

# Analyzing Outlier Detection Techniques with Hybrid Method

Shruti Aggarwal

Assistant Professor

Department of Computer Science and Engineering  
Sri Guru Granth Sahib World University. (SGGSWU)  
Fatehgarh Sahib, Punjab, India.

Janpreet Singh

M.Tech Research Scholar

Department of Computer Science and Engineering  
Sri Guru Granth Sahib World University. (SGGSWU)  
Fatehgarh Sahib, Punjab, India.

**Abstract:** Now day's Outlier Detection is used in various fields such as Credit Card Fraud Detection, Cyber-Intrusion Detection, Medical Anomaly Detection, and Data Mining etc. So to detect anomaly objects from various types of dataset Outlier Detection techniques are used, that detects and remove the anomaly objects from the dataset. Outliers are the containments that divert from the other objects. Outlier detection is used to make the data knowledgeable, and easy to understand. There are various outlier detection techniques used now day that detects and remove outliers from datasets. The proposed method is used to find outliers from the numerical dataset with the mean of Euclidean and Manhattan Distance.

**Keywords:** Data mining, Outlier Detection, K-Mean, Euclidean Distance, Manhattan Distance.

## I. INTRODUCTION

Data and Information or Knowledge has a significant role on human activities. Data mining is the knowledge discovery process by analyzing the large volumes of data from various perspectives and summarizing it into useful information. Due to the importance of extracting knowledge/information from the large data repositories, data mining has become an essential component in various fields of human life [1]. Mining knowledge from large amounts of spatial data is known as spatial data mining. It becomes a highly demanding field because huge amounts of spatial data have been collected by various applications. The databases can be clustered in many ways depending on the clustering algorithm employed, parameter settings used, and other factors. Multiple clustering can be combined so that the final partitioning of data provides better clustering [2].

### 1.1 Data Preprocessing

Preprocessing is the first step of Knowledge discovery. Data are normally preprocessed through data cleaning, data integration, data selection, and data transformation and prepared from the data warehouses and other information repositories [3].

### 1.2 Data Mining Functionalities

Data mining functionalities [3] are used to specify the kind of patterns to be found in data mining tasks. There are various types of databases and information repositories on which data mining can be performed. There are different data mining functionalities such as,

- Concept/Class Description: Characterization and Discrimination
- Classification and Prediction
- Cluster Analysis
- Evolution and Deviation Analysis
- Outlier Analysis.

## II. Literature Survey

### 2.1 Outlier Detection

Outliers are the objects that are not same as the other objects in the cluster or in the database. In other words outliers are the containment objects that divert from other objects or clusters in some way. Outlier Detection is used in various domains such as data mining. This has resulted in a huge and highly diverse literature of outlier detection techniques. A lot of techniques have been developed in order to solve the problems based on some of the particular features [4].

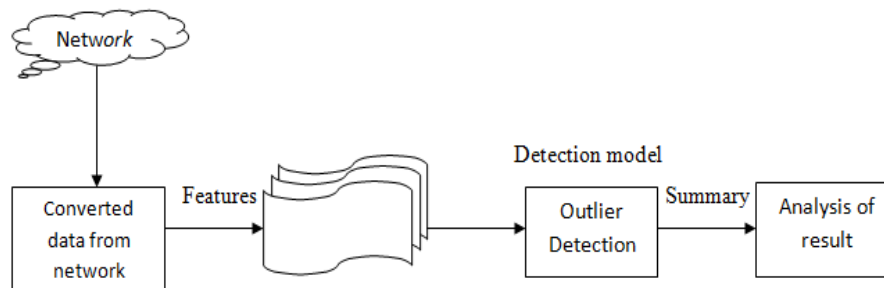


Fig. 2.1: Outlier Detection Process in Data Mining [5]

## 2.2 Partitioning Methods

Partitioning based technique is a well known method to cluster the dataset into group of similar objects. Some of the well known techniques are (A) Centroid - Based Technique and (B) Object - Based Technique. Both of them are mainly used techniques because they are very effective in finding the clusters.

