**	**Enhanced K-Means Clustering Algorithm using Weight Based Center Approach**
Cancer Data Set	**0.29246**	**0.09541**
Fisher's Iris Data Set	**0.13256**	**0.08321**
Medical diabetes Data Set	**0.56497**	**0.37597**
Liver Disorder Data Set	**0.60723**	**0.40975**

Figure 2, 3, 4 and 5 explicates performance of enhanced K-means algorithm using Weight Based Center approach and Distance Based Approach to detect outliers over various Data Sets.

PERFORMANCE OF THE PROPOSED ALGORITHM FOR OUTLIER DETECTION OVER VARIOUS DATA SETS

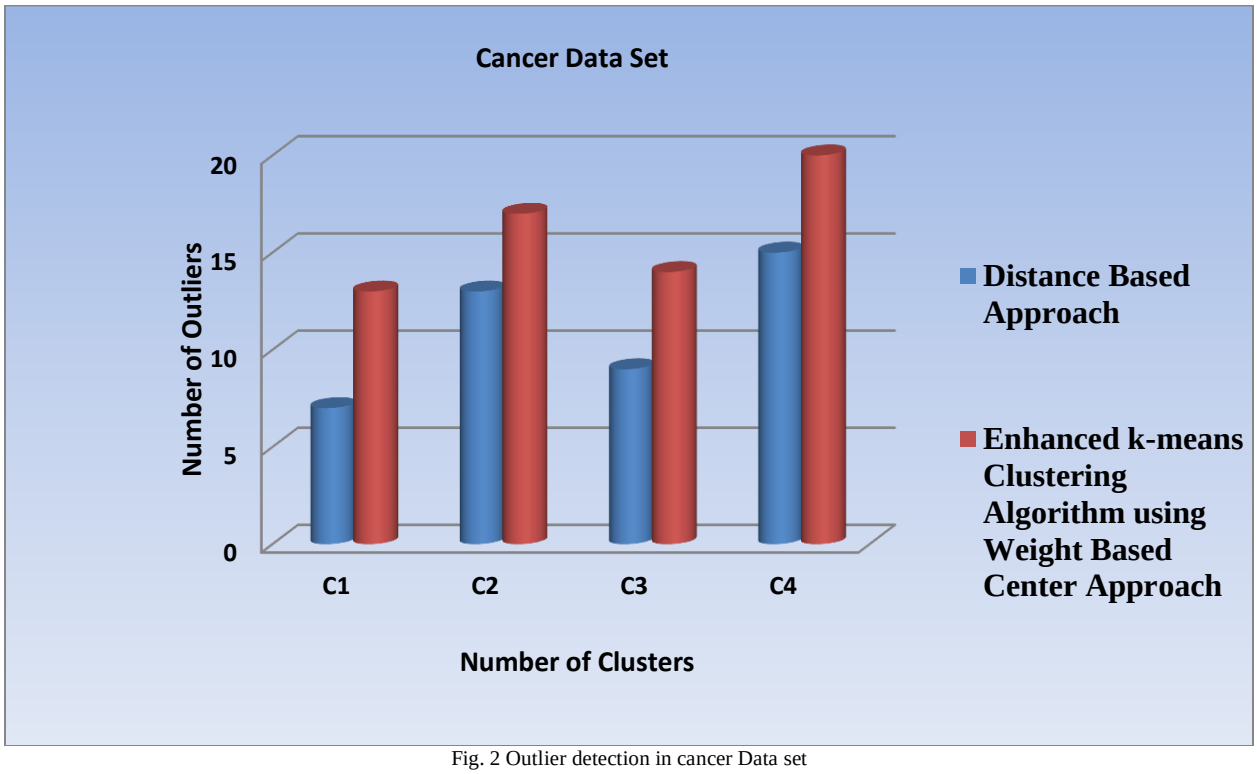


Fig. 2 Outlier detection in cancer Data set

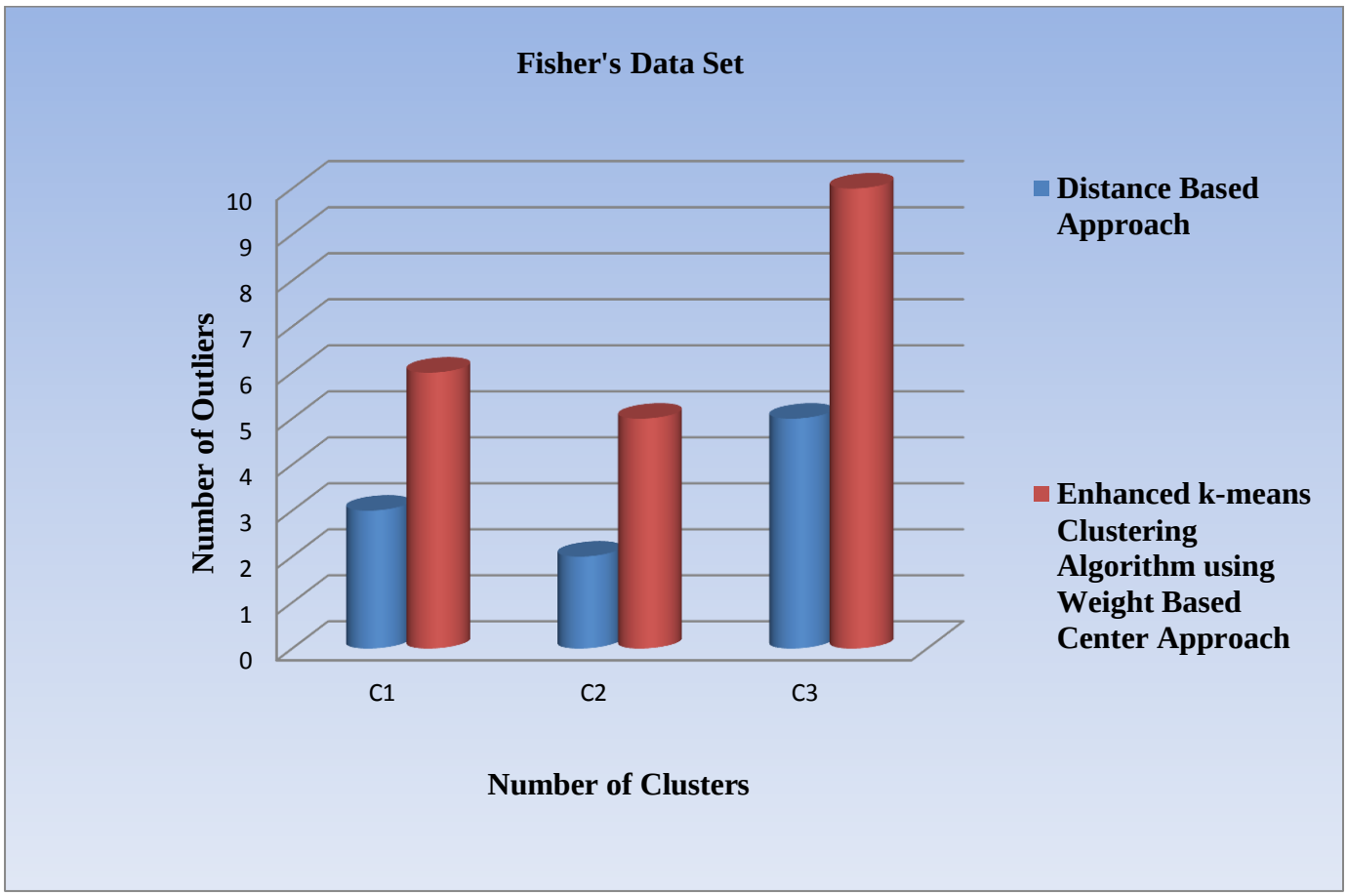


Fig. 3 Outlier detection in Fisher's Data Set

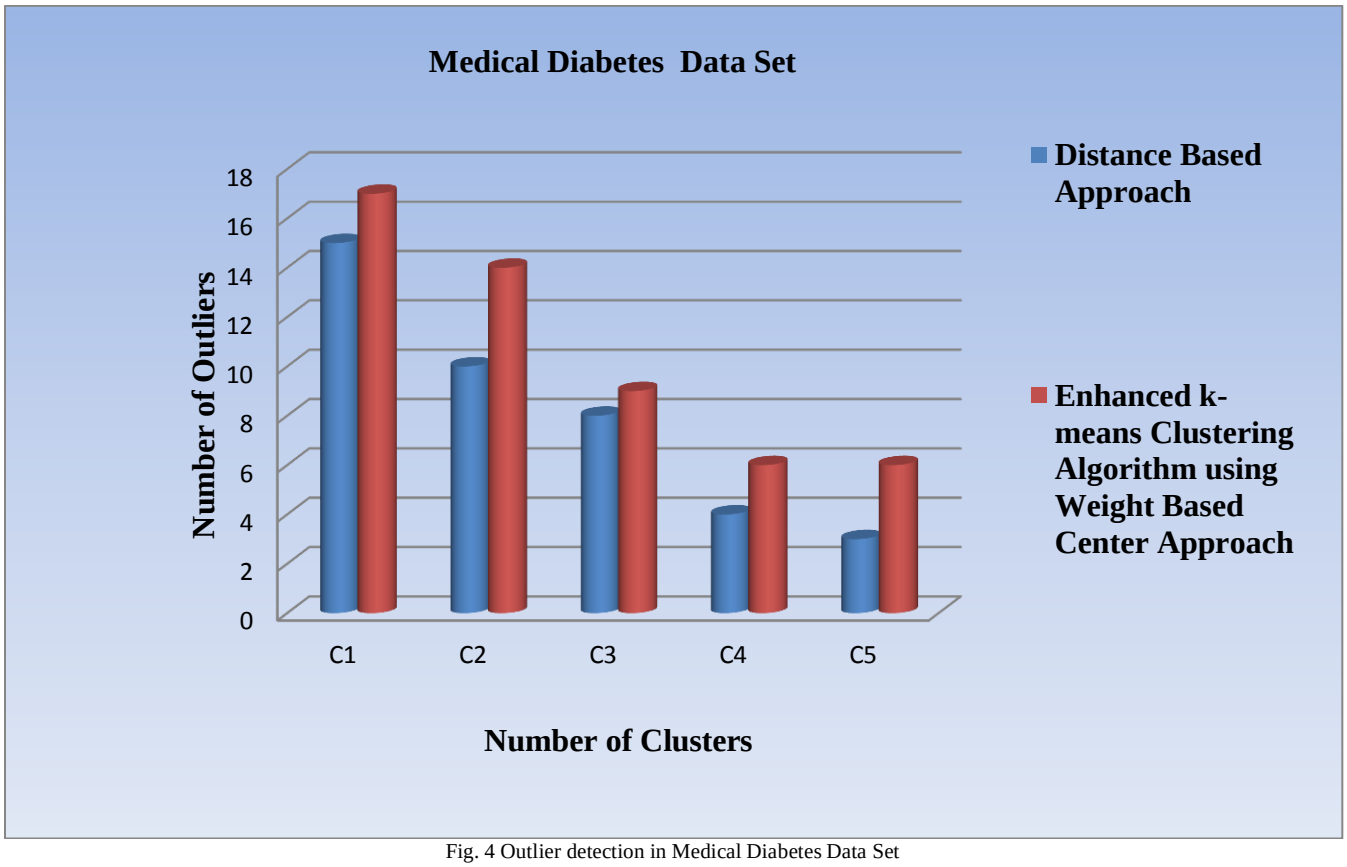


Fig. 4 Outlier detection in Medical Diabetes Data Set

