Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 12, Number 5 (2016), pp. 4405–4418
© Research India Publications
http://www.ripublication.com/gjpam.htm

Time Complexity of K-Means and K-Medians Clustering Algorithms in Outliers Detection

Paul Inuwa Dalatu

*Department of Mathematics,
Faculty of Science,
Adamawa State University, P.M.B. 25 Mubi,
Adamawa State, Nigeria.*

Abstract

Data mining responsibilities show the broad features of the data in the database and also examine the current data in order to determine some arrangements. While, clustering establishes an important part of professed data mining, a procedure of exploring and analyzing large volumes of data in order to determine valuable information. Outliers are points that do not conform with the common performance of the data. Therefore, we applied the K-Means and K-Medians clustering algorithms to calculate the run time complexity analysis in identification of outliers in clustering analysis. The result shows that the K-Medians clustering algorithm with the use of medians served as robust for its faster and effective run time facilities better performance compared to the K-Means clustering algorithm in detecting outliers.

AMS subject classification:
Keywords: Outliers, Clustering, Time complexity, Simulation.

1. Introduction

Data mining responsibilities show the broad features of the data in the database and also examine the current data in order to determine some arrangements Dhiviya and Jayanthi (2015). Data mining is the knowledge discovery procedure by investigating the large bulks of data from various viewpoints and making it short into important information Chuchra (2012). Data mining, can be seen as correctly interdisciplinary issue, can be defined in many different means. According to Vijayarani and Nithya (2011), data mining is becoming an essential tool to transform the data into information. It is normally used in extensive series of profiling practices, such as marketing, fraud detection and scientific discovery. A lot of persons give data mining as a substitute for another commonly used term, knowledge discovery from data, or KDD, while some understand data mining as solely an important step in the procedure of knowledge discovery. The knowledge discovery procedure are listed according Han et al. (2011), as follows:

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
5. Data mining (an essential process where intelligent methods are applied to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures)
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users).

This shows that data mining as one step in the knowledge discovery procedures, albeit an important one because it reveals hidden patterns for evaluation. Hence, we implement a extensive observation of data mining functionality: Data mining is the method of discovering motivating patterns and knowledge from large aggregates of data. Clustering and classification are the two leading methods of data mining followed by association rules, predictions, estimations and regressions Sairam et al. (2011). Clustering is the procedure of partitioning a set of data objects into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters Han et al. (2011). It is observed that, different clustering techniques may produce different clustering on the same data set. The separating is not done by humans, but by the clustering algorithm. Therefore, clustering is very important in that it can lead to the detection of earlier unidentified groups within the data. According Han et al. (2011), Sairam et al. (2011), and Vijayarani and Nithya (2011), clustering has been broadly used in several applications such as machine learning, data mining, bioinformatics, business intelligence, market, image pattern recognition, image analysis, Web search, biology, and security. The clustering can be reflected as the best essential unsupervised learning issue; as each issue of this class, it deals with searching arrangement in a pool of unlabeled data Vijayarani and Jothi (2013).

Outliers are points that do not conform with the common performance of the data. By definition Vijayarani and Nithya (2011), outliers are infrequent existences and which characterize small part of the data. As stated by Han et al. (2011), an outlier is a data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism. Also, Sairam et al. (2011), that outliers are not part of a cluster, they are definite points which behave very different from the model. It occurs due to the changes in the system behavior, human error or instrument error. Vijayarani and Jothi (2013), that dependent on diverse presentation areas those irregular arrangements are often referred to as outliers, anomalies, discordant observations, faults, exceptions, defects, aberrations, errors, noise, damage, surprise, novelty, peculiarities or impurity.

Outliers can be classified in two sections Kim (2015); global and local outliers. Global outlier defines that it is far isolated from the center of the data set, this means that observation inconsistent with rest of the data set. Local outlier is an observation inconsistent with its neighborhoods. It is well-known that global outliers are not always outliers since it can be seen as another cluster if there are some cases over some threshold compared to the total number of cases. In clustering, outliers are measured as points that should be removed in order to make clustering more consistent Patel and Mehta (2011).

The rest of this paper is structured as follows. Section 2 presents section A: brief concepts of data clustering, section B: clustering algorithms literature, and section C: time complexity analysis. Section 3 presents description of algorithms which involves; section A: K-Means, section B: K-Medians, and section C: criteria for convergence. Section 4 presents results and simulation with the following; section A: K-Means initial clustering, section B: output of outliers, and section C: principles of convergence. Finally, section 5 concludes the paper and presents the possibility for future plan work.