### A. Centroid - Based Technique

Centroid-Based Technique mainly uses the K – Mean algorithm [6]. The algorithm takes the input parameter, K, and partitions a set of N objects into K clusters so that the resulting intra cluster similarity is high but the inter cluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid [3]. It is much simple and mostly used algorithm for finding clusters. So to find the cluster of the two or more data values or any two or more attributes it uses the addition of the two or more subtracted squared values until the value does not get mean of the all the values.

$$d = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2}$$

Equation of K- Mean Algorithm [3]

Example: K- mean algorithm is mostly used to cluster the dataset. let's suppose that a data set with many black points as shown in Figure 2.2 (A) and user wants to cluster the data into 3 parts then the value of K = 3, now the K – Mean clustering algorithm will partition the points into 3 parts, and each time it will update the mean value of the cluster and at last there are three clusters as shown in Figure 2.2(C).

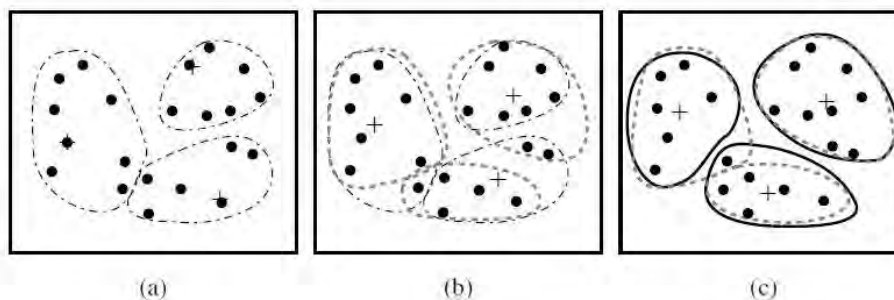


Fig. 2.2 Example of K- Means Clustering [3]

Steps by step process of K- Mean Algorithm

- Randomly select k data objects from dataset D
- Repeat step 3 and 4 until there is no change in the centre of clusters 3. Calculate the distance between each data object and all k cluster centers and assign to the closest cluster.
- Recalculate the cluster centre of each Cluster.

The k-means algorithm has the following important properties:

- It is efficient in processing large data sets.
- It often terminates at a local optimum
- It works only on numeric values.
- The clusters have convex shapes [7].

Khaled Alsabti et al [8] has shown effectiveness by explaining the efficiency of k –mean and using it with the different dataset and have also compared it by reducing the dimensionality of the data set and noted down the overall time taken by it.

S. D. Pachgade et al [9] have used the simple k – mean method to partition the data into group of different clusters and then get the threshold value from the user and find the outliers using the Euclidean distance method. Above the user defined threshold value data is an outlier; they also showed the result with the real data set that how efficient the method is in finding the outliers.

Hautamaki [10] have presented the new method named Outlier Removal Clustering that uses the clustering technique to find out the outlier by first removing the far vectors of data from the grouped dataset then analyzing the remaining clusters for outlier removal and recalculating the k – mean for grouping the vectors of the dataset, thus k – mean is used for efficient clustering techniques and for outlier removal method with Euclidean distance based methods.

#### *B. Object - Based Technique*

K – Medoid [11] is the well known partitioning based method. Instead of taking the mean value of the objects in a cluster, a reference point can be taken. To represent cluster, actual objects can be chosen from the clusters, using one representative object per cluster. Each remaining object is clustered with the representative object to which it is the most similar. And it can be seen from the K – Medoid equation [4] as show below that is the addition of the two subtracted points.

$$d(X_1, X_2) = \sum_{i=1}^K |X_1^i - X_2^i|$$

Equation of K- Medoid Algorithm [3]

Ali Al-Dahoud et al [38] has proposed a new method that uses the partitioning around medoid method with advance distance finding technique. In the proposed method they have firstly used k – medoid for finding cluster mean then removing those points that are far from the mean point then using the ADMP method for finding outlier from the rest of the clustered part of the dataset.

