

A Review Paper on Comparison of Clustering Algorithms based on Outliers

Shivanjli Jain

Research Scholar

*Department of Computer Science & Engineering
Punjab Technical University Baba Banda Singh Bahadur
Engineering College, Fatehgarh Sahib, Punjab, India*

Amanjot Kaur

Assistant Professor

*Department of Computer Science & Engineering
Punjab Technical University Baba Banda Singh Bahadur
Engineering College, Fatehgarh Sahib, Punjab, India*

Abstract

Data mining, in general, deals with the discovery of non-trivial, hidden and interesting knowledge from different types of data. With the development of information technologies, the number of databases, as well as their dimension and complexity, grow rapidly. It is necessary what we need automated analysis of great amount of information. The analysis results are then used for making a decision by a human or program. One of the basic problems of data mining is the outlier detection. The outlier detection problem in some cases is similar to the classification problem. For example, 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. In this thesis, the ability to detect outliers can be improved using a combined perspective from outlier detection and cluster identification. In proposed work comparison of four methods will be done like K-Mean, k-Medoids, Iterative k-Mean and density based method. Unlike the traditional clustering-based methods, the proposed algorithm provides much efficient outlier detection and data clustering capabilities in the presence of outliers, so comparison has been made. The purpose of our method is not only to produce data clustering but at the same time to find outliers from the resulting clusters. The goal is to model an unknown nonlinear function based on observed input-output pairs. The whole simulation of this proposed work has been taken in MATLAB environment.

Keywords: Data Mining, Clustering, Outlier, Data Mining Methods

I. INTRODUCTION

Data mining also known as knowledge-discovery in databases (KDD) is process of extracting potentially useful information from raw data. A software engine can scan large amounts of data and automatically report interesting patterns without requiring human intervention [10, 11]. Other knowledge discovery technologies are Statistical Analysis, OLAP, Data Visualization, and Ad hoc queries. Unlike these technologies, data mining does not require a human to ask specific questions.

Here is the list of areas where data mining is widely used:

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

In general, Data mining has four major relationships. They are:

- Clustering.
- Classification.
- Association.
- Sequential Pattern.

II. CLUSTERING

Clustering can be said as identification of similar classes of objects [10]. Clustering is a data mining technique of grouping set of data objects into multiple groups or clusters so that objects within the cluster have high similarity, but are very dissimilar to objects in the other clusters. Maximizing intra-class similarity & minimizing inter-class similarity as shown in figure 1.

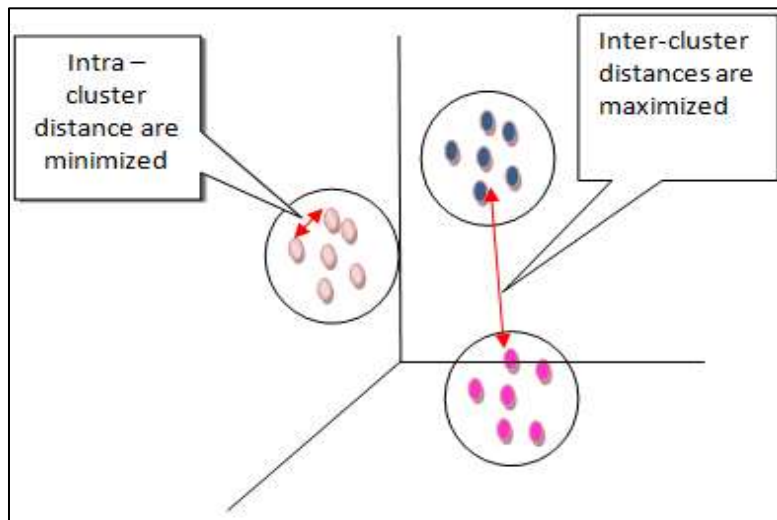


Fig. 1: Clustering Principle

III. CLUSTERING METHODS

The clustering methods can be classified into following categories:

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

A. K-Means Clustering

It is a partition method technique which finds mutual exclusive clusters of spherical shape. K-Means algorithm organizes objects into k – partitions where each partition represents a cluster [8]. We start out with initial set of means and classify cases based on their distances to their centers. Next, we compute the cluster means again, using the cases that are assigned to the clusters; then, we reclassify all cases based on the new set of means. We keep repeating this step until cluster means don't change between successive steps. Finally, we calculate the means of cluster once again and assign the cases to their permanent clusters.

