International Journal of Database Theory and Application
Vol.9, No.5 (2016), pp.275-282
http://dx.doi.org/10.14257/ijdta.2016.9.5.28

Filtered Clustering Based on Local Outlier Factor in Data Mining

[1]Vishal Bhatt, [2]Mradul Dhakar and [3]Brijesh Kumar Chaurasia

[1,2,3]Deptt. of Computer Science & Engg.
ITM University Gwalior, India
er.vishalbhatt@gmail.com, mraduliitm@gmail.com, bkchaurasia.itm@gmail.com

Abstract

In this paper, the impact of $\boldsymbol{k}$-means and local outliner factor on data set is studied. Outlier is the observation which is different from or inconsistent with the rest of the data. However, the main challenges of outlier detection are increasing complexity due to variety of datasets and size of dataset. To evaluate the outlierness and catch similar outliers as a group are also issues of this technique. The concept of $\boldsymbol{LOF}$ (Local Outlier Factor) is presented in this work. The paper describes comparative study of five different methodologies using $\boldsymbol{K}$-means as the base algorithm along with the various distances method used in finding the dissimilarities between the objects hence to analyze the effects of the outliers on the cluster analysis of dataset in data mining.

***Keywords**: Outlier; K-means algorithm; LOF*

1. Introduction

Data mining in general, process of exploration and analysis of large amount and different varieties of data in order to discover meaningful patterns and rules [1]. Due to the present technological advancement, large amount of data is generated which is communicated or stored over the web. But still there is a requirement to generate or dug out the information out of the huge amount of data available on the web. The other sources of data available are social sites, media, web pages *etc.* In short abundance of data is available. Though abundance of data, but still lagging in information. Information which can be utilized for various purposes like: decision making, marketing, future interpretation, customer behavior and so on. The field which deals in the extraction of knowledge from the available data is known as data mining. Various techniques are available in data mining for the extraction of knowledge. These are classification, prediction, estimation, association or affinity grouping, description and visualization and clustering [2]. In clustering, the cluster is organized in such a manner that the objects with similar properties lies in a single cluster or in other words the objects in two different clusters may possess different properties. The various types of clustering are: partition based, hierarchical based, density based, grid based and model based. Under classical partitioning method we have k-means clustering method. In k-means clustering method, k represents the number of clusters. All the objects in a given dataset are assigned to these clusters. Clustering helps in finding out outliers. The outlier detection is one of the basic problems of data mining addresses in this paper. An outlier is an observation of the data that deviates from other observations so much that it arouses suspicions that it was generated by a different and unusual mechanism [3-4]. In general, outliers are not considered to be of great use. But as it is known "One's garbage can is of other's use". Detection of outlier is helpful in numerous applications as network intrusion detection, credit card frauds, activity monitoring, financial applications, voting irregularity analysis, severe weather prediction, and geographic information systems *etc.* [5-6]. Different methods are available for outlier detection such as statistical distribution based, distance based, density based and deviation based. Statistical and distance based method detects

ISSN: 2005-4270 IJDTA

the global outliers *i.e.* either an object is outlier or not where as density based method detects local outlier *i.e.* how much an object is outlier with respect to its neighbor. The extension of k-means clustering in terms of accuracy and efficacy is presented in [7]. In this proposed work, the methods are used here finds the better initial centroids along with provide an efficient way for assigning data points to appropriate clusters with reduced time complexity. The Euclidean distance is used to find the distance between the centroids and data point to produce good clusters in less amount of computational time. The initial centroids algorithm is presented in [15]. The proposed algorithm selects k objects randomly from the given data set as the initial centroids. However, accuracy of output by the standard k-means algorithm can be affected if different initial values are given for the centroids. The issue is addressed in [8] the initial centroids calculated systematically. The initial centroids algorithm based on k-means that have avoided alternative randomness of initial center is presented in [9]. To find better initial cluster centers for k means cluster and hierarchical algorithms is presented in [10]. In this proposed method all the clustering results of k-means are utilized in certain times. After that, the results are transformed by combining with hierarchical algorithm so as to find the better initial cluster centres for k-means clustering algorithm.

