Interpolation-based outlier detection for sparse, high dimensional data

Wanghu Chen*, Zhen Tian and LiZhi Zhang

College of Computer Science and Engineering, Northwest Normal University, Lanzhou, 730070, China

*Corresponding author’s e-mail: chenwh@nwnu.edu.cn

Abstract: The clustering-based approach has always been an important research direction of outlier detection technology. The current detection approaches based on clustering of outliers are able to overcome the shortcoming of traditional test approach of outliers, however, most of the existing approach based on clustering to improve the choice of initial clustering center, this does not solve poor due to the sparse data clustering effect, and so cannot be radically improve the accuracy of detecting outliers. The data in the outlier detection problem can be regarded as a high degree of mixing between the normal point and the abnormal point. On the premise of reducing the loss of normal points, the set of outliers is the N samples farthest from the clustering center containing the most outliers. Inspired by this idea, an outlier detection approach based on genetic clustering for high-dimensional sparse data was proposed. By applying genetic algorithm based on traditional k-means, this approach processed the difference of original data and solved the problem of poor clustering effect caused by high-dimensional data sparsity. The experimental results show that compared with several improved k-means-based clustering approaches and outlier detection approaches, the proposed approach can reduce the loss of normal points and distinguish normal data and outlier more accurately.

1. Introduction

Data-driven approaches play an increasingly important role in regular exploration and trend prediction. Since data quality has a great impact on the results of data analysis, it is often necessary to filter the anomalous data present in the original data set before analysing the data. Outliers are often thought of as individuals that deviate from other objects, so that people suspect that it is produced by another mechanism [1]. It looks like a subset and seems to be inconsistent with the rest of the set of data[2]. Therefore, outlier detection is the core problem in abnormal data detection[2], which has important research significance.

Supervised learning is a widely used outlier detection approach [3][4][5]. However, in the actual outlier detection task, the provision or selection of abnormal samples often has certain difficulty. If the coverage of the outlier features is insufficient, the accuracy of the detection is seriously affected. The unsupervised outlier detection approach is more versatile and more applicable because it does not depend on the tag data and is compared with the supervised approach. Cluster-based approaches have long been an important research branch of unsupervised outlier detection. In such approaches, most of the current research approaches rely on different strategies to improve the clustering effect by improving the initial clustering center and so on, thereby improving the accuracy of outlier detection. However, from the perspective of the spatial distribution of data, high-dimensional data is often sparse. Therefore, clustering based on distance or density thoughts tends to cause some non-outlier data to be excluded.
from each cluster obtained by clustering, reducing the recall rate of outliers; while increasing the distance or density threshold, it is easy to lead the interleaving of group point data with normal data reduces the accuracy of outlier detection. Therefore, this paper proposes an unsupervised outlier detection approach based on the idea of spatial data interpolation based on the idea of spatial data interpolation.

The following content is arranged as follows: The second part introduces the current related work, the third part proposes a high-dimensional sparse data outlier detection approach based on interpolation, and the fourth part gives the experimental results, and the experimental results are developed. Discussed in detail. Finally, the fifth part discusses the research results and looks forward to the future work.

2. Related Work

Supervised outlier detection approaches typically learn from sample data and establish patterns for normal or outlier samples. By establishing the hypothesis model $h(x)$ and determining the threshold $\rho$, when given the sample $x$, if $h(x) \geq \rho$, then $x$ is considered a normal sample, otherwise it is an outlier, and the outlier is in many cases. The lower is also considered to be an abnormal sample, and the threshold $\rho$ is often set according to empirical error [5]. Most existing outlier detection approaches can be considered based on the outlier detection framework, and many supervised outliers are evolved for the specific construction approach of the hypothesis model. Representative work includes [5][6][7][8].

The outlier detection approach based on supervised learning has the accuracy of the glue, but its accuracy has a high dependence on the number of samples and the quality of the annotation. To some extent, the number of samples and the quality of the annotation, determines the effect of the test. However, the labelling of many samples is a challenging task, posing many difficulties for supervised outlier detection. In many cases, the practical application of such approaches is also limited.