- Input: n data points and the number of cluster (K)
- Output: K clusters
 - 1) Initialize the K cluster centers
 - 2) while termination condition is not satisfied do
 - Determine the distance of each object to the centroids
 - Group the object based on minimum distance (find the closest centroid)
 - 3) end while

1) Disadvantage of K-mean

- Randomly choosing of the cluster center cannot lead us to the fruitful result.
- Applicable only when mean is defined i.e. fails for categorical data.
- Unable to handle noisy data and outliers.
- Algorithm fails for non-linear data set.
- Work only on numeric dataset.

B. K-Medoids

To reduce the complexities of K-means and to improve the detection rate, K-Medoids algorithm has been proposed. The K-Medoids algorithm basically divide the objects into k different clusters having same features. The K-Medoids algorithm is more robust to noise as well as to the outlier detection. A medoid is mainly the dataset which has minute dissimilarity. The algorithm is described as below [9].

1) Begin

- Obtain features of high priority than less priority from cluster size.
- Take medoids and use Euclidean distance to measure the dissimilarity between the clusters. After this sort out the clusters in an ascending order.
- Map each object with medoid that has close value and also find the optimal value from large number of objects

- Exchange the current medoid with the medoid that has minimum value of dissimilarity.
- Again Exchange the current medoid with the medoid that has minimum value of dissimilarity. But if the value is same as the previous then algorithm will be stopped otherwise repeat step 4.

2) End

C. DBSCAN

Density based clustering algorithm has played a vital role in finding non linear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm. It uses the concept of density reachability and density connectivity [10].

- Density Reachability - A point "p" is said to be density reachable from a point "q" if point "p" is within ϵ distance from point "q" and "q" has sufficient number of points in its neighbors which are within distance ϵ .
- Density Connectivity - A point "p" and "q" are said to be density connected if there exist a point "r" which has sufficient number of points in its neighbors and both the points "p" and "q" are within the ϵ distance. This is chaining process. So, if "q" is neighbor of "r", "r" is neighbor of "s", "s" is neighbor of "t" which in turn is neighbor of "p" implies that "q" is neighbor of "p".

Let $X = \{x_1, x_2, x_3, \dots, x_n\}$ be the set of data points. DBSCAN requires two parameters: ϵ (eps) and the minimum number of points required to form a cluster (minPts).

- Start with an arbitrary starting point that has not been visited.
- Extract the neighborhood of this point using ϵ (All points which are within the ϵ distance are neighborhood).
- If there are sufficient neighborhoods around this point then clustering process starts and point is marked as visited else this point is labeled as noise (Later this point can become the part of the cluster).
- If a point is found to be a part of the cluster then its ϵ neighborhood is also the part of the cluster and the above procedure from step 2 is repeated for all ϵ neighborhood points. This is repeated until all points in the cluster is determined.
- A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.
- This process continues until all points are marked as visited.

D. Iterative K means Algorithm

The k -means approach to clustering performs an iterative alternating fitting process to form the number of specified clusters. The k -means method first selects a set of n points called cluster seeds as a first guess of the means of the clusters. Each observation is assigned to the nearest seed to form a set of temporary clusters. The seeds are then replaced by the cluster means, the points are reassigned, and the process continues until no further changes occur in the clusters. When the clustering process is finished, you see tables showing brief summaries of the clusters. The k -means approach is a special case of a general approach called the EM algorithm; E stands for Expectation (the cluster means in this case), and M stands for maximization, which means assigning points to closest clusters in this case.

The k -means method is intended for use with larger data tables, from approximately 200 to 100,000 observations. With smaller data tables, the results can be highly sensitive to the order of the observations in the data table.

K-Means clustering only supports numeric columns. K-Means clustering ignores model types (nominal and ordinal), and treat all numeric columns as continuous columns.

IV. OUTLIER

Very often, there exist data objects that do not comply with the general behaviour or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining data, are called outliers.

