Pak. J. Agri. Sci., Vol. 54(1), 227-231; 2017
ISSN (Print) 0552-9034, ISSN (Online) 2076-0906
DOI: 10.21162/PAKJAS/17.4743
http://www.pakjas.com.pk

MULTIVARIATE OUTLIER DETECTION: A COMPARISON AMONG TWO CLUSTERING TECHNIQUES

Muhammad Zafar Iqbal*, Mehvish Riaz and Waqar Nasir

Department of Mathematics and Statistics University of Agriculture, Faisalabad, Pakistan.
***Corresponding author's email: mzts2004@hotmail.com**

In data analysis, the first step for broad analysis is outlying items. Outlying items cause false results, biased parameter estimation and model misspecification. Some of the methods for detection of outlier are: Distance based method, Distribution based method, Density based method and Clustering based method. In this research, clustering based method is used for outlying items. The purpose of this study is to detect outlier in multivariate data by performing cluster analysis. The K means and Partitioning Around Medoid (PAM) methods were performed. After cluster the data outliers are detected by measuring distance. The comparison of the clustering techniques in outlier detection methods are analyzed. The secondary data was used for analysis. In this data set impact of zinc phosphide with three sub lethal doses (concentrations) on different body tissues of rat are used to find any anomalies. *R* software was used for data analysis.

Keywords: Cluster analysis, K means, PAM, mahalanobis distance, outlier detection

INTRODUCTION

In several data analysis procedures, an enormous number of variables are reported or tested. The first step for comprehensive data analysis is to find out the outlying items. However, outlying items are usually treated as noise or error, they can convey meaningful knowledge. A detected outlier is essential for abnormal data that can skeptically lead to biased parameter estimates, model misspecification and false consequences. Hence, this one is meaningful to find them preceding to modeling and analyzing (Ben-Gal, 2010).

Hawkins (1980) defined that an outlier is an observation which is different from other items and produce doubt that it was made by a diverse tool. Johnson (1992) described in a data set an outlier is an observation which one looks to stand uneven with the rest of that data. Explained methods for detection of outliers had been proposed for several presentations, like that credit card fake recognition, medical trials, elective anomaly investigation, severe weather prediction, environmental data structures, player presentation analysis, data cleaning and further data mining assignments.

There are many methodologies for finding outliers. These methodologies are: Distance based method, Distribution based method, Density based method and Clustering based method (Kaur and Kaur, 2013).

In clustering based methods for outlier detection, usually two clusters were constructed. A cluster consisted of minimum data points would likely to have outlier(s). The advantage of clustering based methods is that there is no need to supervise classification. Furthermore, clustering based methods are accomplished for implemented in an incremental manner (Jaykumar and Thomas, 2013).

Clustering is a most important task in analysis of data and applications of data mining. It is the task of mixture a set of items so that item in the identical set is more associated to each other than those in other sets. Cluster is a well-ordered list of data which have the familiar features. Cluster analysis is done by finding similarities among data with respect to the structures establish in the data set and group the similar data objects into clusters. Clustering is a process of unsupervised classification. In which a post -handling step after clustering is involved to decide the size of the clusters. The distance of the clusters is then computed, with which the outliers are detected. These methods use the distance measure among two items and clustering is based by grouping items which have a minimum distance from the center of the cluster.

High superiority groups with high between class similarity and low within class similarity show that it is good clustering method. The perfection of a clustering result depends on equally the similarity measure used by the method and its application. The perfection of a clustering method is calculated by its ability to find some or all of the unseen patterns. The similarity of a cluster can be expressed by the distance function. The types of data that are used for analysis of clustering are Binary variables, Interval scale variables, Nominal, ordinal, and ratio variables, mixed type variables.