#### *2.3 Hierarchical Method*

A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach is also called the bottom-up approach that starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one (the topmost level of the hierarchy), or until a termination condition holds. The divisive approach is also called the top-down approach, and starts with all of the objects in the same cluster. In an each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds. Hierarchical methods suffer from the fact that once a step (merge or split) is done, it can never be undone [3].

#### *2.4 Density - Based Method*

Density based technique detect the cluster of the high density and they also to detect the cluster of the low density, and on the bases of the density the outliers are find. Density based technique is also used for high dimensional data set, as well as for low to medium dimensional data set, density based algorithms such as DBSCAN, CLIQUE and DENCLUE have shown to find clusters of different sizes and shapes, although not of different densities.

Density based method works on assumptions that the density around a normal data object is similar to the density around its neighbors while the density around an outlier is considerably different to the density around its neighbors. The density-based approach compares the density around a point with the density around its local neighbors by computing an outlier score. Thus Raghuvira Pratap et al [2] have proposed an improved density based method with k- Medoid technique to detect the outlier and improved the method over the DBSCAN method.

### **III. Proposed Hybrid Method**

Firstly K – mean clustering algorithm partitions a dataset D into K number of clusters. The objective of the algorithm is to minimize the sum of distances from the observations in the clusters to their cluster means. Suppose we have a dataset D with n observations  $\{X_1, X_2, \dots, X_n\}$ . Assume that we have partitioned D into k clusters  $\{C_1 \dots C_k\}$  with the corresponding means  $m_i, i \in [1, k]$ .

The k - mean algorithm consists of two steps: a clustering step and an update step. Initially, k distinct cluster means are selected randomly. In the clustering step, the k - mean algorithm clusters the dataset using the current k cluster means. A cluster  $C_i$  is a set that consists of all the observations  $X_j$ . The k-mean algorithm runs multiple times until the means converge. However, the number of clusters depends on the parameter k.

Input: Data set

**Step 1:** Randomly select k data object from array of dataset D.

**Step 2:** Repeat step 3 and 4 till no new cluster centers are found or it reaches to the maximum limit of the iteration where the max count value is being set.

**Step 3:** Calculate the distance with the mean of the Euclidean and Manhattan between each data object and all k cluster centers and assign data object to the nearest cluster.

**Step 4:** recalculate the cluster center of each cluster.

**Step 5:** Calculate the distance of each data points of a cluster and the k cluster centers with mean of Euclidean and Manhattan Distance.

**Step 6:** Assign that point to a new array that contains the outliers of all the k clusters.

**Step 7:** Repeat the Steps 5 and 6 till no new outlier is founded or until the distance criteria met.

**Step 8:** Calculate the mean of all data point of outliers detected from the each cluster.

**Step 9:** Calculate the distance of each outlier with mean of outliers.

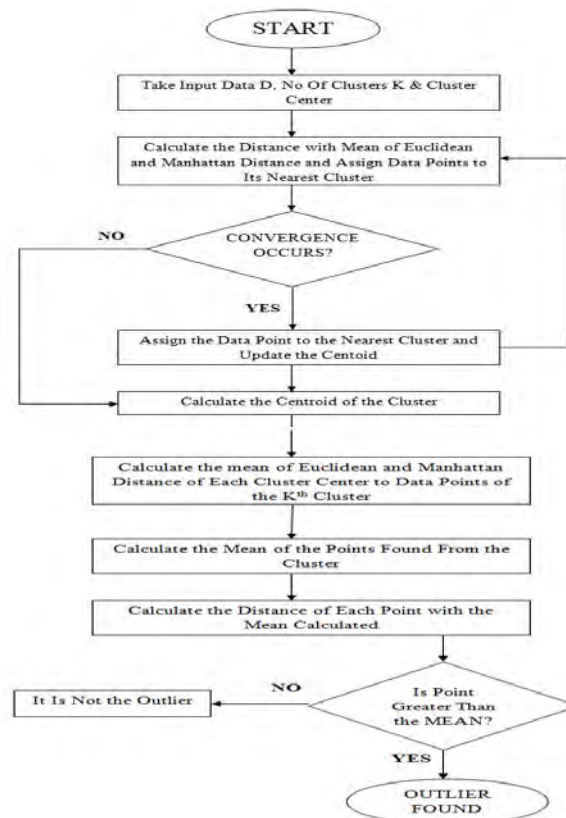
**Step 10:** If the calculated distance is less than the mean then it is not an outlier.

