Three-dimensional Data Outlier Detected by Angle Analysis

Zhongyang Shen
China Mobile
Beijing, China
ORCID: 0000-0002-4826-0966

Abstract— **We present a method to distinguish outliers from spherical cluster distributed three-dimensional data. The angle measurement method 3DOD transforms three-dimensional data to two-dimensional data, and then outliers can be detected by conventional two-dimensional data outlier algorithm.**

Keywords— **outlier, three-dimensional, angle transformation, dimension reduction**

I. INTRODUCTION

In some data application cases, we will meet the requirements related to the outliers of three-dimensional data. To be distinguished from most data, outliers show their special characters in some cases. Currently, there are a variety of methods for abnormal recognition of two-dimensional data, such as LOF, Isolation Forest, OGAD[1], ABOD[3], DBSCAN[4] and other algorithms. For three-dimensional or high-dimensional data, as the number of dimensions increases, the amount of calculations will increase exponentially, which brings challenges to find the correlation of high-dimensional features in outlier detection. There are current algorithms such as ABOD and other algorithms supporting three-dimensional data outlier detection, but it has disadvantages with calculation complexity, training data, and proportion setting. Other algorithms such as LOF with PCA algorithm are processed through dimension reduction, bringing the loss of multidimensional data information to get some obviously abnormal results.

In this paper, an unsupervised algorithm method for outlier detect of three-dimensional data (3DOD) is proposed to solve the problems of computational complexity caused by the increase in dimension. Through the angle transformation method, three-dimensional data is transformed into two-dimensional data, with the characteristics of three-dimensional data are remained, and then the conventional two-dimensional outlier analysis method is used to detect outliers. Experiments show that this method can effectively detect the outliers of spherical cluster distributed three-dimensional data.

This paper verifies that the three-dimensional data can be dimensional reduced to two-dimensional data by angle conversion in outlier recognition, with the data characteristics of the three-dimensional data remained. Also, this paper verifies that the two-dimensional outlier algorithm such as OGAD, LOF still has the ability to identify outliers for three-dimensional data after reducing the dimension to two-dimensional data through angle conversion.

II. ALGORITHM

A. Principle

Outliers are some data that deviate from the normal data area. For the three-dimensional data of spherical cluster distribution, we assume there are a spherical cluster virtual boundary between normal points and outliers. To distinguish the outlier from normal points, we need to identify the virtual boundary of the spherical cluster. This method proposes an angle measurement method to determine the corresponding angle position of each three-dimensional data point by means of angle scanning from a spherical peripheral observation point, then convert the angle data into two-dimensional data to filter out the outliers.

B. Proof

First, we will try to prove that angle measurement method is able to separate the normal data and abnormal data from the spherical cluster-like three-dimensional data when we intersect tangents from observation points to the spherical cluster. In the case, it will form a cone shape with possible normal values inside the cone and outliers outside the cone (Figure 1). After angle scanning from several observation points in the periphery, the possible normal values are in the overlapping areas of all scanning areas, and the outlier values are in the other areas, which separated from normal points.

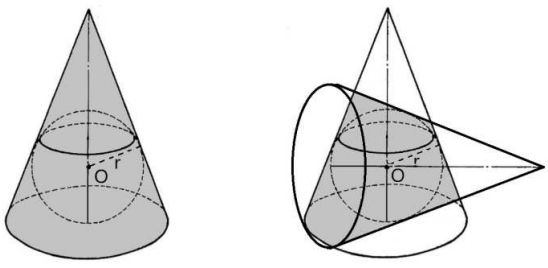

Figure 1. Example of angle scanning from observation points to detect abnormal values: shaded part is possible area of normal points

As an example, we take 14 external observation points with uniform distribution to compare the overlapping volume with the volume of the cluster. After angle scanning from all observation points, the difference between the two volumes is

compared to prove the effect of algorithm 3DOD on the identification and separation of outliers.

To simplify calculation, we assume that the observation point is a point at infinity, thus from the observation point, the scanning volume that intersects with the spherical cluster is to be close to a cylindrical volume. When we select fourteen evenly distributed observation points, these scanning operations will be converted into seven cylindrical volumes, and then the remaining volume will be formed by these seven cylinders intersected and overlapped. Now we can compare the remaining volume with standard spherical cluster size to analyze the difference.

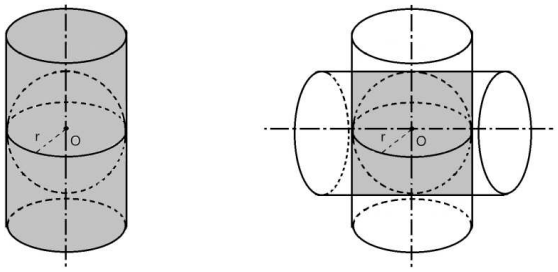

Figure 2. Example of angle scanning from observation point at an infinite distance: shadow portion is possible area of normal points

To calculate and compare the remaining volume, we use the following calculating process.