MacQueen (1967) proposed the term K Means, a non-hierarchical technique, is emerging as a popular choice in the community of data mining. The K means is a simple, prototype based partition based clustering method which tries to discover a user designated k number of groups. Then these groups are categorized by their centers. A cluster mean is naturally the centroid of a point in the cluster. The K Means method is modest to appliance and easy to adjust, track rather fast and common in practice. The procedure consists of two separate steps, in first to select k centers arbitrarily, in which k is fixed in advance. In the second step to assign each data object to the adjacent center cluster means. Recalculation remains done on the average of the clusters, once all the data items are contained within in some clusters. The iterative procedure continuing frequently till the criterion task become minimum (Behera *et al.*, 2012). Kaufman and Rousseeuw (1987) developed a procedure PAM, a centrally located in clusters an arrangement of items is known as medoids. Items which are uncertainly distinct as medoids are located into a set S of particular items. If O is the set of items that the set U = O − S is the set of not selected items. The objective of the procedure is to reduce the usual dissimilarities of items to their neighboring nominated item. Consistently, minimize the sum of dissimilarities in item and their closer nominated item. The purpose of this study is to detect outlier in multivariate data. In this research, clustering based method is used for detection of outliers. The comparison of the clustering techniques in outlier detection methods is analyzed. There are many methods of clustering for detection of outliers. These are K means, PAM, CLARA, and CLARANS. In this research paper, two clustering algorithms are analyzed. After clustering the data outliers are detected. Compare the two methods for outlier detection with biplot of PC analysis. The analysis of outlier detection was done on *R* (Zhao, 2013).

Chawla and Gionis (2013) presented a combined approach for simultaneously clustering and determining outliers in the data set. Aroma and Karkkainen (2006) suggested that clustering the information is a data mining technique and an unsupervised data analysis, which polished more theoretical views to the intrinsic structure of a data set of partitioning it into a number of fuzzy or disjoint groups. Sivaram and Saveetha (2013) proposed that detection of outlier is a very significant task in an extensive selection of application areas. The situation has numerous usage in many submissions. In this research paper, a planned method built on clustering methods for outlier detection is obtainable. This method firstly performed partition methods use these methods PAM, CLARA, CLARANS and CLATIN. The method yields clusters and cluster centers. Small clusters are at that time restrained and declared as outlier clusters. Explained in clustering procedures, outliers are by multiple of clustering procedures and could not rank the importance of outliers. In this study, three methods PAM, CLARA and CLARANS are joined with k medoid distance centered to recover the detection of outlier and removal procedure.

MATERIALS AND MATHODS

Outlier detection terminology refers to the task of finding outliers as per behaviour of data and distribution of data. There are different types of techniques used in data streams called statistic. Clustering is an important task in data mining which group related objects into a cluster. A number of clustering methods have been introduced in recent years. There are several types of clustering methods suitable for several types of applications which are k means, PAM, CLARA, CLARANS. In this paper k Means and PAM are compared. The analysis performed in three steps.

Step 1: Cluster analysis is performed with different clustering algorithm.

Step 2: Detect outliers by using the Mahalanobis distance.

Step 3: Comparison between different methods of clustering for detection of outliers was performed.

A Principal Component analysis is concerned with amplification the variance covariance construction a set of variables over a few linear combinations of these variables. A biplot allows evidence on both variable and sample of a data to be demonstrated graphically. Variables are displayed either as vectors whereas samples are displayed as points, linear axes or nonlinear trajectories. Biplots were described comprehensively by Gabriel (1971).

Partitions based clustering**:** In this method, k partitions of data are created with n data items. It is iterative procedure used to expand the clustering by moving up the items from one group to another. They are characterized by centroid or mediod in order to perform the cluster analysis (Aggrwal and Kaur, 2013).

There are three partitioning methods including K means,
PAM and CLARANS.

K means method**:** The method is self-possessed of these three steps:

Step 1. Partition the items into K primary clusters.

Step 2. Proceed over the list of items conveying an item to the cluster whose mean is closet.

Step 3. Recomputed the center for the clusters getting new entries and for clusters losing the entries.

