Journal of Applied Science and Engineering Innovation, Vol.4 No.3, 2017, pp. 72-75
ISSN (Print): 2331-9062
ISSN (Online): 2331-9070

# Improvement of Outliers Detection Algorithm Based on Density

Nuong Hoang[1], Khon Loi Nguyen[1], Bach Huynh Dunnigan[2]

[1]*Department of Computational and Applied Mathematics, VNU University of Science, 19 Le Thanh Tong, Hanoi, Vietnam*

[2]*Department of Informatics, VNU University of Science, 19 Le Thanh Tong, Hanoi, Vietnam*

**Abstract:** In this paper, firstly two improved algorithm methods are introduced, namely INFLOF and COF, which are based on LOF, then the motivation of each algorithm, the definition of the algorithm and the specific steps of the algorithm are described respectively. Then through summarizing LOF, INFLOF and COF it can find out the intrinsic link between them: INFLOF can solve the problem of edge misjudgment caused by different density cluster's closing to each other in data set, while COF can solve the problem of outliers, but these kinds of two algorithms are from different steps to solve the outlier factor. Finally, the advantages of the these two algorithms are presented, thus the algorithm of this paper is introduced. Moreover, the definition of the algorithm, as well as the specific steps of the algorithm is respectively introduced, besides it also analyzed the time complexity of algorithm.

**Keywords** Outliers detection, LOF algorithm, Outlier factor

## INTRODUCTION

In statistics, an outlier is an observation point that is distant from other observations.[Maciá-Pérez, *et. al.*, 2015][Santiago, *et. al.*, 2015] An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.[Wu, *et. al.*, 2015]

Outliers can occur by chance in any distribution, but they often indicate either measurement error or that the population has a heavy-tailed distribution. In the former case one wishes to discard them or use statistics that are robust to outliers, [Reinoso, *et. al.*, 2015] while in the latter case they indicate that the distribution has high skewness and that one should be very cautious in using tools or intuitions that assume a normal distribution.[Breunig, *et. al.*, 2000] A frequent cause of outliers is a mixture of two distributions, which may be two distinct sub-populations, or may indicate 'correct trial' versus 'measurement error'; [Vidal, *et. al.*, 2015] this is modeled by a mixture model.

In most larger samplings of data, some data points will be further away from the sample mean than what is deemed reasonable. This can be due to incidental systematic error or flaws in the theory that generated an assumed family of probability distributions, [Aggarwal, et. al., 2001] or it may be that some observations are far from the center of the data. [Lozano, et. al., 2005] Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. However, in large samples, a small number of outliers is to be expected (and not due to any anomalous condition).

Data mining is a new subject with the explosive growth of data generated in the information society, it can dig out the effective, novel, potential, useful, and knowledge that can be understood be understood by the people ultimately from the massive data. With the arrival of the era of big data, data mining has been paid more and more attention. At present, data mining has played an irreplaceable role in all aspects of social life.[Aggarwal, *et. al.*, 2005] The traditional data mining out its focus on the model that most of the data are concerned with, such as the frequent pattern and the discovery of association rule, categories of judgment and description and clustering analysis and so on, outlier detection is the relatively sparse and isolated abnormal data mode that is found from massive data. Since LOF is put forward, many scholars put forward the improved algorithm, which can be divided into two aspects: one is to improve the efficiency of outlier detection, the other is to improve the accuracy of outlier detection in the complex data distribution. [Prastawa, et. al., 2004] For the former, it is mainly to remove the class or region which can not contain outlier by clustering or partitioning, so as to reduce the amount of data. In this paper, it studies on the second aspect of the problem, which out its focus on how to improve the accuracy of outlier detection through improved definition of outlier factor, so as to make the data points have outlier factor with higher degree. Wenetal as well as other people proposed an outlier factor based on symmetric neighborhood. INFLOF (Influenced Local Outlier Factor) can define the outlier factor based on the symmetric neighbor relationship, the higher INFLO

value of the data is, the greater possibility of data become the outlier points. [Hautamaki, et. al., 2004] When INFLO calculates the outlier of data points, it should consider not only the nearest neighbor of the data points, but also its inverse nearest neighbors (RNN). [Pham, et. al., 2012] In this way, it can avoid the fact that the edges of the data have misjudgment when data sets are close to each other. Tang et. al. proposed COF (Connectivity based Outlier Factor) based on link outlier factor chain distance can be divided by the average value of chain distance of its all nearest neighbor point distance, so as to define the outlier factor of data. Thus, when the data distribution is sparse and some patterns are distributed, there will be a good effect of outlier detection. Subsequently, Hui Cao and other people proposed DSNOF (Density Similarity Neighbor based Outlier Factor), [Nasraoui, et. al., 1999] which can further strengthen the effect of outlier detection when COF presents the case of deviation. [Ma, 2016]

In this paper, it will firstly introduce the two main algorithm methods based on LOF, namely, INLOF and COF, then putting focus on the proposed improved algorithms according to the shortcomings of these two kinds of algorithms, moreover it analyzes the time complexity of the algorithm, in the next chapter it will analyze the effectiveness of the proposed algorithm through the experiment.