The unsupervised outlier detection approach usually does not require sample labelling, and its application conditions are more relaxed than the outlier detection approach based on supervised learning. A common idea is to estimate the density model of the training sample by parameterized or non-parametric approaches, and set the corresponding density threshold to determine the outliers by judging the local density [9]. Among the current unsupervised outlier detection approaches, representative approaches include detection approaches based on relative density estimation, model-based detection approaches, and support domain-based detection approaches.

The outlier detection approach based on relative density can effectively solve the problem of uneven distribution of data object density [10]. The most representative of the density-based approaches is the LOF algorithm [11]. However, this approach is sometimes sensitive to parameters, and the density estimation approach is difficult to make a correct judgment on the sparse region of the target class data.

Model-based approaches can be divided into three types depending on the shape of the model. The point reconstruction approach uses the distance from the sample point to the nearest cluster center as the reconstruction error to perform outlier detection, for example, outlier detection based on k-means [12] and k-center [13]. The plane reconstruction approach uses the distance of the nearest neighbour hyperplane as the reconstruction error of outlier detection. Considering that the point reconstruction approach is difficult to describe the planar cluster data, the k-plane clustering algorithm is proposed [14]. In recent years, many researchers have proposed outlier detection approaches based on curve, surface and subspace reconstruction. Principal component analysis [15] is a approach based on linear dimension reduction, which can also be extended in nonlinear environments, such as the KPCA approach proposed in [16]. Another nonlinear extension of PCA is based on the main curve [16], which constructs a cylindrical data description centred on the main curve based on the reconstruction error of each point in the data and the threshold set under a certain empirical error constraint.

In practical outlier detection applications, there is usually only one type of sample, so Scholkopf proposes a support domain-based approach [18], a single class SVM for outlier detection. In order to avoid the influence of receiving noise data and orphan data, it is not strictly required that the square of the distance of the training sample to the center of the sphere is less than or equal to R2, but the point larger than R2 will be punished. On this basis, David M.J. Tax and Robert P.W. Duin proposed the
description of support vector data [19].

The above unsupervised outlier detection approach lays a good foundation for the research of this paper. However, in the case of a limited number of samples, the spatial distribution of the data shows high sparsity, which makes it difficult to detect outliers based on density, model and support domain. At the same time, high-dimensional samples tend to have natural sparsity in spatial distribution. However, in the current research, there are not many concerns about outlier detection for high-dimensional sparse data. Therefore, this paper introduces the interpolation idea of high-dimensional data space, and attempts to explore an outlier detection approach for high-dimensional sparse data (ODGA).

3. Interpolation-based sparse data outlier detection

It is generally believed that outliers are significantly different from other objects in that they are suspected to be produced by another mechanism (Hawkins, 1980) and can therefore be used for the detection of anomalous data. If all sample data are clustered, the samples in each cluster have similar characteristics, and when the distance between the sample points and the centroid exceeds a certain range, it is difficult to be classified into any one cluster. At this time, the generation mechanism of the sample may be different from most samples, and therefore belong to the outliers with a large probability.

The unsupervised clustering approach is simple and the dependency conditions are relaxed. Among them, k-means is a representative distance-based clustering algorithm, which is one of the most widely used clustering algorithms. K-means uses distance as the basis for the evaluation of sample similarity. The closer the distance is, the higher the similarity of the sample is, which can be considered as finding a compact, independent cluster. Therefore, the outlier detection approach of this paper firstly detects the outliers based on k-means clustering by judging whether the sample points are greater than a certain threshold.

However, since the k-means algorithm uses the distance average as the clustering center and applies it to the next iteration, when the cluster center is a noise point or an isolated point, this causes the cluster center to deviate from the data set region. Therefore, if the data set contains many isolated points or noise data, the clustering result is greatly affected by noise or isolated points, resulting in inaccurate or even incorrect clustering results, resulting in separated outlier data and correct outlier data. A large deviation is produced. If the outlier detection is directly performed by the traditional k-means approach, many outliers will inevitably affect the clustering result, so that the outliers and normal values cannot be effectively separated, and the more outliers are detected. Increase the loss of normal points. This situation is more pronounced if the sample data is sparse.

