A Survey of Outlier Detection Techniques

Snehal More[1], Shirish Sane[2]
[1]M.E. Student, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education & Research, Nashik, India
[2]Professor, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education & Research, Nashik, India

***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. This paper aims to study various outlier detection techniques its advantages and limitations.**

***Keywords*: Outlier detection, Rough clustering, Scattered data, PCA, Under sampling, Feature extraction, Dimension reduction.**

1. 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 work 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.

Following section II includes the related work done in the domain of outlier detection followed by problem formulation. The Section III concludes the paper.

2. 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 does not generate accurate results for outlier 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 k-

Table 1
System Comparison

		LOF[2]	MiLOF[10]	LDOF[5]	KDEOS[7]	E2DLOS[1]
Outlier Detection Strategies	1.Density based Detection	YES	YES	-	YES	YES
	2. Distance Based Detection	-	YES	YES	-	-
Silent Features	Outlier 1. 1.Detection over Structured dataset	YES	YES	YES	YES	YES
	2.Outlier detection over scatter dataset	-	-	-	-	YES
	3.Clustering	YES	YES	YES	YES	YES
	4.Data summarization and filtering for Efficiency Improvement				YES	YES

nearest 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(k2) 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(k2).

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 under sampling technique. The system only focuses on sample reduction and not the dimension reduction.

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. Following table shows the summarized working in terms of outlier detection system strategies and silent features.

3. Conclusion

Outliers are the byproduct of clustering algorithm. Many existing approaches apply clustering and then based on the clustering results outlier are detected. The efficiency of outlier detection process is depending on the clustering algorithm. To improve efficiency of clustering algorithm data reduction is required in terms of attribute and instances. Most of 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 new system is required for finding outliers over scatter data with the help of degree of deviation and degree of dispersion with efficient execution.

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.

[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. 2003, 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," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 12, pp. 3246-3260, 2016.