2. Related Work

A. Brief Concepts of Data Clustering

Data clustering according to (Gan et al., 2007), also known as cluster analysis, segmentation analysis, taxonomy analysis, or unsupervised classification, is a process of forming groups of objects, or clusters, in such a way that objects in one cluster are very similar and objects in different clusters are dissimilar. Clustering establishes an important part of professed data mining, a procedure of exploring and analyzing large volumes of data in order to determine valuable information (Berry and Linoff, 2000). As listed by (Sairam et al., 2011; Sridhar and Sowndarya, 2010), that clustering algorithms can be categorize into three parts as follows:

- Hierarchical clustering
- Partitional clustering
- Spectral clustering

The simple requirements of clustering in data mining are Han et al. (2011), and Sairam et al. (2011):

- Scalability,

- Ability to deal with different types of attributes,
- Discovery of clusters with arbitrary shape,
- Requirements for domain knowledge to determine input parameters,
- Ability to deal with noisy data,
- Incremental clustering and insensitivity to input order,
- Capability of clustering high-dimensionality data,
- Constraint-based clustering, and
- Interpretability and usability.

The K-Means and K-Medians clustering algorithms are two of the type of algorithms under partitional clustering algorithms. Therefore, partitional techniques need to be given with usual initial seeds which are then enhanced iteratively, with distance metric, the descriptive points are selected at dissimilar iterations can be essential points such as the center of the cluster.

B. Clustering Algorithms Literature

Clustering algorithms according to Patel and Mehta (2011), produce clusters having similarity between data objects based on features belong to same cluster. Clustering algorithms have been widely spread in many fields of publications such as pattern recognition, artificial intelligence, information technology, medical, machine learning, image processing, biology, psychology, and marketing Gan et al. (2007), and Patel and Mehta (2011). Therefore, the objective of clustering is to isolate a finite, unlabeled data set into a finite and separate set of natural, hidden data arrangements, rather than to provide correct representation of undetected samples created the same probability distribution Xu and Wunsch (2009).

Han et al., (2011) stated that, an outlier is a data object that differs considerably from the rest of the objects, as if it were produced by a different mechanism. Also, Barnet and Lewis (1994) show that an outlying observation, is one that seems to depart significantly from other members of the sample in which it occurs.

Therefore, proximity-based outlier detection method has two types: distance-based and density-based methods. A distance-based outlier detection method accesses the neighborhood of an object, which is demarcated by a given radius. An object is taken as an outlier if its neighborhood does not have sufficient other points. A density-based outlier detection method examines the density of an object and that of its neighbors. Here, an object is recognized as an outlier if its density is reasonably much lower than that of its neighbors Han et al. (2011).

Let us consider Clustering-Based Methods. The idea of outliers is extremely associated to that of clusters. Clustering-based methods detect outliers by investigating the relationship between objects and clusters. Instinctively, an outlier is an object that belongs to a small and distant cluster, or does not belong to any cluster. Han et al, (2011), supported it with three common methods to clustering-based outlier detection. Deliberate on an object:

- Does the object belong to any cluster? If not, then it is identified as an outlier.
- Is there a large distance between the object and the cluster to which it is closest? If yes, it is an outlier.
- Is the object part of a small or sparse cluster? If yes, then all the objects in that cluster are outliers.

Therefore, in order to detect outliers with the use of K-Means algorithm Zhao (2011) stated that, with K-Means, data are separated into k groups by allocating them to the nearby cluster centers. Next, we can calculate the distance (or dissimilarity) between object and its cluster center, and choose those with far distances as outliers. In this section, we are going to discuss several methods proposed by lot of researchers to detect outliers in K-Means and K-Medians clustering algorithms.

The paper Sairam et al. (2011), used simulation studies based on time complexity by comparing the performance analysis of K-Means and K-Medians clustering algorithms in detecting outliers. The results shows K-Means takes more time to calculate outliers to K-Medians and in minimizing the errors, K-Medians clustering algorithm is much efficient than K-Means clustering algorithm.

