International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
Available at http://inpressco.com/category/ijcet

Research Article

Local Deviation Coefficient based Outlier detection for Scattered Data

Ms.Snehal Rajendra More and Prof.Shirish S Sane

Department of Computer Engineering K K Wagh College Of Engineering,Nasik.

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

Outlier detection technique is applied in variety of domains like intrusion detection, health care monitoring, human gait analysis, etc. There are 2 main types of outliers: Global and local. Global outliers are the extreme data values in a dataset whereas local outliers are the data points within a range but much less or higher than other dataset values. Lot of work has been done in the domain of outlier detection. LOF, LOF with incremental approach, Memory efficient LOF with streaming data are well known outlier detection techniques. Existing approaches focuses on finding of the degree of deviation of local outlier points from the clustered data and failed to find degree of dispersion. The proposed system focuses on finding local outliers on scattered data. For outlier detection, degree of deviation and degree of dispersion is calculated. To make the system memory efficient, the dimensionality reduction and undersampling is applied. For dimension reduction principal component analysis is applied. A rough clustering based on multi-level queries algorithm is applied for safe non-outliers points' elimination. This is undersampling method. This reduces the number of points for further processing. A density-based local outlier detection for scattered data (E2DLOS) algorithm is applied on non-safe points and top m outlier are identified. The system will be tested on various datasets downloaded from UCI repository.

Keywords: *Outlier detection, rough clustering, scattered data, PCA, undersampling, feature extraction, dimension reduction*

Introduction

Outlier points are unexpected behavioral points in dataset. The outlier points create a special subset of overall data with significant distinct behavior. The subset size is very small as compared to the overall dataset. There are 2 main types of outliers: Global and local. Global outliers are the extreme data values in a dataset whereas local outliers are the data points within a range but much less or higher than other dataset values. Form large amount of generated data such special informative, unexpected points are extracted in outlier detection process. The outlier detection process has high importance in data mining domain. This technique is useful for researcher and scientist for detailed data analysis. Outlier detection technique is applied in variety of domains like intrusion detection, health care monitoring, human gait analysis, etc. Generally outliers are the by-product of clustering algorithm. The points which are far away from cluster centroid or far away from its nearest neighbors are treated as outliers. Initially whole clustering process is get executed and then along with the cluster result outlier points are extracted. To remove the dependency of outlier detection technique over clustering algorithm, some new techniques are proposed. These techniques work on the efficiency improvement of outlier detection process.

The initial outlier detection techniques works on finding only global outliers from whole dataset. But in real world scenario, the structure of data and user needs changes the focus of outlier detection technique at local level. The real time generated data is always incomplete in terms of time and space. As compared with the global outlier detection technique, the local outlier detection technique only compared the data points with subset of data i.e. with its nearest neighbor and with the entire dataset. The local outliers factor (LOF) is widely used local outlier detection technique. Based on the basic LOF new techniques are proposed to deal with real time requirements. The iLOF is extended version of LOF for streaming data analysis. MiLOF is the memory efficient local outlier detection technique over streaming data. The efficiency of execution is improved at two level:

1. Finding local neighbors of test object.
2. Comparison of test object with its neighbors.

These algorithms generate good results on regular dataset. These techniques focus on finding the object deviation from other data entries but do not consider the degree of dispersion of dataset points. These techniques failed to generate accurate outlier points set over scattered structured dataset objects.

For outlier detection process, nearest neighbors needs to be identified. The nearest neighbor of each point is identified from complete dataset. This is time consuming process. For nearest neighbor identification the system complexity raises to O(n^2). This affects the system efficiency and infeasible for large volume dataset. In this research work, a new outlier detection system is proposed which mainly deals with two problems:

1. Find outlier over scatter data
2. Improve efficiency of outlier detection process without hampering the accuracy of detection process.

For scatter data handling a degree of dispersion is checked for every suspicions point. For efficiency improvement, dimensionality reduction and undersampling techniques are used. A principal component analysis(PCA) is a feature extraction technique that reduces the dimensional space by projecting n dimensional data to k dimensional space where k< n. The undersampling technique removes the safe datapoints and generates a subset of data. The outlier detection process works on subset of data and hence reduces the processing overhead. The Rough Clustering based on MultiLevel Queries (RCMLQ) algorithm. This algorithm removes the safe points in a dataset and generates the suspicions candidate outlier list.

Following section II includes the related work done in the domain of outlier detection followed by problem formulation in section III. Based on the problem formulation a new system details are mention in section IV. Section V includes result and discussion and Section VI concludes the paper.