2. Outliner Detection Technique

An outlier is one that appears to deviate markedly from other members of the sample in which it occurs [11]. In [12], an outlier as an observation in a data set which appears to be inconsistent with the remainder of that set of data. An outlier detection technique can be divided between parametric method and non-parametric method. Figure 1 is depicted by taxonomy of outlier detection technique.

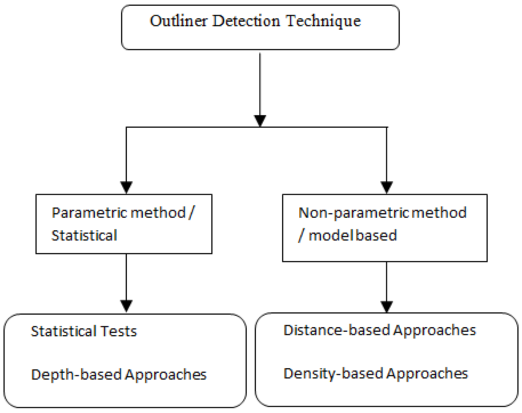
Outliner Detection Technique
Parametric method / Statistical
Non-parametric method / model based
Statistical Tests
Depth-based Approaches
Distance-based Approaches
Density-based Approaches

Figure 1. Outlier Detection Techniques

The parametric or statistical method may also subcategorize into statistical test approach, depth based approaches and deviation based approaches.

Similarly, the non-parametric or model based or spatial proximity method may also subcategorize into distance-based approaches, density-based approaches and high dimensional approaches.

A. Statistical Approach:

This approach is based on numeric values. LRT (Linear Regression Technique) and CCT (Control Chart Technique) [16] are under the categories of statistical approach.

B. Distance-Based Approach:

In this approach, the objects are being classified as outlier based on their distance with rest of the objects in a given dataset. One of such approach is given in paper [17] where author propose the Class Outlier Factor (COF) which measures the degree of being a Class Outlier for a data object.

C. Density-Based Approach:

The density-based approach estimates the density distribution of the data and identifies outliers as those lying in low-density regions. There are several approaches based on density. Some of these approaches are LOF, LOF', LOF'', MDV [18],[19], [20] and [21].

D. Space-Based Approach:

This approach is based on the concept of spatial and non-spatial attributes. This approach is well discussed in the paper [22] where author propose two spatial outlier detection methods which integrate the impact of spatial properties to the outlierness measurement.

3. Proposed Approach

The proposed approach addresses the concept of filtered clustering to enhance the quality of clusters been formed. In the filtered clustering we have use the simple k-means algorithm. The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the intercluster similarity is low [4]. The algorithm of proposed approach is explained below:

Algorithm:

The k-means algorithm for partitioning, where each clusters center is represented by the mean value of the objects in the cluster.

```
Input:
    1. $k$: the number of clusters
    2. $D$: a dataset containing $n$ objects.
Output: A set of $k$ clusters

Method:
    1. arbitrarily choose $k$ objects from $D$ as the initial cluster centers;
    2. repeat
    3. (re)assign each object to the cluster to which the object is the most similar,
       based on the mean value of the objects in the cluster;
    4. update the cluster means, i.e., calculate the mean value of the objects for
       each cluster;
    5. until no change;
```

The filtered used in k-means algorithm is local outlier factor (LOF). There are several situations where we have to define the degree of outlierness of an object in a given database. This degree of outlierness is well expressed in [17]. They named this degree of outlierness as LOF.

Necessary definitions to explain LOF concept are as:

Definition 1: (k-distance of an object p): This is the distance between object p and its k^{th} nearest neighbour.

