Simultaneous K-Means and Outlier Detection

Software Design Specification
(Session 2016-2017)

Guided By:

Dr. G.L. Prajapati

Submitted By:

Nitin Mittal (13C7032)

Pulkit Maloo (13C7037)

Tanay Kothari(13C7054)

Department of Computer Engineering
Institute of Engineering & Technology
Devi Ahilya Vishwavidyalaya, Indore (M.P.)
(www.iet.dauniv.ac.in)
October - 2016

Recommendation

The software design specification entitled **“Simultaneous K-means & Outlier Detection”** submitted by **Pulkit Maloo, Nitin Mittal** & **Tanay Kothari** is a satisfactory account of the bonafide work done for Project Phase I of project work and is recommended for approval.

Date:

Dr. G.L. Prajapati
Project Guide

Software Design Specification Approval Sheet

The software design specification entitled **"Simultaneous K-means & Outlier Detection"** submitted by **Pulkit Maloo, Nitin Mittal** & **Tanay Kothari** is approved for Project Phase I of Project work.

Internal Examiner	**External Examiner**
Date:	Date:

INDEX

Table of Contents Page No.

1. Introduction……………………………………………………………………………5

    1.1 Overview…………………………………………………………………………5

    1.2 Problem Definition………………………………………………………………….5

    1.3 Proposed Solution……………………………………………………………...5

2. Literature Survey…………………………………………………………………………6

    2.1 Unsupervised Learning……………………………………………………………...6

    2.2 Clustering……………………………………………………………………………6

    2.3 K-Means……………………………………………………………………………..7

    2.4 Anomaly Detection/Outlier Detection………………………………………………8

3. Related Work……………………………………………………………………………10
4. Proposed Solution……………………………………………………………………….11
5. Conclusion………………………………………………………………………………12

1. INTRODUCTION

1.1 OVERVIEW

Outlier is defined as a noisy observation, which does not fit to the assumed model that generated the data. In clustering, outliers are considered as observations that should be removed in order to make clustering more reliable. Some clustering algorithms such as the k-means algorithm is extremely sensitive to outliers, and such outliers may have a disproportionate impact on the final cluster configuration. The result of outliers in the clusters also leads to distorted centroids positions of these clusters.

Despite their close complementarity, clustering and anomaly detection are often treated as separate problems in the data-mining community. This distinction is not without justification. Often applications are defined in terms of outliers (like in fraud detection, network anomaly detection, etc.) in which case a direct approach is likely to be more efficient.

1.2 PROBLEM DEFINITION

The main concern of clustering-based outlier detection algorithms is to find clusters and outliers, which are often regarded as noise that should be removed in order to make more reliable clustering. Some noisy points may be far away from the data points, whereas the others may be close. The faraway noisy points would affect the result more significantly because they are more different from the data points. It is desirable to identify and remove the outliers, which are far away from all the other points in cluster.

1.3 PROPOSED SOLUTION

So, to improve the clustering process of such algorithms, we propose a clustering-based technique to produce data clusters and simultaneously remove outliers. Proposed outlier detection process at the same time is effective for extracting clusters and very efficient in finding outliers. K-means is an iterative clustering algorithm widely used in pattern recognition and data mining for finding statistical structures in data. We propose to integrate outlier removal into K-means clustering. This method employs both clustering and outlier discovery to improve estimation of the centroids of the generative distribution.as well as keeping in mind performance of this method compared to the standard K-means algorithm.

2. LITERATURE SURVEY

2.1 UNSUPERVISED LEARNING

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.

The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data. The clusters are modeled using a measure of similarity which is defined upon metrics such as Euclidean or probabilistic distance.