- Step 1: 14 uniformly distributed observation directions are selected, which the connecting lines from observation direction to the center of sphere were separated by equal angle.
- Step 2: Choose an observation direction.
- Step 3: The cylinder and the sphere are intersected on the surface of the sphere, which radius of the cylinder is equal to the radius of the sphere, and the height is the diameter of the sphere. Then the volume of cylinder is calculated.
- Step 4: The remaining volume will be the overlapped part with the above calculated cylinder volume and the previously calculated volume.
- Step 5: Go to Step 2 until calculation of all observation points are finished.
- Step 6: Calculate the remaining volume.

As calculation result, the ratio between remaining volume and spherical volume is 1.021:1, that is, the remaining volume is 2.1% larger than the spherical volume. Since the data in practical applications is not evenly distributed, the difference value can be reduced by increasing observation points or adjusting virtual boundary of the cluster, the fluctuation of 2.1% can be considered as an acceptable range of differences in the approximate calculation. When observation points are increased from 14 to 26 points or more, the difference value will be tended to be smaller.

C. Thought of Algorithm

Based on above result, we can see that outliers can be detected by calculating each three-dimensional data from the peripheral observation point.

In this paper, we present a new angle conversion calculation method by converting three-dimensional data into two-dimensional. Two-dimensional data is formed by a certain three-dimensional observation point, the method is, from the observation point, we can form an angle in the direction of XY plane and another angle in the direction of Y axis, and then these two angle values are composed as two-dimensional data. When all two-dimensional data are formed, we can use conventional two-dimensional outlier algorithm to detect the outliers.

Algorithm steps are showed as following:

- Step 1: Select 14 or more external observation points that are evenly distributed relative to the three-dimensional data cluster, which the axis between each observation point and the center of the spherical cluster is spaced equal angle apart from each other.
- Step 2: Select an observation point and calculate the two-dimensional data angles of each measured point. We mark the line from observation point to measured point as LINE0, and mark LINE0 projection line on x, y plane as LINE1.

  Angle A: Angle between X axis and LINE1.

  Angle B: Angle between LINE0 and LINE1.

  The Angles is rounded to simplify algorithm calculation.

  Two-dimensional data are formed by angle A and angle B, which is used for outlier identification.
- Step 3: Use conventional two-dimensional outlier identification methods such as OGAD or LOF to calculate outliers of the two-dimensional data formed by angle A and angle B.
- Step 4: Go to Step 1, move to next point until all observation points are calculated.
- Step 5: Collect all the calculated values to sort the result of outliers.

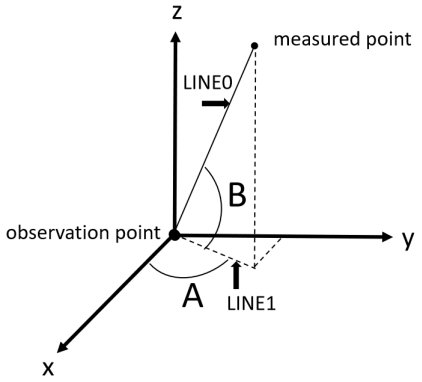

Figure 3. Example of converting angles into two-dimensional data

III. PSEUDO-CODE

The following pseudo code is based on the thought of algorithm described above.

Algorithm Program

```
//Get the barycentre position and radius length;
m ← count(Measured points) ;
Barycentre(x0,y0,z0)= 1/m ∑_{i=0}^{m} point(x,y,z);
for each point(x,y,z) ∈ Measured points do
    radius=max(distance(Barycentre(x0,y0,z0), point(x,y,z)));
end for;
//Get observation points which evenly distributed around measured points;
p ← quantity of observation points;
for i=0 to p do
  obser(xi,yi,zi)= position(
     base:Barycentre,
     length: (1+ratio）* radius,
     direction: evenly distributed around measured points
     );
  //Get Density for every angle from observation point;
  for each point(x,y,z) ∈ Measured points do
     angleA =integer (
          vertex: obser(xi,yi,zi) ,
          sideline: x-axis,
          obser(xi,yi,zi) to point(x,y,z) projection line on x,y plane
          );
     angleB =integer (
          vertex: obser(xi,yi,zi) ,
          sideline: obser(xi,yi,zi) to point(x,y,z),
          obser(xi,yi,zi) to point(x,y,z) projective line on x,y plane
          );
     pointAngle( αi, βi) ← (angleA,angleB);
  end for;
  LOF(pointAngle( αi, βi)) or OGAD(pointAngle( αi, βi)) →
          RankingAnomaly2D(pointAngle( αi, βi));
  RankingAnomaly2D(pointAngle( αi, βi)) →
          RankingAnomaly3D(point(x,y,z));
end for;
```

IV. EXPERIMENTAL DATA AND ANALYSIS

The following experiments show experimental results and comparison results. As different algorithms, ABOD and LOF with PCA dimensional reduction are used in the experiments as comparisons.