Therefore, based on the basic idea of k-means approach, this paper introduces the interpolation idea of spatial data, and tries to prevent sparse data from being merged at the time of clustering to improve the detection effect of outliers. The basic idea of this approach is as follows:

a) All samples were taken as the initial population in the existing sample set, and the initial cluster center was randomly selected to complete a round of clustering.

b) In the cluster generated by clustering, rely on the fitness function to find out the class with the worst clustering effect in the sub-cluster

c) Interpolate the sample space of the above class and enter the next round of clustering.

d) In the final clustering result, it is considered a possible theoretical point to detect that the distance between the sample point and the centroid is too large.

The ODGA algorithm can be described as follows:
4. Sparse Data Interpolation-based on Genetic Algorithm

Since the genetic algorithm is a global optimization algorithm, the genetic algorithm is chosen to optimize the k-means clustering algorithm. It overcomes the fact that the general iterative approach is easy to fall into the local minimum trap and the "infinite loop" phenomenon, so that the iteration cannot continue. Compared with the traditional optimization approach (enumeration, heuristic, etc.), the genetic algorithm has a good prototype based on biological evolution. Convergence. When calculating the accuracy requirements, it has the advantages of short calculation time and strong robustness. In addition to the above advantages, the biggest advantage of applying genetic algorithm to this approach is that it compensates for the incompleteness of data information in the outlier detection of high-dimensional sparse sample sets. Increase data density by generating new data samples in an informed manner, helping to find a more favourable centroid.

Genetic algorithm is an adaptive global optimization probability search algorithm that simulates biological genetics and evolutionary processes in the natural environment. Genetic algorithms treat search and variables as finite-length strings. These strings are called chromosomes. Individual elements or features of these strings are called genes, and the value of the gene is called the allele [17]. In ODGA, we use binary encoding to treat each piece of data as a single chromosome, with each chromosome consisting of one gene, where l is the attribute. In the binary case, each bit of the encoded chromosome is initialized to a random 0 or 1.
In the experiment we used a single point crossover and a fixed crossover probability \( p_c \), which means that the population is first randomly paired, then the intersection is set, and finally some of the genes between the paired chromosomes are exchanged. In a single point crossing, the two "parents" are cut once at the corresponding points and the cut gene fragments are exchanged. Mutations are designed to help introduce diversity into the chromosomes of the offspring and introduce new genetic constructs by modifying some genes. In the binary case, a simple mutation can be accomplished by inverting the value of each gene with a small probability, usually the mutation probability \( p_m \) is obtained from \( p_m = \frac{1}{L} \), where \( L \) is the length of the chromosome.

Constructing an appropriate fitness function has an important influence on the performance of the optimization algorithm. In our experiments, we use the ratio of the distance within the class to the distance between the classes to define the fitness. Obviously, the smaller the distance between sample points in the same cluster is, the larger the fitness function is, and the better the clustering effect is.

We represent the sample set for each cluster as \( x_i \) each piece of data \( s \in x_i \) in the cluster.

The average distance \( d_c(x) \) within each cluster is expressed as:

\[
    d_c(x) = \frac{\sum_{s \in x_i}^{|x_i|} x_{s-i}}{|x_i|}
\]

We use \( i \) to represent the centroid, \( j \) to represent the number of clusters, and \( k_{ij} \) to represent the class spacing. The fitness function is:

\[
    J = d_c(x) \cdot \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{1}{k_{ij}} \quad (i \neq j)
\]

Through the fusion iteration of the two stages of genetic variation and clustering, we interpolate the sparse part of the original data, and use these new data to fill the sparse part of the original sample set, which improves the sparse data and is easily clustered. The problem of merging is to perform outlier detection after obtaining better clustering results. The algorithm in this part is described as follows:

```
Algorithm 3 Genetic Algorithm
0: function GENETIC_ALGORITHM(scope, genetic_rate, variation, column, Gafile, Tfile)
1:     gedata[ ][ ] = readData(Gafile, genetic_rate);
2:     parents[2][ ] = nextParents(gedata);
3:     an = getAttributionNumber();
4:     while parents != null do
5:         crosspoint = random(0, an);
6:         newdata[ ][ ] = regroup(parents, crosspoint);
7:         writeData(newdata, Tfile);
8:     parents[2][ ] = nextParents(gedata);
9: end while
10: while i < vadata.length do
11:     variapoint = random(0, an);
12:     variedata[i][ variapoint] = random(scope[0][ variapoint], scope[1][ variapoint]);
13:     writeData(newdata, Tfile);
14:     parents[2][ ] = nextParents(gedata);
15: end while
16: end function=0
```