The paper Sridhar and Sowndarya (2010), presents the performance of K-Means clustering algorithm, in mining outliers from large datasets. Adopting three methods of algorithms, the run time of the third method takes longer runtime, although is more efficient in detecting the outliers. The second method is similar to first method as both methods are less efficient; all these are based on simulation results.

The paper Patel and Mehta (2011), presented the performance of modified K-Means clustering algorithm with data preprocessing method consisting cleaning and normalization methods with the purpose to check the impact of outlier detection with instinctive initialization of seed values on datasets. Therefore, the performance analysis of calculated MSE for Mk-Means and Mk-Means using the three normalization methods with outlier removal displays the finest and effective results for Mk-Means which create lowest MSE and increase the effectiveness and excellence of results created by this algorithm.

The paper Vijayarani and Nithya (2011), presented some algorithms as PAM, CLARA, and CLARANS; with new proposed clustering algorithm called ECLARANS for outliers detection. The experimental result indicates that, the proposed algorithm ECLARANS increases the accurateness of detection and CLARANS decreases the time complexity when related with other algorithms.

The paper Vijayarani and Jothi (2013), presented two clustering algorithms; BIRCH with K-Means Birch with CLARANS were used for clustering the data substances and ruling the outliers in data streams. Based on the analysis of the clustering and outlier presentation of BIRCH with CLARANS and BIRCH with K-Means clustering algorithm for outliers detection. The experimental results shown that the clustering and outlier detection accurateness is more capable in BIRCH with CLARANS clustering compare to BIRCH with K-Means clustering.

The paper Pamula et al. (2011), proposed a clustering based technique to seizure outliers. By applying K-Means clustering algorithm to partition the data set into groups, where points which are lying near the center of the cluster may not point for outlier and such points can be pruned out from each cluster. Due to the reduction in the size of the data set, the computation time reduced considerably. Local distance-based outlier factor was used to quantify the amount an object departs from its neighborhood. The proposed method for accuracy of detecting outliers gives higher than the existing methods, although, some points were pruned.

The paper Vu and Gopalkrishnan (2009), presented a new technique for detecting distance-based outliers, with the objective of decreasing implementation time related with the detection procedure. The proposed method, MIRO, contains of two pruning stages of treating which lead to amortized effectiveness. In the first stage, partition-based technique is engaged to remove point clusters for the later treating step. Further advantage of the first stage is to calculate an initial value of the outlier cutoff threshold which is employed in the nested-loop stage. In the second stage of MIRO, two pruning guidelines are engaged to further decrease the complete temporal cost.

The paper Zhou et al. (2009), with three-stage K-Means algorithm of $O(nkt)$ polynomial time is proposed to effectively cluster the numerical data points and check out the outliers. The first-stage clustering can preliminarily define the k clusters. The second-stage clustering can check out the local outliers of each clusters and improve centroids after eliminating the impact of the local outliers. The data points not fitting to any of the clusters are finally recognized as global outliers. The last stage cluster-merging combines the clusters, those alike density and have mutual data points, into one cluster. The rule backings the separating of the data into clusters of dissimilar densities and dissimilar sizes.

The paper Kim (2015), proposed an automated K-Means clustering process, which combines the VS-KM algorithm and serve as the proof of identity of outliers, also applied to the variable selection procedure. The Automated K-Means clustering process comprises of three stages:

(i) routinely computing the cluster number and initial cluster center each time new variable is added,

(ii) detecting outliers for each cluster subject on used variables,

(iii) choosing variables outlining cluster organization in a forward manner.

To detect outliers, we used a hybrid technique joining a clustering based technique and distance based technique. Simulation outcomes show that the proposed automated K-Means clustering process is effective to choose variables and detect outliers.

The paper Ahmed and Mahmood (2013), proposed a method for modification of the known K-Means algorithm to detect outliers. The identified outliers are removed from the dataset to modify clustering accurateness. The method is authenticated by comparing against current methods and bench performance. Experimental outcomes on benchmark datasets indicate that, the proposed method outperforms present techniques on numerous processes.

The paper Dhiviya and Jayanthi (2015), proposed a technique to investigate imperfectly labeled data, which has candidature value towards the normal and abnormal categories The Cuckoo-SVDD handles imperfectly labeled data and effectively identify outliers and offers global optimal result based on the learned classifier. The reason is that the kernel function used for cross validation and kernel LOF-based technique to compute the candidature values. Density based method reliably achieves on the data with variable density. The proposed technique widely contracts with the imperfectly labeled data and reaches the accurateness of identifying outliers.