First, we design a set of simulation three-dimensional data, define specifically normal values and abnormal values, and then use algorithm 3DOD, LOF with PCA dimensional reduction and ABOD to identify respectively. In the experiments two-dimensional outlier algorithm OGAD is used in 3DOD as comparison to LOF with PCA dimensional reduction.

Table I and Table II show the experiment results with radius of standard sample set as 200 distance unit, and each experiment tests 200 times.

Table I. Experiments designed: number of standard sample set to 600, number of outliers set to 10

Outlier radius range (distance unit)	number of observation points(3DOD)	Times that match 10 outliers exactly			times match 9 outliers			times match 8 outliers		
		3DOD	LOF with PCA	ABOD	3DOD	LOF with PCA	ABOD	3DOD	LOF with PCA	ABOD
200-400	14	131	55	153	191	137	195	200	182	200
	26	139			195			200		
250-400	14	190	71	196	199	145	200	200	190	200
	26	193			200			200		
300-400	14	197	88	200	200	170	200	200	196	200
	26	198			200			200		

Table II. Experiments designed: number of standard sample set to 605, number of outliers set to 5

Outlier radius range (distance unit)	number of observation points(3DOD)	Times that match 5 outliers exactly			times match 4 outliers			times match 3 outliers		
		3DOD	LOF with PCA	ABOD	3DOD	LOF with PCA	ABOD	3DOD	LOF with PCA	ABOD
200-400	14	162	109	181	200	181	198	200	195	200
	26	172			200			200		
250-400	14	198	125	200	200	189	200	200	199	200
	26	198			200			200		
300-400	14	200	146	199	200	192	200	200	200	200
	26	200			200			200		

To verify the results, we analyze and compare the difference results of the above experiment data. In these cases, we select the case which outlier distribution is 200-400 distance units, number of normal values is 605, number of outliers designed is 5, and observation points of algorithm 3DOD is 14 points while the results are the same at 26 points.

The following Figure 4(a) and Figure 4(b) show difference between algorithm 3DOD and LOF with PCA dimensional

reduction, and Figure 4(c) and Figure 4(d) show difference between algorithm 3DOD and ABOD.

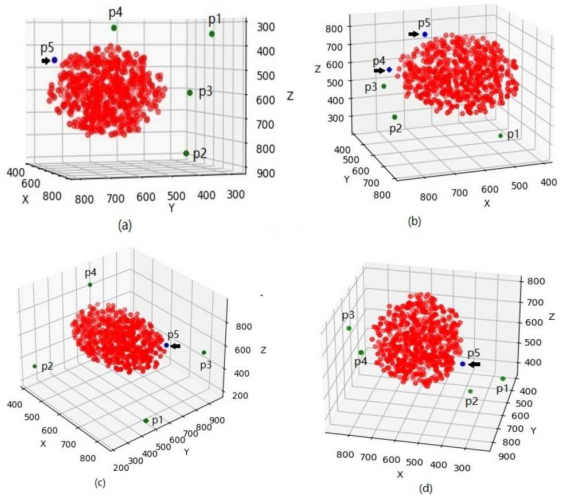

Figure 4. Outlier Cases

- Case 1: In Figure 4(a), algorithm 3DOD accurately identifies the defined 5 outliers (P1-P5), among which the blue point (P5) is ranked 5th in the outlier ranking, while in LOF with PCA dimensional reduction, 4 outliers are identified, and the blue point (P5) is ranked 216th in the outlier ranking, shown as in the normal value range. Result shows in this case algorithm 3DOD is more reasonable than LOF with PCA dimensional reduction.
- Case 2: In Figure 4(b), algorithm 3DOD accurately identified 5 defined outliers (P1-P5), among which 2 blue points (P4, P5) ranked 4th and 5th in outliers ranking, while 3 outliers are identified in LOF with PCA dimensional reduction, and 2 blue points (P4, P5) are ranked 14th and 507th respectively, which both showed in normal range. Result shows in this case algorithm 3DOD is more reasonable than LOF with PCA dimensional reduction.
- Case 3: In Figure 4(c), 4 outliers (P1-P4) are identified both by ABOD and algorithm 3DOD. The blue point (P5) is ranked 176th in ABOD and ranked 26th in algorithm 3DOD, which are in normal value range. In this case algorithm 3DOD and ABOD are similar in outliers detect.
- Case 4: In Figure 4(d), algorithm ABOD identifies 4 outliers (P1-P4), and algorithm 3DOD identifies 5 outliers (P1-P5). The blue point (P5) is ranked 11th in ABOD, while is within normal value range, and blue point (P5) is ranked 5th in algorithm 3GOD. Results show that in this case algorithm 3DOD is better than ABOD in outliers detected.