2.2 CLUSTERING

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

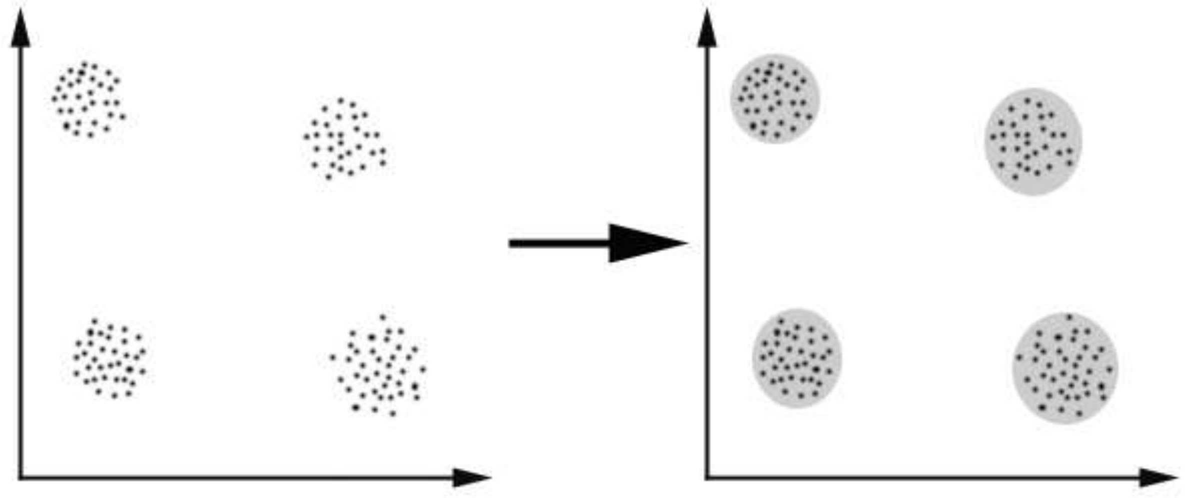

Fig. 1

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called distance-based clustering.

Clustering problems arise in many different applications, such as data mining and knowledge discovery, data compression and vector quantization, and pattern recognition and pattern classification. The notion of what constitutes a good cluster depends on the application and there are many methods for finding clusters subject to various criteria, both ad hoc and systematic.

2.3 K-MEANS

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because different locations cause different results. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words, centroids do not move any more.

Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2 ,$$

where $\left\|x_i^{(j)} - c_j\right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center c_j, is an indicator of the distance of the *n* data points from their respective cluster centers.
The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.

3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

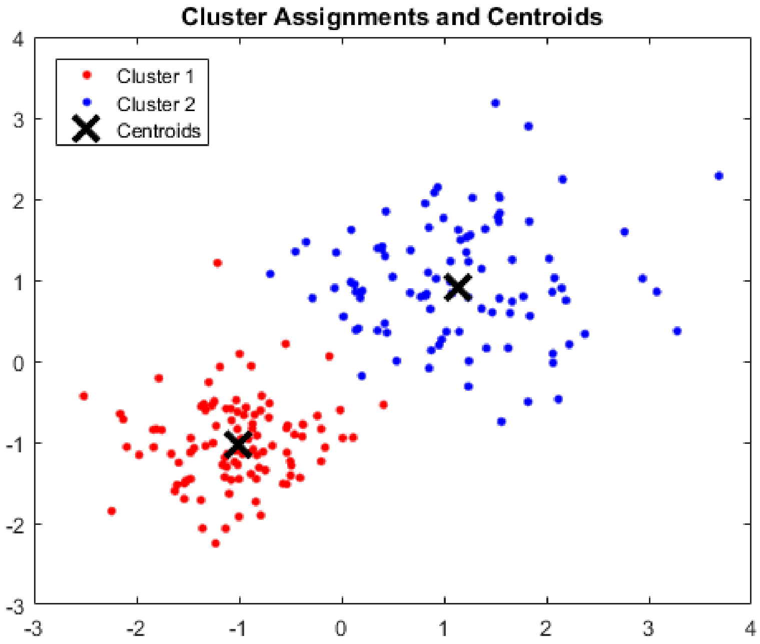

Fig. 2

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centers. The k-means algorithm can be run multiple times to reduce this effect.

2.4 ANOMALY DETECTION / OUTLIER DETECTION

In data mining, ‘anomaly detection’ (also outlier detection) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset. Typically, the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions.

Anomaly detection is applicable in a variety of domains, such as intrusion detection, fraud detection, fault detection, system health monitoring, event detection in sensor networks, and detecting Eco-system disturbances. It is often used in preprocessing to remove anomalous data from the dataset.

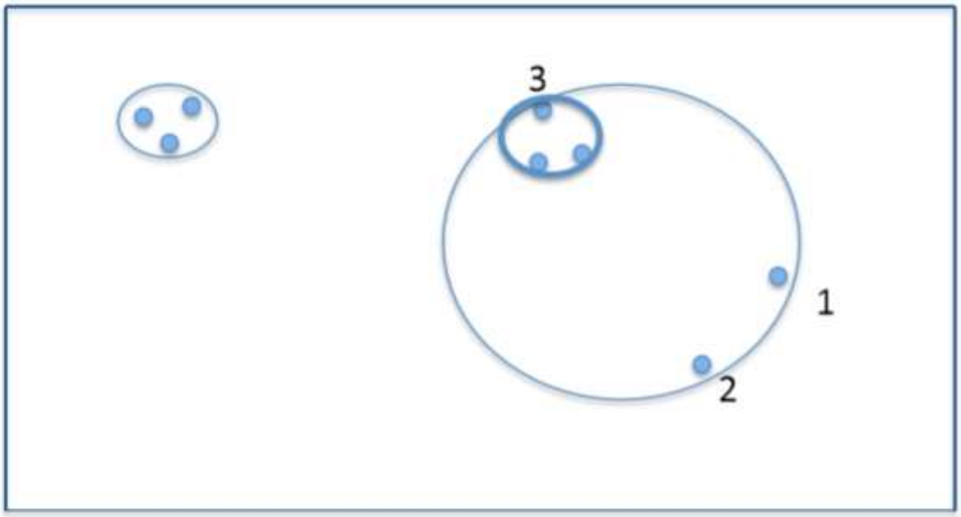

Fig. 3

The k-means algorithm is extremely sensitive to outliers. By removing two points (1) and (2), we can obtain much tighter clusters (the bold circle near 3).