Related Work

Breuing et al. [2] proposes a concept of local outlier. Based on this concept, Local outlier Factor LOF algorithm is proposed. This is distance based outlier detection technique. Local outlier Factor value is calculated with the help of nearest neighbor search, k-distance, reachable distance and reachable density. The LOF is useful for finding outlier local outliers in dataset with uneven density distribution. The LOF algorithm has following limitation:

1. The system do not generate accurate results for oulier detection over scatter data.
2. The LOF works only on numerical datasets.
3. The efficiency of algorithm is depending on threshold value of k for KNN.

M-tree [3], R-tree [4] are the techniques for efficient knearest neighbor search. These techniques are applied in LOF for finding nearest neighbors to improve efficiency of LOF algorithm.

Zhang et al. [5] proposes a local distance-based outlier Factor (LDOF) based on the local outlier factor. In LDOF algorithm initially average distance of each point with its k nearest neighbor is identified then average distance among all nearest neighbor is identified. Then the ratio of these two values is called as outlier factor. The complexity of system is O(k^2) because distance between each pair need to be calculated.

Latecki et al. [6] proposes some alteration in LOF algorithm. The distance between each point is replaced by variable-width Gaussian kernel density estimation (KDE). This is a density estimate. Same as LOF, local density factor LDF is introduced. The complexity of this system is similar to the LOF system.

Schubert et al. [7] proposes a kernel density estimation outlier score (KDEOS) algorithm. This technique also focuses on improvement of LOF algorithm using KDE. This technique uses mathematical properties of KDE. This is density based outlier detection technique and uses the normal cumulative density function to calculate KDE. This is applicable on datasets with normal distribution. It is not applicable for all datasets. The system complexity is O(n*k*dk) where dk = kmax − kmin + 1.

Kriegel et al. [8] proposes an Angle-Based Outlier Detection ABOD algorithm and Fast ABOD algorithm for high dimensional datasets. From each data point, its nearest neighbor and angle between point and its neighbor is identified. The system uses weighted variance technique for local outliers' identification. The complexity of the algorithm is O(k^2).

S. Papadimitriou , et. al.[9] proposes a multi-granularity deviation factor(MDEF) algorithm . This algorithm tries to find isolated outliers as well as outlying clusters. This algorithm does not require any user defined threshold value. This technique deals with local density and multiple granularity. This technique is not applicable for scatter dataset.

Mahsa et.al. [10] proposes a combined distance and density based approach for local outlier detection over streaming data. The system mainly works in 3 phases summarization merging and revised insertion. For efficiency improvement a data summarization is performed in terms of clustering. Cluster centroids points are preserved as a representative of data for further steaming data processing. This is a memory efficient technique and can be applied on systems with low configurations.

Most of the existing approaches focus on finding degree of deviation and not the degree of dispersion. Subn Su, et. al.[1] proposes a technique that simultaneously focus on degree of deviation and degree of dispersion. A new Local Deviation Coefficient (LDC) is proposed. The system mainly proposes efficient local outlier detection algorithm (E2DLOS). This method is best suited for outlier detection over scatter data. For efficiency improvement it uses rough clustering algorithm. This algorithm reduces the number of samples for processing. This clustering approach is used as a preprocessing step and this algorithm is work like undersampling technique. The system only focuses on sample reduction and not the dimension reduction.

Table 1: System Comparison

		LOF [2]	**MiLO F[10]**	**LDO F[5]**	**KDE OS[7]**	**E2DL OS[1]**
Outlier Detecti on Strateg ies	1.Density based Detection	YES	YES	-	YES	YES
	2. Distance Based Detection	-	YES	YES	-	-
Silent Feature s	Outlier 1. 1.Detectio n over Structured dataset	YES	YES	YES	YES	YES
	2.Outlier detection over scatter dataset	-	-	-	-	YES
	3.Clusteri ng	YES	YES	YES	YES	YES
	4.Data summariz ation and filtering for Efficiency Improvem ent				YES	YES

Problem Formulation

Lot of work has been done in the domain of outlier detection. The outlier detection is treated as by-product of clustering technique. This hampers the efficiency of outlier detection process. Many techniques in literature are proposed to improve efficiency of outlier detection process. Local outlier detection is important technique in data mining due to the recent need in data analysis. Most of the existing work focuses on degree of deviation and fails to find degree of dispersion. Due to this these systems fails to find accurate results on scattered data. There is need to develop a system that finds outlier points on scattered data in efficient manner.

Proposed System

A. System Architecture

The scatter data is input to the system. The system finds outlier points from dataset using E2DLOS algorithm. Before applying this algorithm dimension reduction and instance reduction process is applied. PCA is used for dimensionality reduction. Clustering based undersampling technique: RCMLQ is used to remove safe instances from dataset. Following figure shows the architecture of the system.

