INTRODUCTION

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

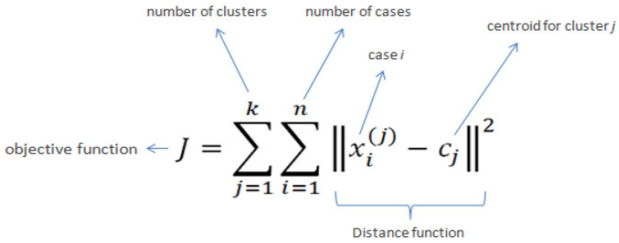
Outlier detection (also known as anomaly detection) is the process of finding data objects with behaviors that are very different from expectation. Such objects are called outliers or anomalies. Outlier detection is important in many applications in addition to fraud detection such as medical care, public safety and security, industry damage detection, image processing, sensor/video network surveillance, and intrusion detection. Outlier detection and clustering analysis are two highly related tasks. Clustering finds the majority patterns in a data set and organizes the data accordingly, whereas outlier detection tries to capture those exceptional cases that deviate substantially from the majority patterns. Outlier detection and clustering analysis serve different purposes.

Cluster analysis should be defined as the process of separating a set of patterns into clusters such that members of one cluster are similar. The goal of such partitioning , or clustering , may be to gain an insight into some structure inherence in the population or to develop a business strategy that is customized to each cluster customer for higher business efficiency. There are many algorithms which is used to form cluster such as kmeans, leader based clustering algorithm.

k-means clustering algorithm

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem.

k-means clustering intends to partition *n* objects into *k* clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly *k* different clusters of greatest possible distinction. The best number of clusters *k* leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:



k-means Algorithm

1. **Clusters the data into *k* groups where *k* is predefined.**
2. **Select *k* points at random as cluster centers.**
3. **Assign objects to their closest cluster center according to the *Euclidean distance* function.**
4. **Calculate the centroid or mean of all objects in each cluster.**
5. **Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.**

Our Approach

In this project, we have modified the well known k-means clustering algorithm and used it for outlier detection.

Outliers can be found from k-means by calculating the average radius for each of the k-clusters. We can then draw circles by taking each of the k-

clusters centre as centre and their corresponding avg. radius as radius. The points lying outside of these circles, can then be classified as outliers.

Our idea is perform k-means, (k+1)-means and (k-1)-means clustering process and then we classify the common set of outliers obtained separately from the three processes as the “better” outliers.

Also, our process of getting from k-means to (k+1) and (k-1)-means adds to the process of finding “better” outliers without addiing significant time to k-means.

k,k+1,k-1-means Algorithm

1. **Clusters the data into *k* groups where *k* is predefined.**
2. **Select *k* points at random as cluster centers.**
3. **Assign objects to their closest cluster center according to the *Euclidean distance* function.**
4. **Calculate the mean of all objects in each cluster.**
5. **Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.**
6. **Calculate average radius of the k clusters, form circles by taking each of the k-clusters centre as centre and their corresponding avg. radius as radius. The points outside this circle are outliers set 1 (o1).**
7. **Calculate the mean of the k cluster centers and assign this mean and the rest of the k cluster centres as initial cluster centers for k+1-means.**
8. **Repeat step 5 for k=k+1.**
9. **Repeat step 6 with k=k+1 for getting outliers set 2(o2).**
10. **Now, drop the cluster with least number of points from k-clusters obtained after step 5. Assign rest k-1 centers as initial cluster center for (k-1)-means.**
11. **Repeat step 5 for k=k-1.**
12. **Repeat step 6 with k=k-1 for getting outliers set 3(o3).**
13. **Take the common subset of o1, o2 and o3. Sort it based on their distance from the centre of their corresponding clusters in k-means.These are required outliers.**