## LOCAL OUTLIERS DETECTION BASED ON INFLUENCED SPACE —INFLOF

LOF algorithm can measure the outlier degree by comparing data object with the density of K neighbor. [Fu, et. al., 2015] [Tu, et. al., 2009] However, if the data distribution is complex, especially when the density of the two different clusters are close to each other, there will be errors, next we can illustrate the specific situation through a concrete example. (Fig.1).

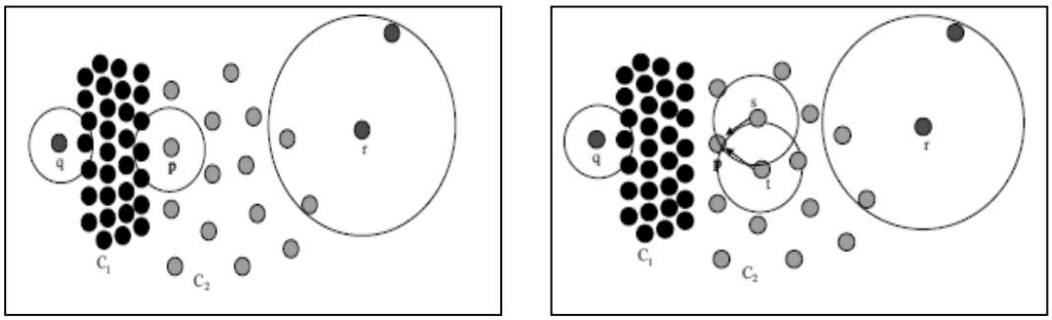


Fig.1 Comparing outlierness of *p*, *q*, *r*

As shown in Fig.1, *p* point is a relatively sparse point near the edge of the area, from the intuitive point of view, both *q* point and *r* point is outlier point that is further than *p* point, however, due to inspecting *k* neighbor of each point, because *p* point is too close to the relatively dense region, making its neighbor *k* come from the populated area, so its outlier factor value, namely the value of LOF will be higher. On the contrary, although *q* point is also close to the dense region, but because its distance is shorter from the dense region, the local reachable density is higher than that of *p* point, so that the outlier factor of the *q* point is smaller than that of *p* point. Now let's look at *r* point, it is obviously a far away from the dense point outlier region, but its neighbors are relatively sparse, so when we the calculate the LOF value of *p* point, [Goldstein, et. al., 2012] the relative calculation method will make the LOF value of r point much higher than that of *p* point. From the above analysis, we can see that when we adopt the traditional LOF value to calculate, it will make *p* point become more outlier than *q* point and *r* point, which is clearly not true.

INFLOF can effectively overcome these shortcomings, when it calculates the outlier factor of data point, it will not only consider KNN but also consider its RNN, the essence of RNN of one data point is that the data included by the nearest *k* neighbors points. We consider the set of points in the data KNN and the points contained in its RNN as the influence space of the data points, so that we can get a more reasonable evaluation when we calculate the outlier of the data points. As shown in Fig.1, the data point *p* of RNN contains data s point and t point from the sparse region, so that the INFLOF of outlier factor of p point can get more reasonable estimation. On the contrary, because *q* point has no RNN, so it's INFLOF of outlier factor is still higher, which is higher than the value of outlier factor of *p* point; although *r* point has RNN, its RNN is from sparse regions, its outlier factor had not too many changes. Therefore, when we adopt RNN to calculate the INFLOF of the data points, it will not appear the misjudgment like before. In the following, we will introduce the calculation method of INFLOF in detail.

## OUTLIERS DETECTION ALGORITHM BASED ON CHAIN DISTANCE—COF AND DSNOF

LOF algorithm puts its focus on the low density outlier, in which the outliers are defined as the density of points in the nearest neighbor region that are higher than their own points. However, when the data set are relatively sparse, when there is small density difference between outlier and normal point, this kind of method will not work according to the density of the points, in order to overcome this problem, Tang as well as other people proposed the concept of "deviation outlier", if there is no big density difference between the outliers and most of the points, then the outlier point can show the characteristics of deviating from the normal point, for those abnormal density of outliers, we also can see them that are caused by outliers from most of the high density area. As shown in Fig.2, o point deviates from the pattern of C1, while all the points in C1 follow the same pattern, but both o point and points in C1 belong to the region of low density.

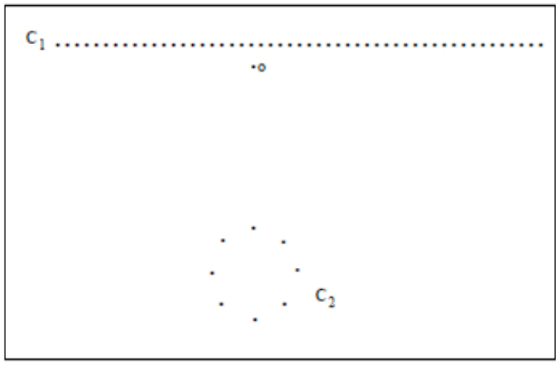


Fig.2 A low density pattern

In a word, the low density outliers are generated from the data points deviating from the high density density patterns, and the deviation outliers are generated by the fact that the data points deviate from the link patterns. COF extends the concept of outliers in LOF, which makes it no longer just refers to those with low density, but also the point deviated from the link mode. In the following we will introduce the specific calculation method of COF algorithm, which is similar to LOF and INFLOF, COF is also obtained through a series of interrelated definitions.