The paper Vijayarani and Jothi (2014), presented two partitioning clustering algorithms as: CLARANS and E-CLARANS (Enhanced Clarans) are used for clustering and detecting outliers in data streams. The aim of the paper is to implement the clustering procedure and identifying outliers in data streams. Two presentation issues such as clustering accurateness and identification of outlier accurateness are used for observation. Through the experimental outcomes, it is detected that the proposed E-CLARANS clustering algorithm presentation is more correct than the current algorithm CLARANS.

The paper Jobe and Pokojovy (2015), proposed a computer-intensive cluster-based technique that integrates a reweighted version of Rousseeuw‘s minimum covariance determinant method with a multi-step cluster-based algorithm that initially screens out possible masking points. The experimental results compare the most robust process, simulation studies has shown that the proposed method is better for outlier identification.

The paper Murugavel and Punithavalli (2011), presented improved hybrid methods, consisting of three partition-based algorithms, PAM, CLARA, and CLARANS were merged with K-Medoid distance based outlier detection to enhance the outlier detection and elimination procedure. The experimental results showed that CLARANS is the finest candidate in view of outlier identification, followed by CLARA and PAM.

The paper Dhaliwal et al. (2010), introduced a new clustering based technique, which splits the stream into chunks and clusters, where each chunk using k-median into variable number of clusters. Instead of keeping whole data stream chunk in memory, is being represented with the weighted medians originate after mining a data stream chunk and pass that information along with the newly arrived data chunk to the next stage. The weighted medians originate in each stage are verified for outlierness and this gives number of stages, which is confirmed as a real outlier or an inlier. Their method is perfectly better than the k-means as it could not fix the number of clusters to k either gives interval to it and offers a more steady and improved result which turns in poly-logarithmic space.

C: Time Complexity Analysis

Definition Black (2016): An abstract measure of the implementation of an algorithm, usually the time or memory needed, given the issue size of n, which is sample number of objects. Casually, given some equation $f(n) = O(g(n))$ it means less than some constant multiple of g(n). The notation is read, "f of n is big *oh* of g of n". Defined as $f(n) = O(g(n))$ means c and k are positive constants, such that $0 \leq f(n) \leq cg(n)$ for all $n \geq k$. The values of c and k have to be fixed for the function f and must not influenced by n. Nopiah et al. (2010), stated that time complexity analysis is portion of computational complex model that is used to describe an algorithm‘s use of computational properties; in this situation, the worst case processing time expressed as a function of its input using big Omicron ($big - O$) notation Black (2008), and Knuth (1976).

The big-O notation is usually expressed the upper bound of the growth rate of a function and frequently define asymptotic performance Knuth (1976). The big-O notation is defined using set notation as follows Nopiah et al. (2010):

$$O(g(n))\{f|\exists c > 0, \exists n_0 > 0, \forall n \geq n_0 : 0 \leq f \leq cg(n)\} \tag{2.1}$$

Alternatively, $f \in (g(n))$ if and only if there exist positive constants c and n_0 such that for all $n \geq n_0$, the inequality $0 \leq f \leq cg(n)$ is satisfied. Approximately, f is big-O of $g(n)$, or that $g(n)$ is an asymptotic upper bound for f Black (2008). Regarding the time complexity analysis, we use the term $T \in O(g(n))$ and say that the algorithm has order $g(n)$ complexity. Therefore, the time taken to compute an issue of size n is in the set of functions denoted by $O(g(n))$.

Therefore, to compute the run time complexity analysis in clustering analysis: Take N tuples in the dataset then, the similarity matrix can be calculated in $O(KNT)$. Let N be the number of tuples in the dataset. K is the number of clusters and T is the time to calculate the distance between two data objects. Time complexity of each iteration is $O(KNT)$. I is the number of iteration in k-means algorithm. Where, I number of iteration in the time complexity of this algorithm is $O(IKNT)$ Patel and Mehta (2011).