Definition 2: (k-distance neighborhood of an object p): Given the $k-distance(p)$, the $Nk(p)$ contains every object whose distance from p is not greater than the $k-distance$.

Definition 3: (reachability distance of p with respect to object o): This is the maximum distance out of $k-distance$ of an object o and real distance between p and o.

Definition 4: (local reachability density of an object p): The local reachability density of an object p can be calculated dividing one by the average reachability distance based on the $MinPts$-distance neighborhood of p.

Definition 5: (local outlier factor of an object p): The LOF of p is defined as

$$LOF_{Minpts}(p) = \frac{\sum_{o\in N_{Minpts}(p)} \frac{lrd_{Minpts}(o)}{lrd_{Minpts}(p)}}{|lrd_{Minpts}(p)|}$$

The different variants used for finding distance are: Euclidean distance, Manhattan distance and Minkowski distance. The definition of these three is as follows [4]:

Euclidean:

$$d(i,j) = \sqrt{(x_{i_1}-x_{j_1})^2 \quad (x_{i_2}-x_{j_2})^2 + \cdots .. + \quad (x_{i_n}-x_{j_n})^2}$$

Where $i = (x_{i_1}, x_{i_2}, \ldots., x_{i_n})$ and $j = (x_{j_1}, x_{j_2}, \ldots., x_{j_n})$ are two n-dimensional data objects.

Manhattan:

$$d(i,j) = |x_{i_1}-x_{j_1}| + |x_{i_2}-x_{j_2}| + \cdots + |x_{i_n}-x_{j_n}|$$

Minkowski:

$$d(i,j) = \left(|x_{i_1}-x_{j_1}|^p + |x_{i_2}-x_{j_2}|^p + \cdots + |x_{i_n}-x_{j_n}|^p\right)^{1/p}$$

Where p is a positive integer.

4. Results and Discussion

We investigate the effect of $k - mean$ with filtered parameters and without filtered parameters. The implementation of hybrid LOF is done using Weka tool. Weka is a collection of machine learning algorithms for data mining tasks [13]. To verify the efficacy of the LOF, the dataset with 12 attributes and 600 instances [14] is used. However, application of such proposed hybrid scheme on huge data with high dimensionality is the future scope. After running the Weka tool on given data set following results are achieved:

Table 1. Number of Iterations Taken by the Methodology

Methodology Used	Number of Iterations
SimpleKMeans without filtered parameter *i.e.* LOF	4
SimpleKMeans with filtered parameter LOF and Euclidean distance	4
SimpleKMeans with filtered parameter LOF and Chebyshev distance	4
SimpleKMeans with filtered parameter LOF and Manhattan distance	9
SimpleKMeans with filtered parameter LOF and Minkowski distance	5

From Table 1 it is observed that the number of iterations taken by Simple K-Means with filtered parameter LOF and Manhattan distance to form the final cluster is much higher with respect to rest of the methodology used. However, the performance in terms of iteration LOF without filtered parameter and Euclidean distance and Chebyshev distance is same.

Table 2. Sum of Squared Errors for the LOF Schemes

Methodology Used	Sum of squared errors
SimpleKMeans without filtered parameter *i.e.* LOF	2048.5301065012936
SimpleKMeans with filtered parameter LOF and Euclidean distance	2067.8933053687697
SimpleKMeans with filtered parameter LOF and Chebyshev distance	2048.5301065012936
SimpleKMeans with	2223.274614128641

filtered parameter LOF and Manhattan distance	
SimpleKMeans with filtered parameter LOF and Minkowski distance	2093.943983269058

Table 2 shows the sum of squared errors for the different methodology used. The sum of squared errors is defined as [4]:

$$E = \sum_{i=1}^{k} \sum_{p \epsilon c_i} |p - m_i|^2$$

Where E is the sum of the square error for all objects in the dataset; p is the point in space representing a given object; and m_i is the mean of cluster c_i (both p and m_i are multidimensional). In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This criterion tries to make the resulting k clusters as compact and as separate as possible. Obviously from the Table 2, we can observed that the sum of squared errors is less for Simple K-Means without filtered parameter *i.e.* LOF and Simple K-Means with filtered parameter LOF and Chebyshev distance.