Partitioning Around Medoid (PAM) method**:** This method is projected to find a sequence of objects called medoids that are centrally positioned in clusters. The aim of this method is to minimize the average dissimilarity of items to their closest selected item, minimize the sum of the dissimilarities between them.

Comparison of Clustering Algorithms**:** At first in data analysis PC performed then analyzed two clustering procedure. In PC analysis after plot the 1st two components find the outliers in data set. Then the result compare with the outlier detection in clustering procedures. Search for clusters graphically by plotting the o bservations. For p>2, plot the data in two dimensions using principal components or Biplots.

Data set**:** The secondary data was taken from the department of Zoology and Fisheries from the University of Agriculture, Faisalabad. In this data set, impact of application of zinc phosphide (three sub lethal doses) on different body tissues of rat was measured. The data was collected in four days and responses were taken as controls and after Appling the three treatments.

RESULTS AND DISCUSSION

The analysis was done in two steps. First, analyzed the data by PC analysis and then plotting first two principal components in order to detect the outliers

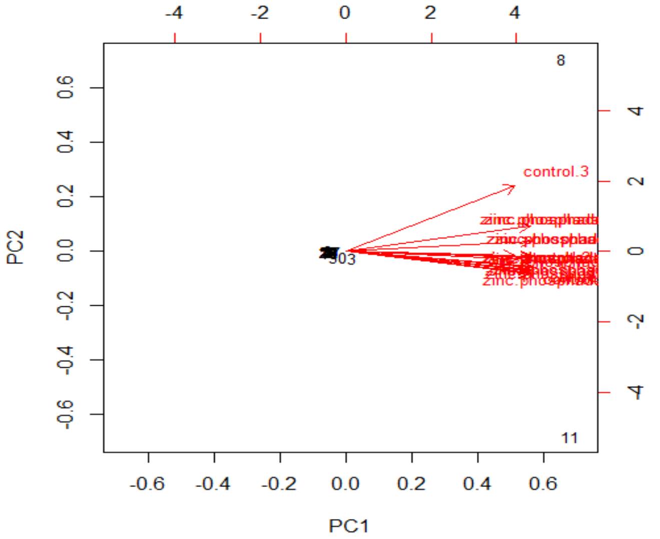

Figure 1. Biplot of first two PC's showing the outliers.

The principal component analysis was performed to construct the bi- plot with the first two pc's. The x-axis and y-axis show the first and second pc, respectively. The treatments and two outliers labeled with row numbers were shown by arrows. In order to choose k (number of clusters), first two PC's are plotted which show at least two clusters. The response on number 8 is Thrombocytes and response on number 11 is Monocytes.

K Mean Clustering: The method of k means clustering captured the two clusters of sized 28 and 3. Thrombocytes and Monocytes are found in one cluster and other cluster is based on 28 responses.

Within cluster sum of squares by cluster:

SSC_1= 9.935979

SSC_2 = 11.541572

Two aspects were considered to detect clusters with outlier(s): prefer within cluster sum of squares to be small and between clusters sum of squares to be large which means that there are two clusters considered as outlier in this data set.

In Figure 2, the cluster plot of k means showed two clusters in which first cluster is consisted of twenty-eight responses while the second has two responses.

In Figure 3, original observations, clusters and their centers are plotted which are showing the two structures of this data set. According to this plot, two original observations are far away from the others, hence two clusters with their centers are made.