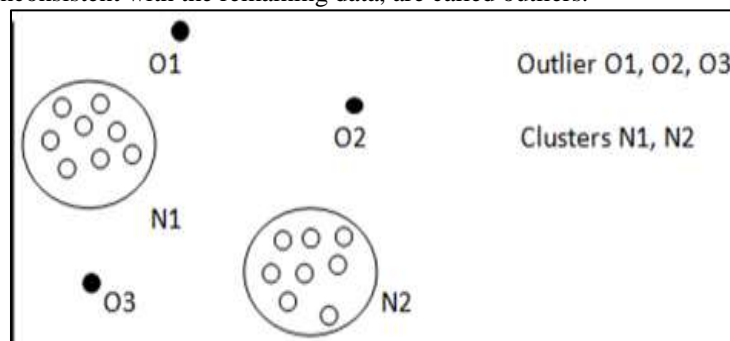


Fig. 2: Outliers

Outliers are patterns in data that do not conform to a well-defined notion of normal behavior. Figure 1.2 shows outliers in a simple 2-dimensional data set. The data has two normal regions, N1 and N2, since most observations lie in these two regions. Those points that are sufficiently far away from the regions, e.g., point's O1 and O2, and O3, are outliers [9]. Outliers might be

induced in the data for a variety of reasons, such as malicious activity, e.g., credit card fraud, cyber-intrusion, terrorist activity or breakdown of a system, but the common point of all is that they are interesting to the analyst. The “interestingness” or real life relevance of outliers is a key feature of outlier detection.

V. LITERATURE SURVEY

In this section, we will give a brief overview on a few representative works.

Hans et.al [1] propose a novel outlier detection model to find outliers that deviate from the generating mechanisms of normal instances by considering combinations of different subsets of attributes, as they occur when there are local correlations in the data set. This model enables to search for outliers in arbitrarily oriented subspaces of the original feature space. Author show how in addition to an outlier score, their model also derives an explanation of the outlierness that is useful in investigating the results. Their experiments suggest that their novel method can find different outliers than existing work and can be seen as a complement of those approaches.

Shu wu et.al [2] propose a formal definition of outliers and an optimization model of outlier detection, via a new concept of holoentropy that takes both entropy and total correlation into consideration. Based on this model, author define a function for the outlier factor of an object which is solely determined by the object itself and can be updated efficiently. author propose two practical 1-parameter outlier detection methods, named ITB-SS and ITB-SP, which require no user-defined parameters for deciding whether an object is an outlier. Users need only provide the number of outliers they want to detect. Experimental results show that ITB-SS and ITB-SP are more effective and efficient than mainstream methods and can be used to deal with both large and high-dimensional data sets where existing algorithms fail.

Hermine et.al [3] proposed the combination of two statistical techniques for the detection and imputation of outliers in time series data. An autoregressive integrated moving average with exogenous inputs (ARIMAX) model is used to extract the characteristics of the time series and to find the residuals. The outliers are detected by performing hypothesis testing on the extrema of the residuals and the anomalous data are imputed using another ARIMAX model. The process is performed in an iterative way because at the beginning the process, the residuals are contaminated by the anomalies and therefore, the ARIMAX model needs to be re-learned on “cleaner” data at every step. author test the algorithm using both synthetic and real data sets and author present the analysis and comments on those results.

Haizau et.al [4] propose a robust method for robust local outlier detection with statistical parameters, which incorporates the clustering based ideas in dealing with big data. Firstly, This method find some density peaks of dataset by 3s standard. Secondly each remaining data object in the dataset is assigned to the same cluster as its nearest neighbor of higher density. Finally, author use Chebyshevs inequality and density peak reachability to identify local outliers of each group. The experimental results demonstrate the efficiency and accuracy of the proposed method in identifying both global and local outliers, Moreover, the method also proved more robust analysis than typical outlier detection methods, such as LOF and DBSCAN.

Mihiri et.al [5] propose a Betweenness Centrality (BEC) as novel to determine the outlier in network analyses. The Betweenness Centrality of a vertex in a graph is a measure for the participation of the vertex in the shortest paths in the graph. The Betweenness centrality is widely used in network analyses. Especially in a social network, the recursive computation of the betweenness centralities of vertices is performed for the community detection and finding the influential user in the network. In this paper, author propose that this method is efficient in finding outlier in social network analyses. Furthermore author show the effectiveness of the new methods using the experiments data.