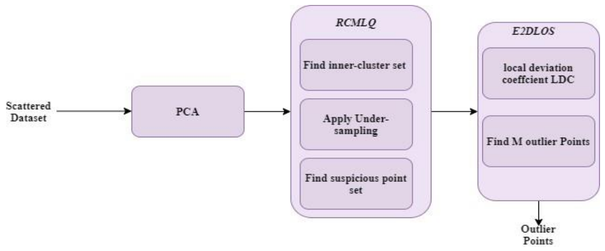

Fig System Archetcture

B. Preliminaries

1. Attribute Normalization

The attribute value can be normalized using following equation:

$$\mathrm{F(aji)} = \frac{a_{ji}}{\sum_{j=1}^{n} a_{ji}}$$

Where a_{ji} is the value of j^{th} attribute for instance i.

2. k-neighborhood average distance:

The k-neighborhood average distance of object o can be calculated as:

$$N_{k-adist}(o) = \frac{\sum_{i=r+1}^{k-r} |op|}{|N_k(o)| - 2r}$$

Where, |Nk (o)| : number of nearest neighbor of object o R = p % |N(0)|,
where p is threshold value varies from 5 to 20

3. k-Neighborhood Dispersion of Object: $N_{k\text{-}disp}$ of object O is calculated as:

$$Nk - disp(o) = (Nn_{k-adist}(O) + 1)^{Nn_{k-adist}(O) * Nn_{k-vari}(O)}$$

Where

$N_{nk\text{-}adist}$ Normalized k-Neighborhood Average Distance

$N_{nk\text{-}vari}$ normalized k-Neighborhood Variance of Object

4. Local Deviation Coefficient of Object:

The LDC of object O is calculated as:

$$LDC(o) = \frac{\frac{\sum_{p \in N(o)} N_{k-dens}(p)}{N_k(o)}}{N_{k\ \ dens}(o)} \quad (4)$$

Where $N_{k\text{-}dens}$(p) is k-Neighborhood Density of Object p.

5. *System Working*

The system takes scatter data as an input. Initially a preprocessing step is applied on the dataset. In this step dimension count of data is reduced using principal component analysis algorithm. This is a feature extraction technique. This technique project the dataset on to the different subspace with reduced dimension count. As we decrease the number of dimensions in the dataset, the time required for processing the data also gets reduced.

After implementing the dimensionality reduction the rough clustering algorithm is applied. This is undersampling technique. Generally the dataset contains very few outlier points. The elimination of safe non-outlier points in a dataset will reduce the dataset size for further outlier detection process. The RCMLQ algorithm generates rough clusters and add label to the inner cluster points as a safe points. The points which are not accommodated in a cluster are called as suspicious points. The suspicious point list is generated using RCMLQ algorithm.

In RCMLQ algorithm core clusters are created. It follows the iterative process and absorbs the points in core clusters based on distribution density of the data. The distance threshold is increased gradually to accommodate more points in a core cluster. The core cluster points are eliminated from the dataset and remaining points are called as suspicious points.

Local deviation coefficient LDC is calculated for all suspicious points in E2DLOS algorithm. Outlier points have high degree of LDC value. The list of suspicious point is sorted based on LDC value and top m outliers are extracted as an output.

6. Algorithm:

Algorithm 1: E2DLOS Algorithm
Input: Dataset with *O objects*,
Dcnt: Dimension count
K: Nearest neighbor count
MNDis: Threshold value
MinN: Threshold value
M-outlier: Outlier count Output: Top-*M outlier list*

Processing:

```
1.  Apply PCA and get reduced dimensions
2.  Apply RCMLQ(O, K, NNDis, MinN) and find suspicious point set Odoubt
3.  For each point in Odoubt
4.  Otemp : Find the k-neighborhood objects
5.  Calculate k-neighborhood average distance in normalize form
6.  normalize k-neighborhood average distance
    in normalize form
7.  Calculate the k-neighborhood variance
8.  k-neighborhood dispersion
9.  calculate the k-neighborhood
10. Oldc: calculate the local deviation coefficient LDC
11. sort the objects of the dataset using LDC value 12. Find top m-outlier objects.
```

Algorithm 2: RCMLQ Algorithm
Input: Dataset with *O objects*,
K: Nearest neighbor count
NNDis: Threshold value
MinN: Threshold value
Moutlier: Outlier count Output: *Odoubt* Processing:

```
1.  Create dataset matrix
2.  Normalize each column value
3.  calculates the NNDis neighborhood of all objects
4.  For all object in O
5.  if |ONNDis| >= k then
6.  create core cluster C
7.  add neighbors in cluster C
8.  else
9.  remove obect from Odoubt
10. End for
11. Add remaining points in Odoubt which are not part of C
12. merge all directly reachable-clusters in core cluster set C
13. if |Odoubt j >=MinN do
14. update NNDis = 2 * NNDis
15. go to step 4
16. else
17. return C and Odoubt
```

Result and Discussions

The system is implemented on windows environment with 4gbram and I5 processor. For development jdk 1.8 is used. The Netbeans 8.1 IDE is used for development.