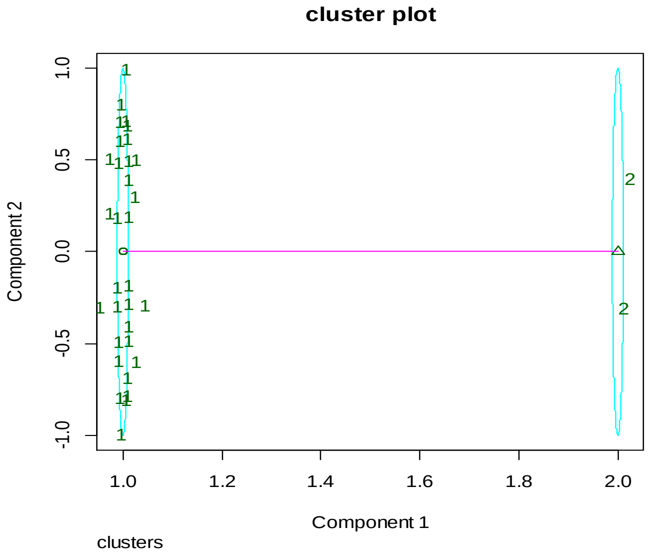

Figure 2. Cluster plot by K mean clustering.

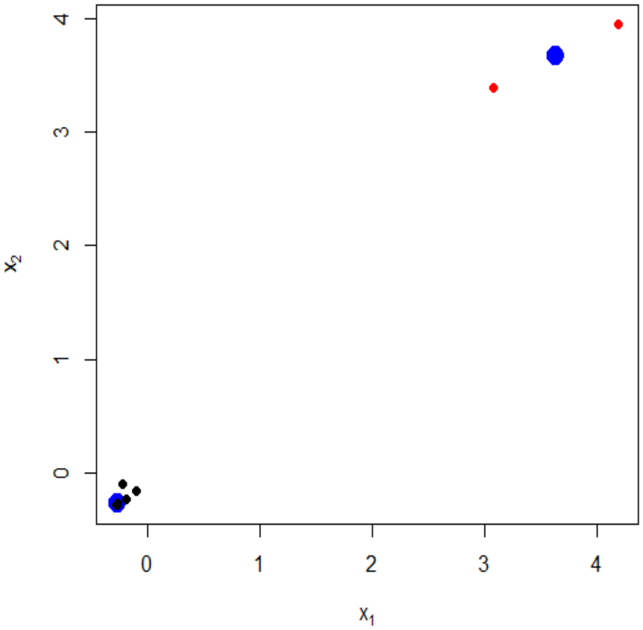

Figure 3. Plot of clusters centers and observations.

From Table 1, it is edident that the response Thrombocytes and Monocytes are outliers and hence the values of these two responses are presented here. The values in response Monocytes and Thrombocytes are compared to other values and found to be larger. Furthermore, these values are far away from other responses, therefore considered as outliers.

Partitioning around Medoids: In PAM method, Medoid or centers are taken while in k means method, means of the clusters are found for the analysis. Therefore, PAM method is more robust than k means method.

Table 1. Outliers detected by K Means method.

Outliers	Thrombocytes	Monocytes
control1	3.0781577	4.1950632
Zp10.025	3.3864518	3.9454996
Zp10.05	3.2181623	4.0896654
Zp10.75	3.8457123	3.5027209
Control2	3.0309167	4.2309581
Zp20.025	3.4434187	3.8967039
Zp20.05	3.1695279	4.0611055
Zp20.75	2.9532400	4.2003475
Control 3	3.0987140	3.3275365
Zp3 0.025	3.4726357	3.8283423
Zp3 0.05	3.1206750	4.1654916
Zp3 0.75	3.4629249	3.8822041
Control 4	4.8960594	1.7843407
Zp4 0.025	4.1936481	3.0818459
Zp4 0.05	4.1915518	3.0826189
Zp4 0.75	3.8409941	3.4651462
Control 5	3.0309167	4.2309581
Zp5 0.025	3.4726357	3.8283423
Zp5 0.05	4.1915518	3.0826189
Zp5 0.75	3.8457123	3.5027209