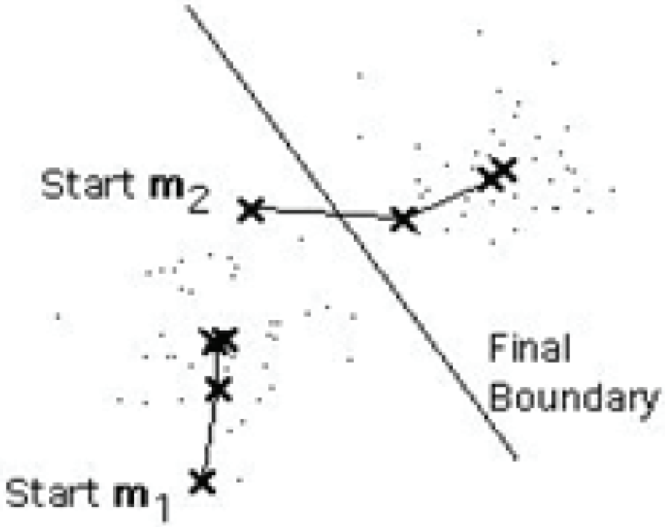
3. Description of Algorithms

A. The K-Means Algorithm

Aggarwal and Reddy (2013), Gan et al. (2007), Sairam et al., (2011) and, Xu and Wunsch II (2009):

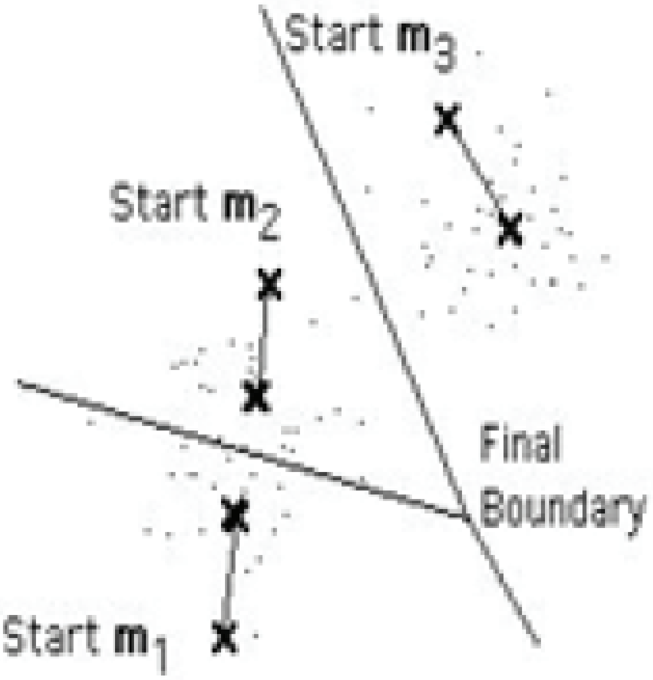
- Allocate initial values for means $m1, m2, \ldots, mk$,
- Allocate each object to the cluster with nearest mean, and
- Compute new mean for each cluster until no change for each cluster.

Here, illustrating the means $m1$ and $m2$ move into the centers of two clusters.



Source: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans_html

Remarks 3.1. It is a simple type of K-Means process. It can be seen as a greedy algorithm for partitioning the n samples into k groups, so to minimize the sum of the squared distances to cluster centers.



Source: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans_html

Remark 3.2. The above illustrating three means, is the same algorithm used to the same data that resulted in 3-means. It is being observed either better or worse than the 2-means clustering. Regrettably, there is no broad theoretical explanation to find the optimum number of cluster for any given data set.

B. The K-Medians Algorithm

Aggarwal and Reddy (201) and Sairam et al. (2011):

- Allocate initial values for means $m1, m2, \ldots, mk$,
- Allocate each object to the cluster with nearest mean,
- Compute new median for each cluster until no change for each cluster,
- If the total number of points in the cluster ends with an even number, then take the middle two values and calculate the new median, and
- If the total number of points in the cluster ends with an odd number then take the middle value as median.

C. Criteria for Convergence

:

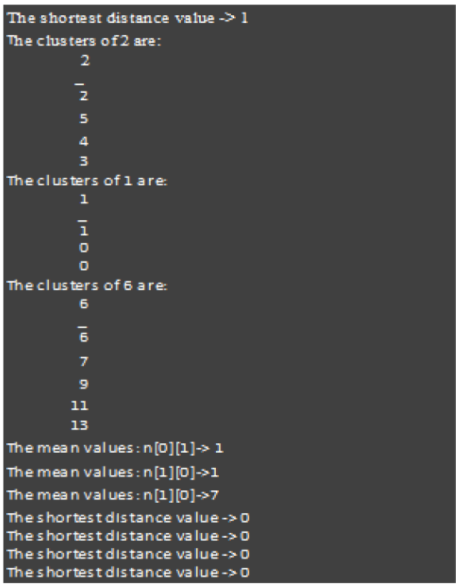
The criteria for convergence has to follow strictly, with the objective of K-Means clustering is to minimize the squared error function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \parallel x_i - c_j \parallel^2 \tag{3.2}$$

Therefore, when the previous mean value and current value becomes equal, then, it is said the criteria for convergence is achieved.