Table 3. LOF for the Methodologies Used

Methodology Used	LOF
SimpleKMeans without filtered parameter *i.e.* LOF	NA
SimpleKMeans with filtered parameter LOF and Euclidean distance	1.0357
SimpleKMeans with filtered parameter LOF and Chebyshev distance	1
SimpleKMeans with filtered parameter LOF and Manhattan distance	1.0481
SimpleKMeans with filtered parameter LOF and Minkowski distance	1.0318

LOF Indicates the local outlierness of an object with respect to rest of object in the given dataset. The Table 3 reflects the local outlierness of given transactions with each other. So, the LOF for Simple K-Means with filtered parameter LOF and Manhattan distance is higher with respect to the rest of the methodology used.

Table 4. Execution Time for Various Methodologies

Methodology Used	Time taken to build Model
SimpleKMeans without filtered parameter *i.e.* LOF	NA
SimpleKMeans with filtered parameter LOF and Euclidean distance	0.96 seconds
SimpleKMeans with filtered parameter LOF and Chebyshev distance	0.23 seconds
SimpleKMeans with filtered parameter LOF and Manhattan distance	0.91 seconds
SimpleKMeans with filtered parameter LOF and Minkowski distance	1.15 seconds

Table 4 indicates the execution time consumed by different methodology for creating the clusters. Results show Simple K-Means with filtered parameter LOF and Chebyshev distance took the least time.

Table 5. Incorrectly Clustered Instances for the LOF Scheme

Methodology Used	Incorrectly clustered instances	
		Percentage
SimpleKMeans without filtered parameter *i.e.* LOF	29.0	4.8333 %
SimpleKMeans with filtered parameter LOF and Euclidean distance	28.0	4.6667 %
SimpleKMeans with filtered parameter LOF and Chebyshev distance	23.0	3.8333 %
SimpleKMeans with filtered parameter LOF and Manhattan distance	27.0	4.5%
SimpleKMeans with filtered parameter LOF and Minkowski distance	26.0	4.3333 %

Table 5 shows the number of instances that are incorrectly clustered by the above different methodology. It is evidently by the results that the Simple K-Means with filtered parameter LOF and Chebyshev distance is well suited for clustering in data mining.

5. Conclusion

The paper introduced and examined the impact of $k - means$ with and without filtered parameters. The work followed by the creation of clusters on a given bank dataset. Afterword, it was found that Simple K-Means with filtered parameter LOF and Chebyshev distance is creating the clusters in a more efficient manner since the number of iterations, sum of squared errors, LOF , time taken to build model for this methodology is less in compare to other methodologies been used for the above experiment. The number of incorrectly clustered instances for Simple K-Means with filtered parameter LOF and Chebyshev distance is also less with respect to other methodology. Finally, hybrid LOF is suited to create clusters in datamining and needs to be studied and evaluated in a data with high dimensionality.

References

[1] M. S. Chen, J. Han and P. S. Yu, “Data mining: an overview from a database perspective,” IEEE Transactions on Knowledge and Data Engineering, vol. 08, no. 6, **(1996)**, pp. 866-883.

[2] M. J. A. Berry and G. S. Linoff, “Mastering Data Mining – The art and science of customer relationship management”, John Wiley & Sons Inc, U. K., **(2000)**.

[3] S. Hawkins, H. X. He, G. J. Williams and R. A. Baxter, “Outlier detection using replicator neural networks”, In Proceedings of the Fourth International Conference and Data Warehousing and Knowledge Discovery (DaWaK02), vol. 2425, **(2002)**, pp.17-180.

[4] D. Hawkins, “Identification of Outliers”, Chapman and Hall, London – New York, **(1980)**.