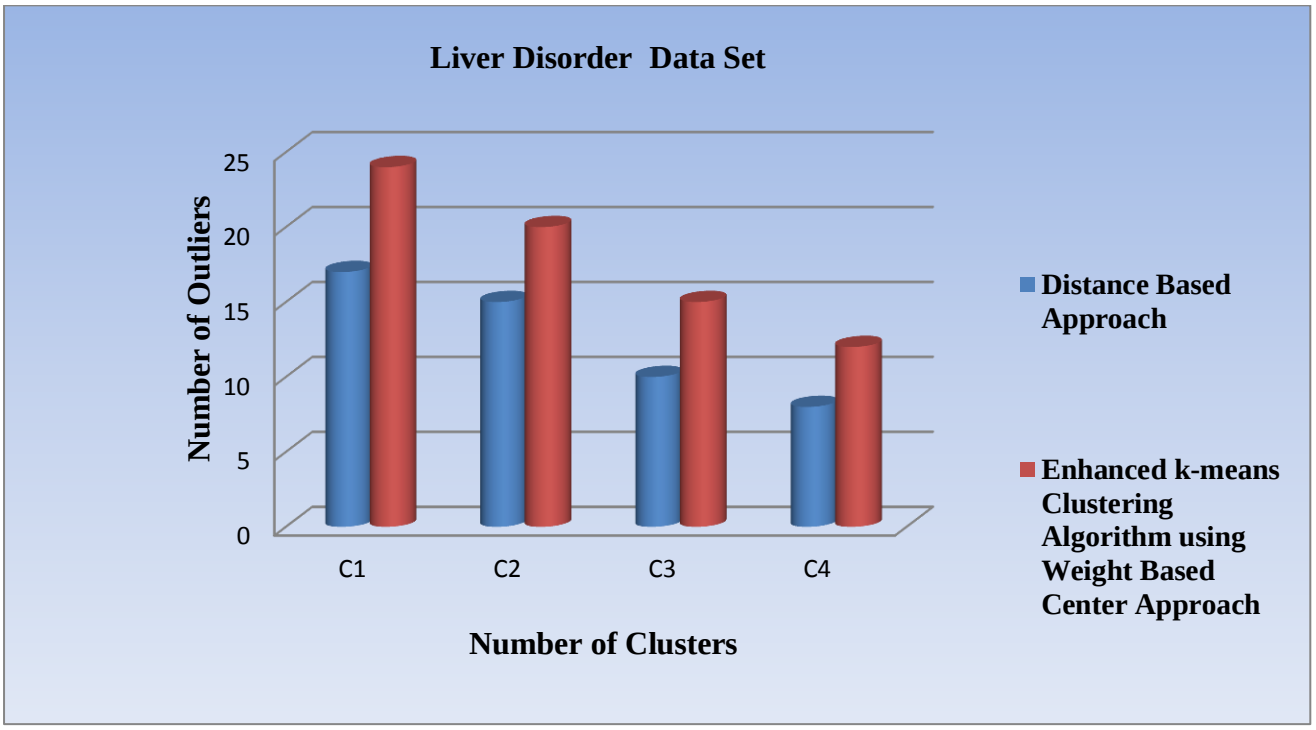


Fig. 5 Outlier detection in Liver Disorder Data Set

The figure 6 depicts the elapsed time taken by the enhanced K-means algorithm using weight based center approach and Distance Based Approach to detect an outlier over various data sets.