**Step 11:** Else, the data point will be considered as real outlier.

Simply these clusters are represented by their centroids. A cluster centroid is typically the mean of the points in the cluster. The mean (centroid) of each cluster is then computed so as to update the cluster center. This update occurs as a result of the change in the membership of each cluster.

The processes of re-assigning the input vectors and the update of the cluster centers is repeated until no more change in the value of any of the cluster centers. Also, it is possible that the k-means algorithm won't find a final solution. In this case it would be a good idea to consider stopping the algorithm after a pre-chosen maximum of iterations.

The K Means clustering method can be considered, as the cunning method because here, to obtain a better result the centroids, are kept as far as possible from each other. This algorithm is simple to implement and run, relatively fast, easy to adapt, and common in practice.



Flow Diagram of Hybrid Method

#### IV. Conclusions

The new approached method provides efficient outlier detection and data clustering capabilities in the presence of outliers. The proposed method first finds out the user defined number of clusters with the mean of Euclidean plus Manhattan Distance then outlier are detected from each cluster.

- A. Hybrid method can cluster the data according to user need and find the outliers that differ from the other data in the dataset.
- B. The approached method can be widely used with the multidimensional datasets.
- C. This approach makes those two problems solvable for less time, using the same process and functionality for both clustering and outlier identification.

#### V. Future Scopes

Several interesting areas of future research have opened up from the work described. In the future, more modifications can be made to the proposed method such as dimension reduction, reducing the number of iterations etc.

- A. Hybrid method can be further extended to make less number of Iterations.
- B. It can be used with the multidimensional alphanumeric dataset.
- C. It can also be extended with different methods to partition the dataset more efficiently into group of clusters.

#### VI. References

- [1] Venkatadri.M, Dr. Lokanatha C. Reddy "A Review on Data mining from Past to the Future" International Journal of Computer Applications (0975 –8887) Volume 15– No.7, February 2011.
- [2] Raghuvira Pratap, K Suvarna, J Rama Devi, Dr.K Nageswara Rao "An Efficient Density based Improved K- Medoids Clustering algorithm " International Journal of Advanced Computer Science and Applications, Volume 2, No. 6, 2011.
- [3] J. Han and M. Kamber. "Data Mining: Concepts and Techniques" (2nd Ed.). Morgan Kaufmann, San Francisco, CA, 2006
- [4] H.S.Behera, AbhishekGhosh, SipakkuMishra "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.
- [5] T1-Svetlana Cherednichenko (2005)"Outlier Detection in Clustering" University of Joensuu 2005.
- [6] T. Soni Madhulatha "An Overview On Clustering Methods" IOSR Journal of Engineering Apr. 2012, Volume 2(4) pp. 719-725.
- [7] M.Vijayalakshmi, M.Renuka Devi "A Survey of different issue of different clustering algorithms used in large data sets" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 3, March 2012.
- [8] Khaled Alsabti, Sanjay Ranka, Vineet Singh"An Efficient K-Means Clustering Algorithm".
- [9] S. D. Pachgade, S. S. Dhande "Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, June 2012.
- [10] Ville Hautamaki, Svetlana Cherednichenko, Ismo Karkkainen, Tomi Kinnunen, and Pasi Franti " Improving K-Means by Outlier Removal" Springer-Verlag Berlin Heidelberg 2005 LNCS 3540, pp. 978–987, 2005.
- [11] T. Soni Madhulatha "An Overview On Clustering Methods" IOSR Journal of Engineering Apr. 2012, Volume 2(4) pp. 719-725.