Outlier detection is a deeply researched problem in both communities of statistics and data mining — but with different perspectives. In data mining, Knorr and Ng proposed a definition of distance-based outlier, which is free of any distributional assumptions and is generalizable to multidimensional datasets. Intuitively, outliers are data points that are far away from their nearest neighbors. However, the outliers detected by these methods are global outliers, i.e., the “outlierness” is with respect to the whole dataset. Breunig et al. have argued that in some situations local outliers are more important than global outliers and cannot be easily detected by standard distance-based techniques.

3. RELATED WORK

In this section we present a solution by Sanjay Chawla for clustering data with outliers. We now define the problem of clustering with outliers. We consider two parameters: k the number of clusters, and ℓ the number of outliers. As in the k-means problem our goal is to find a set of k cluster center points C, which can be subset of X or points in the underlying space. Additionally we aim at identifying a set L ⊆ X of outliers and we require |L| = ℓ and minimizing the error E(X, C, L) = E(X\L, C).

<u>Algorithm –</u>

```
**Input:** Set of points
    $X = \{\mathbf{x}_{1,\dots,}\mathbf{x}_n\}$
    A distance function $d: X \times X \to \mathsf{R}$
    Numbers $k$ and $\ell$
**Output:** A set of $k$ cluster centers $C$
    A set of outliers, $L \subseteq X$
$C_0 \leftarrow \{k$ random points of $X\}$
$i \leftarrow 1$
**while** (no convergence achieved) **do**
    Compute $d(\mathbf{x} \mid C_{i-1})$, for all $\mathbf{x} \in X$
    Re-order the points in $X$ such that $d(\mathbf{x}_1 \mid C_{i-1}) \geq \dots \geq d(\mathbf{x}_n \mid C_{i-1})$
    $L_i \leftarrow \{\mathbf{x}_1,\dots,\mathbf{x}_{`}\}$
    $X_i \leftarrow X \setminus L_i = \{\mathbf{x}_{`+1,\dots,,}\mathbf{x}_n\}$
    **for** $(j \in \{1,\dots,k\})$ **do**
        $P_j \leftarrow \{\mathbf{x} \in X_i \mid c(\mathbf{x} \mid C_{i-1}) = \mathbf{c}_{i-1,j}\}$
        $\mathbf{c}_{i,j} \leftarrow \mathrm{mean}(P_j)$
    $C_i \leftarrow \{\mathbf{c}_{i,1},\dots,\mathbf{c}_{i,k}\}$
    $i \leftarrow i + 1$
```

4. PROPOSED SOLUTION

We propose a method for simultaneous outlier detection and clustering. Our proposed method consists of two consecutive stages, which are repeated several times. In the first stage, we perform K-means algorithm until convergence, and in the second stage, we assign an outlying-ness factor for each vector. Here we introduce a relative distance factor termed as “Outlier Factor". Its value lies between 0 & 1. It defines outlying-ness of each data point in cluster and depends on distance of point from the cluster centroid. Then algorithm iterations start, with first finding the data points with maximum distance to the partition centroid.

$$\text{Outlier Factor} = \frac{\text{Distance of a point from cluster centroid}}{d_{max}}$$

We introduce a threshold value. Threshold is used to determine the points that will be regarded as outliers and removed from the datasets. Act as a bounding value for outlier factors. Points with outlier factor greater then threshold are treated as outliers and thus removed from datasets.

Algorithm –

```
1. Run K-means until convergence, pick best Cluster Centroids C
2. For no. of iterations
        2.1 For every Cluster Ci
                a. Calculate outlier factor for each point
                b. Remove all data points with Outlier Factor > Threshold, and update the dataset
        2.2 Run K-Means over the updated dataset and update C
```

5. CONCLUSION

Clustering and outlier detection are often treated as separate problems. However, both these problems are tightly coupled. For example, outliers can have a disproportionate impact on the shape of clusters which in turn can mask obvious outliers. The proposed method was also compared with the standard K-means without outlier removal, and a simple approach in which outlier removal precedes the actual clustering. The proposed method employs both clustering and outlier discovery to improve estimation of the centroids of the generative distribution. The advantages of combining clustering and outlier selection include:

1. The resulting clusters tend to be compact and semantically coherent
2. The clusters are more robust against data perturbations; and
3. The outliers are contextualized by the clusters and more interpretable.