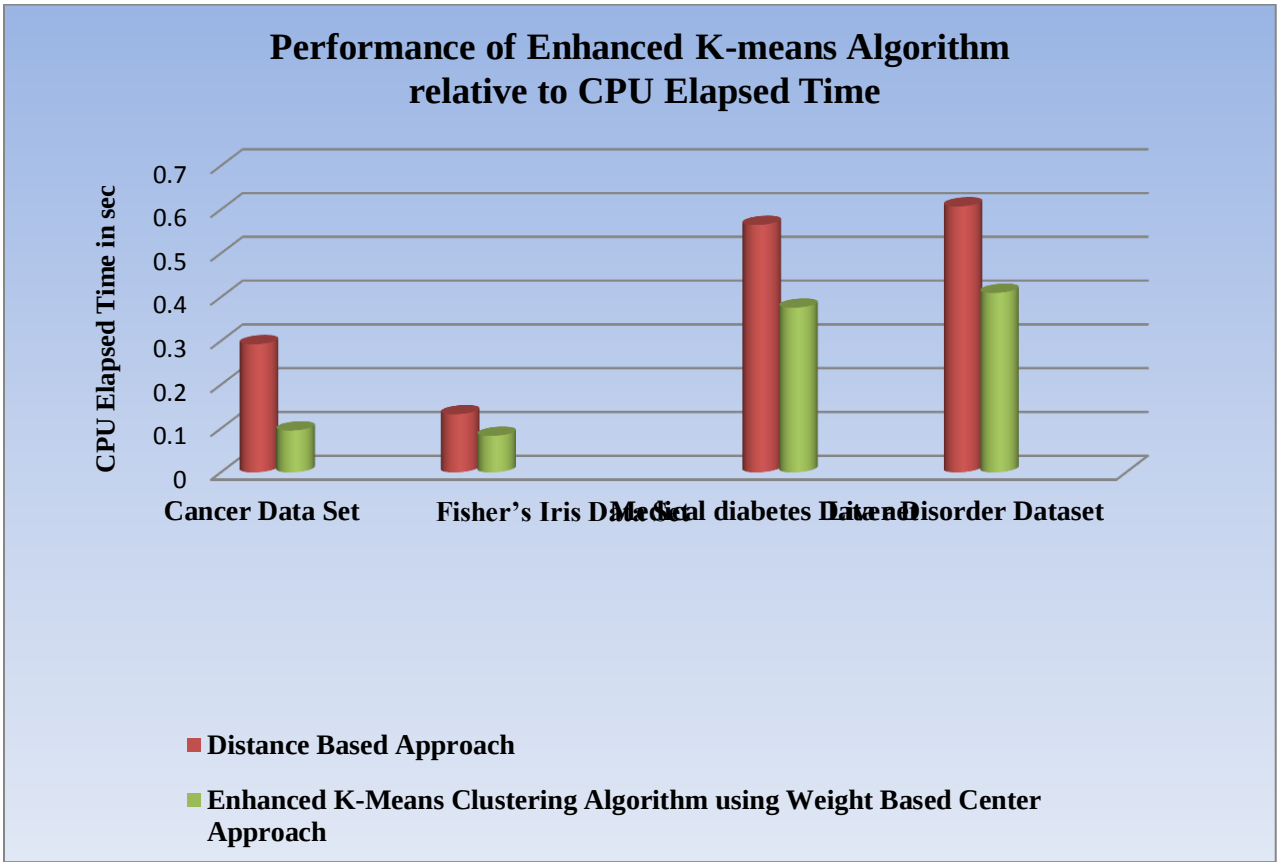


Fig. 6 CPU Elapsed Time for Outlier Detection

VII. CONCLUSION

Detection of outlier is an important task in Data Mining. Discovering outliers is the task that finds data objects that are dissimilar or inconsistent with respect to remaining data. The proposed algorithm efficiently detects outliers inside the clusters by using enhanced k-means clustering algorithm and weight based approach. In this work, we first group the data items in to number of clusters based on similarities between them. The computation time reduced considerably due to reduction in size of dataset. After that the outlier can be detected in each cluster using threshold value that can be calculated programmatically. The enhanced approach takes less computation time to discover outliers inside the clusters. The experimental results using the enhanced k-means clustering algorithm and weight based center approach with different datasets depict the elapsed time required to discover the outliers inside the clusters are comparatively less than the Distance based approach. So the enhanced k-means clustering method optimally detects the outlier in less time. Experimental results shows that the enhanced algorithm generates better results than the distance based approach relative to time and accuracy.

Enhanced approach is only deals with numerical data, so future work requires modifications that can make applicable for textual mining also. Future work has need of approach applicable for varying datasets.