[5] E. Acuna and C. A. Rodriguez, “Meta analysis study of outlier detection methods in classification”, Technical paper, In proceedings IPSI 2004, **(2004)**, pp. 1-25.

[6] K. I. Penny and I. T. Jolliffe, “A comparison of multivariate outlier detection methods for clinical laboratory safety data”, The Statistician, vol. 50, no. 3, **(2001)**, pp. 295-308.

[7] K. A. Abdul Nazeer and M. P. Seba Stian, “Improving the accuracy and efficiency of the k-means clustering algorithm”, In International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE-2009), vol. 1, **(2009)**, pp. 1-5.

[8] F. Yuan, Z. H. Meng, H. X. Zhangz and C. R. Dong, “A New Algorithm to Get the Initial Centroids”, In proceedings of the 3rd International Conference on Machine Learning and Cybernetics, vol. 2, **(2004)**, pp. 1191-1193.

[9] C. Zhang and Shixiong Xia, “K-means Clustering Algorithm with Improved Initial center”, In Second International Workshop on Knowledge Discovery and Data Mining (WKDD), **(2009)**, pp. 790-792.

[10] K. Arai and A. R. Barakbah, “Hirerachical K-means: an algorithm for Centroids intialization for k-means”, Reports of the Faculty of Science and Engineering, Saga University, vol. 36, no.1, **(2007)**.

[11] V. Barnett and T. Lewis, “Outliers in Statistical Data”, John Wiley, **(1994)**.

[12] R. Johnson, “Applied Multivariate Statistical Analysis”, Prentice Hall, **(1992)**.

[13] Oneline available at: http://www.cs.waikato.ac.nz/ml/weka/

[14] http://facweb.cs.depaul.edu/mobasher/classes/ect584/WEKA/k-means .html

[15] M. H. Nabil and K. M. Saad, “Class Outliers Mining: Distance-Based Approach”, International Journal of Computer, Electrical, Automation, Control and Information Engineering, vol. 1, no. 9, **(2007)**, pp. 2792-2805.

[16] Z. Bakar, R. Mohemad, A. Ahmad and M. Deris, “A Comparative Study for Outlier Detection Techniques in Data Mining”, In Proeeding IEEE conference oncybernetics and Intelligent Systems, **(2006)**, pp 1-6.

[17] M. M. Breunig, H.-P. Kriegel, R. T. Ng and J. Sander, “LOF:Identifying density-based local outliers”, Proceeding ACM SIGMOD International Conference on Management of Data, **(2000)**, pp. 93-104.

[18] A. Chiu and A. Fu, “Enhancements on Local Outlier Detection”, In Proceeding the seventh International Database Engineering and Applications Symposium (IDEAS’03), **(2003)**, pp. 298-307.

[19] K. G. Sharma, A. Ram and Y. P. Singh, “Efficient Denity Based Outlier Handling Technique in Data Mining”, In Proceeding 1st International Conference on Computer science and Information Technology, CCSIT, Part 1, **(2011)**, pp. 542-550.

[20] D. Jain, P. Khatri, R. Soni, and B. K. Chaurasia, “Hiding Sensitive Association Rules without Altering the Support of Sensitive Item(s)”, The Third International Conference on Wireless & Mobile Networks (WiMoNe-2012), N. Meghanathan et al. (Eds.): CCSIT 2012, Part I, LNICST 84, **(2012)**, pp. 500-509.

[21] M. M. Breunig, H. P. Kriegel, R. T. Ng and J. Sander, “LOF:Identifying density-based local outliers”, In Proceeding ACM SIGMOD International Conference on Management of Data, **(2000)**, pp. 93-104.

[22] Y. Kou, C. T. Lu and D. Chen, “Spatial Weighted Outlier Detection”, In Proceedings of SIAM Conference on Data Mining, **(2006)**, pp. 613-617.