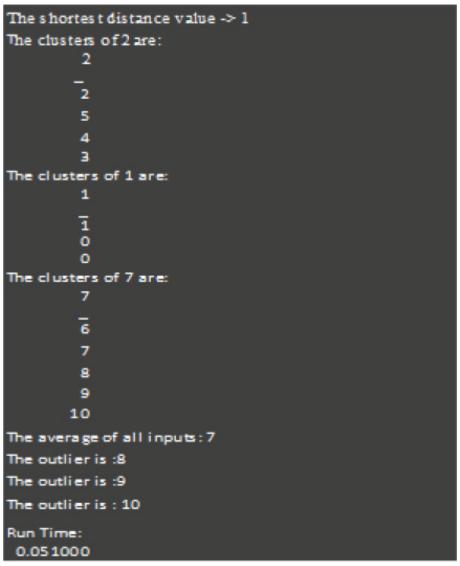
4. Results and Simulation

A: K-Means initial clustering



```
The shortest distance value -> 1
The clusters of 2 are:
          2
         _
          2
          5
          4
          3
The clusters of 1 are:
          1
         _
          1
          0
          0
The clusters of 6 are:
          6
         _
          6
          7
          9
         11
         13
The mean values : n[0][1]-> 1
The mean values : n[1][0]->1
The mean values : n[1][0]->7
The shortest distance value -> 0
The shortest distance value -> 0
The shortest distance value -> 0
The shortest distance value -> 0
```

B: Output Outlier



```
The shortest distance value -> 1
The clusters of 2 are:
          2
         _
          2
          5
          4
          3
The clusters of 1 are:
          1
         _
          1
          0
          0
The clusters of 7 are:
          7
         _
          6
          7
          8
          9
         10
The average of all inputs : 7
The outlier is :8
The outlier is :9
The outlier is : 10
Run Time:
 0.051000
```

5. Conclusion and Future Plan

Based on the dataset obtained from Buzzi-Ferraris and Manent (2011), with the following inputs: 31.1, 31.6, 31.2, 31.2, 31.3, 311.1, 31.3, 31.1, 31.4, 31.3, 32.1, 31.0. The K-Means clustering algorithm is much more slower in using more time in detecting the outlier which its only identified 311.1, while, the K-Medians clustering algorithm is

Table 1: Principles of Convergence

Total No. of Inputs	Total No. of Clusters	K-Means: Run time	K-Medians: Run time
12	2	0.114300	0.111200
12	3	0.112100	0.110100
12	4	0.112200	0.111100
12	5	0.114300	0.111200

Source of data: Buzzi-Ferraris, (2011).

much faster in calculating of time for the identification of outliers, against the K-Means, the K-Medians had correctly identified three outliers as 31.6, 32.1, and 311.1.

Therefore, in errors minimization, the K-Medians clustering algorithm is more effective probably because it uses median as robust for computing the clustering compare to the K-Means which uses mean which is sensitive to outliers. In our future plan, we want to extend this work to hierarchical partitioning with different set of algorithms.

References

[1] Aggarwal, C. C., and Reddy, C. K. (Eds.), 2013, “Data clustering,” algorithms and applications. CRC Press.

[2] Ahmed, M., and Naser, A., 2013, June, “A novel approach for outlier detection and clustering improvement,” In Industrial electronics and applications (ICIEA), 2013 8th IEEE conference on, pp. 577–582.

[3] Barnett, V. and Lewis. T., 1994, “Outliers in statistical data,” John Wiley and Sons, 920, l.

[4] Black, P. E., 2008, “Big-O notation, Dictionary of Algorithms and Data Structures” [online], U.S. National Institute of Standards and Technology, (accessed 16 March, 2016). Available from: http://www.itl.nist.gov/div897/sqg/dads/HTML/bigOnotation.html

[5] Buzzi-Ferraris, G., and Manenti, F., 2011, “Outliers detection in large data sets,” Computers and chemical engineering, 35(2), pp. 388–390.