References

[1] Pallavi Purohit "A new Efficient Approach towards k-means Clustering Algorithm", International journal of Computer Applications,Vol 65-no 11,march 2013.

[2] Wang Shunye , "An Improved K-means Clustering Algorithm Based on Dissimilarity", 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)Dec 20-22,2013, Shenyang, China IEEE.

[3] FAHIM , SALEM A.M, TORKEY F.A, RAMADAN M.A, "An efficient enhanced k-means clustering algorithm", Journal of Zhejiang University SCIENCE A ISSN 1009-3095 , ISSN 1862-1775.

[4] Juntao Wang & Xiaolong Su ,"An improved K-Means clustering algorithm" 2011 IEEE.

[5] Md. Sohrab Mahmud, Md. Mostafizer Rahman, and Md.Nasim Akhtar ,"Improvement of K-means Clustering algorithm with better initial centroids based on weighted average", 2012 7th International Conference on Electrical and Computer Engineering 20-22 December, 2012, Dhaka, Bangladesh, 2012 IEEE.

[6] H.S.Behera Abhishek Ghosh and Sipak ku. Mishra, " A New Hybridized K-Means Clustering Based Outlier Detection Technique For Effective Data Mining" , International Journal of Advanced Research in Computer Science and Software Engineering , Volume 2, Issue 4, April 2012 ISSN: 2277 128X.

[7] M. H. Marghny and Ahmed I. Taloba , " Outlier Detection using Improved Genetic K-means", IEEE TRANSACTIONS ON COMMUNICATIONS, VOL. 38, NO. 11, July 2013.

[8] 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, VOLUME 2, ISSUE 2, FEBRUARY 2010, ISSN: 2151-9617.PAGES 74-80.

[9] Mr. Raghav M. Purankar and Prof. Pragati Patil ,"A Survey paper on An Effective Analytical Approaches for Detecting Outlier in Continuous Time Variant Data Stream " , International Journal Of Engineering And Computer Science ISSN: 2319-7242, Volume 4 Issue 11 Nov 2015, Page No. 14946-14949.

[10] M. Gupta, J. Gao, C. C. Aggarwal, and J.Han, "Outlier Detection for Temporal Data", in Proc. Of the 13th SIAM Intl.Conf. on Data Mining (SDM), 2013.

[11] Prashant Chauhan and Madhu Shukla , " A Review on Outlier Detection Techniques on Data Stream by Using Different Approaches of KMeans Algorithm ", 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA) , IMS Engineering College, Ghaziabad, India @ 2015 IEEE.

[12] T. Divya and Dr. T. Christopher , "A Study of Clustering Based Algorithm for Outlier Detection in Data streams", International Journal Of Advanced Networking and Applications (IJANA), ISSN 0975-0282, March 2015.

[13] Neeraj Chugh, Mitali Chugh and Alok Agarwal , "Outlier Detection in Streaming Data A research Perspective", International Journal of Science, Engineering and Technology Research (IJSETR)Volume 4, Issue 3, March 2015.

[14] Safal V Bhosale et. al, "Outlier Detection in Straming data Using Clustering Approached " ,International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 5 (5) , 2014, 6050-6053.

[15] J.James Manoharan and Dr.S.Hari Ganesh ,"Improved K-means Clustering Algorithm using Linear Data Structure List to Enhance the Efficiency", International Journal of Applied Engineering Research, ISSN 0973-4562, 2015, Vol. 10, No. 20.

[16] J.James Manoharan and Dr. S.Hari Ganesh, "Initialization of optimized K-means Centroids using Divide-and-Conquer Method", ARPN Journal of Engineering and Applied Sciences", Vol. 11, No. 2, ISSN 1819-6608, January 2016.

[17] J.James Manoharan and Dr. S.Hari Ganesh, "A Framework for Enhancing the efficiency of K-means Clustering Algorithm to Avoid formation of Empty Clusters", Middle-East Journal of Scientific and Research (MEJSR),unpublished.