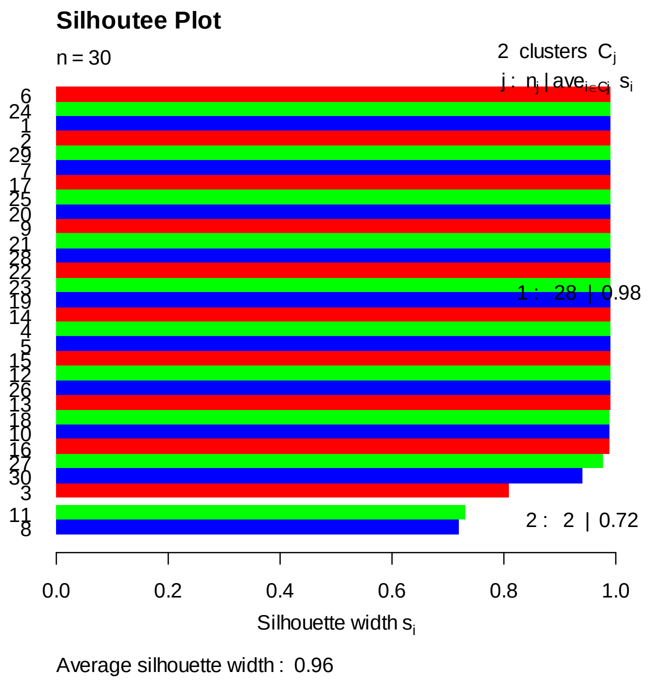

Figure 4. Silhouette plot PAM method.

Silhouette plot is used for a cluster analysis with two clusters. The plot expresses the silhouette values for the points in the two clusters. Average silhouette which is 0.96, which shows that mostly points in the first cluster have a large silhouette value, greater than 0.6, indicating that the cluster is somewhat separated from neighboring cluster. The second cluster covered two points with small silhouette standards showing the two clusters remain well disjointed.

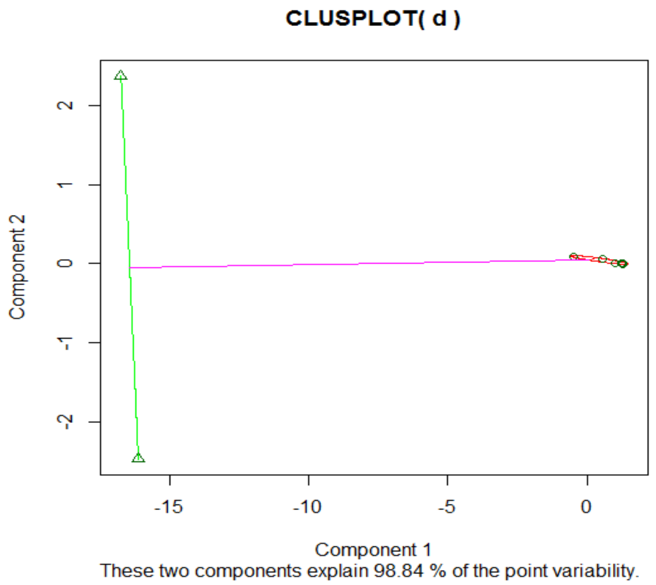

Figure 5. Two-dimensional projection of the cluster points for the dataset.

In Figure 5 Cluster plot was constructed by using the first two PC's in which most of the response lie in one side and some values on the opposite direction. There are only two responses namely, white blood cells and thrombocytes.

Table 2. Outliers detected by PAM method.

Outliers	white blood cells	Thrombocytes
control1	-0.2167138	3.0781577
Zp10.025	-0.0936212	3.3864518
Zp10.05	-0.1841858	3.2181623
Zp10.75	-0.1737184	3.8457123
Control2	-0.1011915	3.0309167
Zp20.025	0.1103377	3.4434187
Zp20.05	0.3878056	3.1695279
Zp20.75	0.4216116	2.9532400
Control 3	2.3359740	3.0987140
Zp3 0.025	0.2641611	3.4726357
Zp3 0.05	-0.1645280	3.1206750
Zp3 0.75	-0.1862012	3.4629249
Control 4	0.1662471	4.8960594
Zp4 0.025	0.1646166	4.1936481
Zp4 0.05	-0.0702787	4.1915518
Zp4 0.75	0.2579114	3.8409941
Control 5	-0.1011915	3.0309167
Zp5 0.025	0.2641611	3.4726357
Zp5 0.05	-0.0702787	4.1915518
Zp5 0.75	-0.1737184	3.8457123