A] Datasets:
UCI[9] and TAO[10] benchmark data sets datasets are used for system testing. Following table gives the detailed description of dataset.

Table 2 : Dataset Description

Sr. No.	Dataset	Number of instances	Number of attributes
1.	Forest Cover (FC)	581,012	54
2.	TAO	575,648	3

Two small two-dimensional synthetic datasets are created with scattered structure. as described in [1]. *B. Performance Measures:*

1. Evaluation Time: Execution Time local outlier detection process is captured with PCA and undersampling technique, with only undersampling technique.
2. Outliers and its neighbors: The top k outlier with its k nearest neighbors is identified with PCA and undersampling technique, with only undersampling technique.

C. *Implementation Status:*
The RCMLQ Algorithm is implemented and Odoubt i.e. Suspicious points from dataset are extracted. Following figure shows the results of RCMLQ Algorithm on two synthetic datasets. The suspicious points are labeled as O.

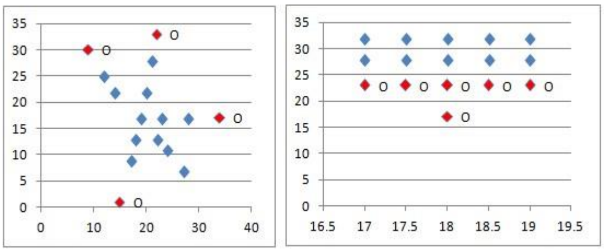

Fig RCMLQ Results

Conclusions

In this research work, the system generates a list of top m outlier points. The outlier points are extracted from scatter dataset. The outlier are identified using local outlier coefficient(LDC). To improve the efficiency of outlier detection process 2 pre-processing steps are applied: in first step dimension in the dataset are reduced and in second step, the amount of data that is needed to be processed for outlier detection process is reduced using rough clustering RCMLQ algorithm.

In future system will be implemented on hybrid dataset containing numerical as well as nominal attributes. The system can also be extended for streaming data.

References

[1]. Su, Shubin Xiao, Limin Ruan, Li Fei, Gu Li, Shupan Wang, Zhaokai Xu, Rongbin. (2018). An Efficient Density-Based Local Outlier Detection Approach for Scattered Data. IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2886197.

[2]. M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in Proc. ACM SIGMOD Int. Conf. Manage. Data, Dallas, TX, USA, May 2000, pp. 93-104.

[3]. P. Ciaccia, M. Patella, and P. Zezula, "M-tree: An efficient access method for similarity search in metric spaces," in Proc. 23rd Int. Conf. Very Large Data Bases (VLDB), Athens, Greece, 1997, pp. 426-435.

[4]. S. T. Leutenegger, M. A. Lopez, and J. M. Edgington, "STR: A simple and efficient algorithm for R-tree packing", in Proc. Int. Conf. Data Eng. (ICDE), Birmingham, U.K., Apr. 1997, pp. 497506.

[5]. K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlier detection approach for scattered real-world data," in Proc. Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining, 2009, pp. 813-822.

[6]. L. J. Latecki, A. Lazarevic, and D. Pokrajac, "Outlier detection with kernel density functions," in Proc. Int. Conf. Mach. Learn. Data Mining Pattern Recognit. (MLDM), Leipzig, Germany, Jul. 2007, pp. 61-75.

[7]. E. Schubert, A. Zimek, and H. P. Kriegel, "Generalized outlier detection with flexible kernel density estimates," in Proc. SIAM Int. Conf. Data Mining, 2014, pp. 542-550.

[8]. H. P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2008, pp. 444-452.

[9]. S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral," in Proc. 19th Int. Conf. Data Eng., Mar. 2013, pp. 315 -326.

[10]. Mahsa Salehi , Christopher Leckie , James C. Bezdek , Tharshan Vaithianathan and Xuyun Zhang, ""Fast Memory Efficient Local Outlier Detection in Data Streams," EEE Transactions on Knowledge and Data Engineering, vol. 28, no. 12, pp. 3246 - 3260, 2016.

[11]. S. Hettich and S. D. Bay, The UCI KDD Archive. Irvine, CA, USA: Univ. of California, Department of Information and Computer Science, 1999. Accessed: May 28, 2018. [Online]. Available: http://kdd.ics.uci.edu

[12]. Pacific Marine Environmental Laboratory. Pacific Ocean TAO. Accessed: May 26, 2018. [Online]. Available: http://www.pmel.noaa. gov/tao/