## IMPROVED LOCAL OUTLIERS DETECTION ALGORITHM BASED ON DENSITY—ISSDOF

By combing the outlier detection algorithm based on density, we find that the traditional LOF algorithm is only for those sparse outliers, that is to say if the density of one data point is sparse compared with its neighbors, the data points are considered as outliers, however when different density of clusters are closer to each other,the edge points will be regarded as outliers, INFLO method can effectively solve this problem. However, when there is no big difference between the density of outlier location and surrounding points, it belongs to the situation of model deviation, at this moment, neither the traditional LOF algorithm nor INFLO method can solve this problem, but COF method can solve this problem, in turn, COF method can not solve the problem of different density of clusters mutually closing to each other situation. The algorithm in this paper can combine the advantages of INFLO and COF, which can effectively mine out the local outliers under the circumstances that the different density of clusters are close to each other or they deviate from the model, thereby it can enhance the accuracy and generality of LOF algorithm.

## CONCLUSION

In this paper, an improved local outliers detection algorithm based on density is proposed. Through having in-depth analysis on two improved algorithms of outliers detection algorithm based on density namely, INFLOF and COF, we can find out their shortcomings, through integrating the advantages of two algorithms, an improved algorithm is proposed in this paper, thus the algorithm and specific steps are given, moreover it also analyzes the time complexity of the algorithm in this paper.

## REFERENCES

Aggarwal, C. C., & Yu, P. S. (2001, May). Outlier detection for high dimensional data. In ACM Sigmod Record (Vol. 30, No. 2, pp. 37-46). ACM.

Aggarwal, C., & Yu, S. (2005). An effective and efficient algorithm for high-dimensional outlier detection. The VLDB Journal—The International Journal on Very Large Data Bases, 14(2), 211-221.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In ACM sigmod record (Vol. 29, No. 2, pp. 93-104). ACM.

Fu, S., Lu, S., & Kai, G. (2015). Characteristics and control technology research of three-stage electro-hydraulic servovalve. Journal of Applied Science and Engineering Innovation, 2(2).

Goldstein, M., & Dengel, A. (2012). Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. KI-2012: Poster and Demo Track, 59-63.

Hautamaki, V., Karkkainen, I., & Franti, P. (2004, August). Outlier detection using k-nearest neighbour graph. In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on (Vol. 3, pp. 430-433). IEEE.

Lozano, E., & Acufia, E. (2005, November). Parallel algorithms for distance-based and density-based outliers. In Data Mining, Fifth IEEE International Conference on (pp. 4-pp). IEEE.

Maciá-Pérez, F., Berna-Martinez, J. V., Oliva, A. F., & Ortega, M. A. A. (2015). Algorithm for the detection of

outliers based on the theory of rough sets. Decision support systems, 75, 63-75.

Ma, X. (2016). Research on Condition Analysis of Information Security and Protection Strategy of Mobile Intelligent Equipment. Journal of Applied Science and Engineering Innovation, 3(5), 172-176.

Nasraoui, O., Krishnapuram, R., & Joshi, A. (1999, August). Mining web access logs using a fuzzy relational clustering algorithm based on a robust estimator. In Proc. the Eighth Int'l World Wide Web Conference, Toronto, Canada.

Pham, N., & Pagh, R. (2012, August). A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data. InProceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 877-885). ACM.

Prastawa, M., Bullitt, E., Ho, S., & Gerig, G. (2004). A brain tumor segmentation framework based on outlier detection. Medical image analysis,8(3), 275-283.

Reinoso, J. F., Moncayo, M., & Ariza-López, F. J. (2015). A new iterative algorithm for creating a mean 3D axis of a road from a set of GNSS traces.Mathematics and Computers in Simulation, 118, 310-319.

Santiago, C., Nascimento, J. C., & Marques, J. S. (2015). 2D segmentation using a robust active shape model with the EM algorithm. IEEE Transactions on Image Processing, 24(8), 2592-2601.

Schadt E E, et al. Feature extraction and normalization algorithms for high‐density oligonucleotide gene expression array data[J]. Journal of Cellular Biochemistry, 2001, 84(S37): 120-125.

Tu, L., & Chen, Y. (2009). Stream data clustering based on grid density and attraction. ACM Transactions on Knowledge Discovery from Data (TKDD), 3(3), 12.

Vidal, F. S., Barcelos, A. D. O. P., & Rosa, P. F. F. (2015, June). SLAM solution based on particle filter with outliers filtering in dynamic environments. In Industrial Electronics (ISIE), 2015 IEEE 24th International Symposium on(pp. 644-649). IEEE.

Wu, J., Kong, W., Mielikainen, J., & Huang, B. (2015). Lossless compression of hyperspectral imagery via clustered differential pulse code modulation with removal of local spectral outliers. IEEE Signal Processing Letters, 22(12), 2194-2198.