Through data analysis and comparison, we can find the following analysis results:

- As a conventional dimension reduction outlier algorithm, LOF with PCA dimensional reduction brings loss of effective information also will bring to the deterioration of outlier data. Results show LOF with PCA dimensional reduction has less accuracy in outlier detection than 3DOD and ABOD.
- 3DOD and ABOD have similar result in accuracy.
- For 3DOD, accuracy of outlier detection will be improved with increasing quantity of observation points.

Experimental data indicates that algorithm 3DOD is proposed as a dimension reduction algorithm for outlier recognition of three-dimensional data cluster. It can be seen from the experimental data that algorithm 3DOD is effective in achieving outlier recognition of clustered three-dimensional data, and it is significantly better than algorithm LOF with PCA dimensional reduction in terms of stability and accuracy. Comparing with ABOD, which training data and abnormality ratio needed to be set in advance, unsupervised learning algorithm 3DOD has advantages in computational complexity.

When analyzing the algorithm implementation process, we can see that algorithm 3DOD is based on the mechanism of data accumulation, that is, when data increasing, the new data is added into previous data set instead of full-scale calculations, which bring more effective.

Based on angle analysis, algorithm 3DOD can bring about the study of three-dimensional data and three-dimensional above data cluster identification by calculating the strongest density direction.

V. CONCLUSION

This paper presents a new algorithm 3DOD for three-dimensional data, the core idea and direction is three-dimensional data reduction by angle conversion from several external observation points. Experimental tests verify that the algorithm is reliable and accurate. It can be seen from experiment data that the method presented in this paper can play an effective role in the identification of outliers from three-dimensional data.

REFERENCES

[1] Zhongyang Shen, "Outlier Geometric Angle Detection Algorithm," 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2019, pp. 316-321, doi: 10.1109/ICAIIC.2019.8669090.

[2] Zhongyang Shen, "Cluster Quantity Distinguished by Geometric Angle Measurement," 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2020, pp. 514-519, doi: 10.1109/ICAIIC48513.2020.9065253.

[3] Hans-Peter Kriegel, Matthias Schubert, Arthur Zimek, "Angle-Based Outlier Detection in High-dimensional Data," The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2008.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD),AAAI Press.,1996.

[5] David Arthur, Sergei Vassilvitskii, "k-means++: the advantages of careful seeding," Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 2007, pp 1027 – 1035.

[6] Xie Huajuan, "Unsupervised Learning Methods and Applcations," Publishing House of Electronics Industry, China, 2016.

[7] Jiasi Shen and Martin Rinard, "Robust programs with filtered iterators," Proceedings of the 10th ACM SIGPLAN International Conference on Software Language Engineering (SLE), ACM, 2017, Pages 244 – 255, DOI:https://doi.org/10.1145/3136014.3136030

[8] Masashi Sugiyama, "An Illustrated Guide to Machine Learning," Kodansha Ltd., Japan, 2013.

[9] Toby Segaran, "Programming Collective Intelligence," O' Reilly Media, Inc., 2007.

[10] Pankaj K. Agarwal and Nabil H. Mustafa, "k-means projective clustering," In PODS'04:Proceedings of the twenty-third ACM SIGMODSIGACT-SIGART symposium on Principles of database systems, pages 155 – 165, ACM, 2004.

[11] Koki Saitoh, "Deep Learning from Scratch," O' Reilly Japan, Inc., 2016

[12] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, "A local search approximation algorithm for k-means clustering," Proceedings of the eighteenth annual symposium on Computational geometry, ACM, 2002, DOI:https://doi.org/10.1145/513400.513402.

[13] Bardia Yousefi and Chu Kiong Loo,"Comparative study on interaction of form and motion processing streams by applying two different classifiers in mechanism for recognition of biological movement," The Scientific World Journal, 2014.

[14] Andreas C. Muller, Sarah Guido, "Introduction to machine Learning with Python," O' Reilly Media, Inc., 2016.

[15] Zhou Zhihua, "Machine Learning," Tsinghua University Press, 2016.