***Outlier detected by PAM method*:** Outlier detected by PAM method are white blood cells and Thrombocytes. White blood cells and Thrombocytes are detected as outlier by PAM method and the values of these two responses are presented in Table 2. The values in response white blood cells are very small as compare to others and the values in response Thrombocytes are large than other. These values are far away from other response therefore these are outliers.

DISCUSSION

In this study, the principal component analysis was performed after the standardization of data and Biplot of first two PC's was constructed. For choosing the number of clusters, the Biplot shows that there may be at least two clusters. Furthermore, the Biplot states that the response Monocytes and Thrombocytes are not correlated to other responses. These two responses are significant from the other responses. For well clusters, prefer within cluster sum of squares to be small and between cluster SS to be large. In k mean method, after finding the distances of data and cluster mean. It shows that response Thrombocytes and Monocytes are outliers. Outliers detected by k means methods are same as showed in Biplot. Outlier detected by PAM method are white blood cells and Thrombocytes. These values are far away from other response therefore considered as outlier. By comparing the results, it was investigated that k means method is best for detection of outliers for this data set.

***Conclusion*:** In this study, firstly the data was clustered and secondly outliers were detected. At the end, the two clustering methods have been compared and the best method for outlier detection has been concluded. After comparing the two methods, it was concluded that k means method is better perform than other method. Thrombocytes and Monocytes are significantly differed to other responses and hence are outliers.

REFERENCES

Aggrwal, S. and P. Kaur. 2013. Survey of partition based clustering algorithm used for outlier detection. Int. J. Adv. Res. Eng. Tech. 1:57-62

Ayramo, S. and T. Karkkainen. 2006. Introduction to partitioning-based clustering methods with a robust example. Report., Department of Mathematical Information Technology Series C. Software and Computational Engineering.

Behera, H.S., A. Ghosh and S.K. Mishra. 2012. A new hybridized K-Means clustering based outlier detection technique for effective data mining. Int. J. Adv. Res. Comp. Sci. Soft. Eng. 2:287-292.

Ben-Gal, I. 2010. Outlier detection. P. 131-142 In. O. Maimon and L. Rockach (2nd ed.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Springer London

Chawla. S. and A. Goinis. 2013. K means: A unified approach to clustering and outlier detection. P. 189-197.In: Proceedings of the SIAM International Conference on Data Mining. USA.

Gabriel, K. 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58:453-467.

Hawkins, D.1980. Identification of Outliers. Chapman and Hall, London. P.188.

Johnson, R A and W.D Wichern. 2007. Applied Multivariate Statistical Analysis. Pearson Prentice Hall, USA.

Jhakumar, G. S. D. S. and B. J. Thomas. 2013. A new procedure of clustering based on multivariate outlier detection. J. of Data Sci. 11:69-84

Kaufman, L. and P. Rousseeuw. 1987. Clustering by means of Medoids. Statistical Data Analysis Based on the L_1-Norm and Related Methods. .405-416.

Kaur, N. and K. Kaur. 2013. Comparison between two approach based on threshold and entropy based approach. Int. J. Adv. Res. Comp. Sci. Soft. Eng. 3:1081-1086.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Mathematical Statistics and Probability; Berkeley, CA: Uni. of California Press, pp.281-297.

Sivaram, K. and D. Saveetha. 2013. An effective algorithm for outlier detection. Glob. J. Adv. Eng. Tech. 2:35-40.

Zhao, Y. 2013. Rand data mining: Examples and case studies. Academic Press,Netherlands..