[6] Chuchra, R., 2012, “Use of Data Mining Techniques for the Evaluation of Student Performance, A Case Study,” International Journal of Computer Science and Management Research, 1(3), pp. 425–433.

[7] Dhaliwal, P., Bhatia, M. P. S., and Bansal, P., 2010, “A Cluster-based Approach for Outlier Detection in Dynamic Data Streams,” KORM: k-median Outlier Miner.

[8] Dhiviya, S. and Jayanthi, P., 2015, "An Experimental Approach for Outlier Detection with Imperfect Data Labels," International Journal of Computer Science and Engineering Technology (IJCSET), 6(3), pp. 138–147.

[9] Gan, G., Ma, C., and Wu, J., 2007, "Data clustering: theory, algorithms, and applications," Vol. 20.

[10] Han, J., Kamber, M., and Pei, J., 2011, "Data mining: concepts and techniques," Elsevier.

[11] http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans_html, 2015.

[12] Jobe, J. M., and Pokojovy, M., 2015, "A Cluster-Based Outlier Detection Scheme for Multivariate Data," Journal of the American Statistical Association, 110(512), pp. 1543–1551.

[13] Kim, S. S., 2015, "Variable Selection and Outlier Detection for Automated K-means Clustering," Communications for Statistical Applications and Methods, 22(1), pp. 55–67.

[14] Knuth, D. E., 1976, "Big omicron and big omega and big theta," ACM Sigact News, 8(2), pp. 18–24.

[15] Linoff, G., and Berry, M., 2000, "Mastering Data Mining: The Art and Science of Customer Relationship Management," New York.

[16] Murugavel, P. and Punithavalli, M., 2011, "Improved Hybrid Clustering and Distance-based Technique for Outlier Removal," International Journal on Computer Science and Engineering (IJCSE), 3(1), pp. 333–339.

[17] Nopiah, Z. M., Khairir, M. I., Abdullah, S., Baharin, M. N., and Arifin, A., 2010 February, "Time complexity analysis of the genetic algorithm clustering method," In Proceedings of the 9th WSEAS international conference on Signal Processing, robotics and automation, pp. 171–176.

[18] Pamula, R., Deka, J. K., and Nandi, S., 2011 February, "An outlier detection method based on clustering," In Emerging Applications of Information Technology (EAIT), 2011 Second International Conference on, pp. 253–256.

[19] Patel, V. R., and Mehta, R. G., 2011, "Impact of outlier removal and normalization approach in modified k-means clustering algorithm," IJCSI International Journal of Computer Science Issues, 8(5), pp. 331–336.

[20] Sairam, M., 2011, "Performance Analysis of Clustering Algorithms in Detecting Outliers," International Journal of Computer Science and Information Technologies, 2(1), pp. 486–488.

[21] Sridhar, A., and Sowndarya, S., 2010, "Efficiency of k-means clustering algorithm in mining outliers from large data sets," International Journal on Computer Science and Engineering, 2(9), pp. 3043–3045.

[22] Vijayarani, S., and Jothi, P., 2013, "An efficient clustering algorithm for outlier detection in data streams," International Journal of Advanced Research in Computer and Communication Engineering, 2(9), pp. 3657–3665.

[23] Vijayarani, S. and Jothi, P., 2014, "Partitioning Clustering Algorithms for Data Stream Outlier Detection," International Journal of Innovative Research in Computer and Communication Engineering, 2(4), pp. 3975–3981.

[24] Vijayarani, S., and Nithya, S., 2011, "An Efficient Clustering Algorithm for Outlier Detection," International Journal of Computer Applications, 32(7), pp. 22–27.

[25] Vu, N. H., and Gopalkrishnan, V., 2009, "Efficient pruning schemes for distance-based outlier detection," In Machine Learning and Knowledge Discovery in Databases, pp. 160–175.

[26] Xu, R., and Wunsch, D. C., 2009, "Clustering," Hoboken.

[27] Zhao, Y., 2011, "R and Data Mining: Examples and Case Studies".

[28] Zhou, Y., Yu, H., and Cai, X., 2009 December, "A novel k-means algorithm for clustering and outlier detection," In Future Information Technology and Management Engineering, Second International Conference on, pp. 476–480.