Saif et.al [6] presented that Data Mining is all about data analysis techniques. It is useful for extracting hidden and interesting patterns from large datasets. Clustering techniques are important when it comes to extracting knowledge from large amount of spatial data collected from various applications including GIS, satellite images, X-ray crystallography, remote sensing and environmental assessment and planning etc. To extract useful pattern from these complex data sources several popular spatial data clustering techniques have been proposed. DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a pioneer density based algorithm. It can discover clusters of any arbitrary shape and size in databases containing even noise and outliers. DBSCAN however are known to have a number of problems such as: (a) it requires user's input to specify parameter values for executing the algorithm; (b) it is prone to dilemma in deciding meaningful clusters from datasets with varying densities; (c) and it incurs certain computational complexity. Many researchers attempted to enhance the basic DBSCAN algorithm, in order to overcome these drawbacks, such as VDBSCAN, FDBSCAN, DD_DBSCAN, and IDBSCAN. In this study, author survey over different variations of DBSCAN algorithms that were proposed so far. These variations are critically evaluated and their limitations are also listed.

Huan et.al [7] proposed a method based on DBSCAN data clustering algorithm to stabilize the jitter of digital video with moving objects in it. In order to distinguish the corners on moving objects with those on background, after extracting the corners of each frame, DBSCAN algorithm was used to cluster all the corners by clustering their motion vectors' lengths and directions. Then author compared the dispersion of each cluster to confirm whether the corners in each cluster were belonged to moving objects or background. Simulation experimental results showed that the proposed method had good stabilization effects to stabilize jitter in video sequence with moving objects in it.

VI. METHODOLOGY

To start with, different clustering algorithms are chosen from data mining and implemented in a programming language. MATLAB will be the implementing language here. Each algorithm will be tested with dataset as input data. I am applying different- different clustering algorithms after detection of outliers and compare the results using metrics.

VII. CONCLUSION

There are lots of methods for detecting the outlier in data mining. Every one mostly focuses on trying different-different techniques to detect the outlier for better results. The proposed methods focus on detection of outlier and then compare them. In proposed work four algorithms i.e K-mean, K-medoid, DBSCAN and Iterative K-Mean has been compared based on proposed dataset that contains only text document.

REFERENCES

- [1] H. P. Kriegel, P. Kröger, E. Schubert and A. Zimek, "Outlier Detection in Arbitrarily Oriented Subspaces," 2012 IEEE 12th International Conference on Data Mining, Brussels, 2012, pp. 379-388. doi: 10.1109/ICDM.2012.21.
- [2] S. Wu and S. Wang, "Information-Theoretic Outlier Detection for Large-Scale Categorical Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 3, pp. 589-602, March 2013. doi: 10.1109/TKDE.2011.261.
- [3] H. N. Akouemo and R. J. Povinelli, "Time series outlier detection and imputation," 2014 IEEE PES General Meeting | Conference & Exposition, National Harbor, MD, 2014, pp. 1-5. doi: 10.1109/PESGM.2014.6939802.
- [4] H. Du, S. zhao and D. zhang, "Robust Local Outlier Detection," 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, 2015, pp. 116-123. doi: 10.1109/ICDMW.2015.114.
- [5] H. B. M. Shashikala, R. George and K. A. Shujae, "Outlier detection in network data using the Betweenness Centrality," SoutheastCon 2015, Fort Lauderdale, FL, 2015, pp. 1-5. doi: 10.1109/SECON.2015.7133008.
- [6] S. U. Rehman, S. Asghar, S. Fong and S. Sarasvady, "DBSCAN: Past, present and future," Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference on the, Bangalore, 2014, pp. 232-238.
- [7] Huan Yu and Wenhui Zhang, "DBSCAN data clustering algorithm for video stabilizing system," Mechatronic Sciences, Electric Engineering and Computer (MEC), Proceedings 2013 International Conference on, Shengyang, 2013, pp. 1297-1301.
- [8] Parimala M. et al., "A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases", International Journal of Advanced Science and Technology, Vol. 31, June, 2011.
- [9] Pratap R. et al., "An Efficient Density based Improved K- Medoids Clustering algorithm", International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.
- [10] Rajagopal Dr. Sankar, "CUSTOMER DATA CLUSTERING USING DATA MINING TECHNIQUE", International Journal of Database Management Systems Vol.3, No.4, November 2011.
- [11] Ramageri B. et al., "DATA MINING TECHNIQUES AND APPLICATIONS", Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305.