5. Experiment and Analyses
The outlier detection approach proposed in this paper is suitable for identifying a small number of high-dimensional sparse outliers in the dataset. For the characteristics of this kind of data, we select some real datasets from the UCI machine learning database [18] to verify our approach. This experiment is based on the feasibility of the approach itself and the outlier detection algorithm based on the traditional K-means approach and based on various improved clustering. The data set selects Lymphography, WBC,
Ionosphere and UCI from UCI. In the Parkinson dataset, we sample the raw data and control the number of outliers to about 5% of the total data. The sampling results are shown in Table 1.

| Dataset      | Number of attributes | Number of instances | Number of outliers |
|--------------|----------------------|---------------------|-------------------|
| Lymphography | 18                   | 148                 | 6                 |
| WBC          | 9                    | 479                 | 34                |
| Ionosphere   | 34                   | 237                 | 12                |
| Parkinson    | 22                   | 155                 | 8                 |

The Lymphography data set contains 148 instances, each of which is described by 18 attributes. The data set has a total of four types of labels: normal findings, metastases, malignant lymphatics, and fibrosis. The smallest two classes only account for 4.05% of the total data. So, we use the points in these two classes as outliers, and the points in the other three classes as normal values. For the remaining data sets, we also sample the outliers by selecting the class with the smallest number of samples.

From a certain point of view, the outlier detection problem can also be regarded as a two-category problem. If the outliers are regarded as outliers and the remaining points are normal values, the detection approach can be used as a classifier to divide all the data. For both abnormal and normal categories. For the two-category problem, the sample can be divided into four cases: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). The "confusion matrix" of the classification results is shown in Table 2. The accuracy rate represents the ratio of the number of samples correctly classified by the classifier to the total number of samples in the baseline data set. In general, the higher the accuracy, the better the classifier. The calculation accuracy is calculated as follows [22]:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]  

Table 2. Confusion matrix

| Actual Results | Forecast Result | True | False |
|----------------|-----------------|------|-------|
| True           | TP              |      |       |
| False          | FP              |      | FN    |

The ORC[19], FindCBLOF [20] and ODC[21] approaches are based on the anomaly detection technology of the improved clustering algorithm. We have carried out the three approaches on the Lymphography dataset from the perspective of the classifier. For comparison, Table 3 shows the final classification results for these four approaches. It can be seen from Figure 1 that our approach (83.11%) is significantly better than ODC (66.22%), FindCBLOF (63.51) and ORC (50%) in terms of classification accuracy.

Table 3. Test results of four approaches

|     | ORC | FindCBLOF | ODC | ODGA |
|-----|-----|-----------|-----|------|
| TP  | 6   | 6         | 6   | 6    |
| FP  | 34  | 24        | 22  | 7    |
| TN  | 62  | 68        | 92  | 117  |
| FN  | 40  | 40        | 28  | 18   |
There is usually another standard to evaluate the classifier, TP_rate indicates the proportion of positive samples divided into positive samples, and FP_rate indicates the proportion of negative samples divided into negative samples. The TP_rate and FP_rate calculation approaches are as follows [23]:

\[
TP_{rate} = \frac{TP}{TP+FN} \quad FP_{rate} = \frac{FP}{FP+TN} \quad (4) \quad [23]
\]

Euc represents the distance between the classifier and the ideal classifier on the ROC map. The smaller the Euc value, the better the classifier. Euc is calculated as follows [24]:

\[
Euc = \sqrt{(FP_{rate})^2 + (1-TP_{rate})^2} \quad (5)
\]

Figure 2 shows that our approach achieves the smallest Euc value compared to the other three approaches and is significantly better than the other three approaches.

Since our approach proposes improvement measures based on k-means clustering, we also use the outlier detection approach based on traditional k-means and the ODC[21] approach based on improved k-means. Compare. The ODC [21] approach first assigns each point to the nearest centroid, and then identifies the outlier based on the proposed outlier definition. If an outlier is detected, it is removed from the data set and treated as an outlier. The point is stored separately, and then the centroid is recalculated. Iteration stops when sample points no longer move from one class to another. In this approach, the outlier is defined as: when the distance between the sample point and the centroid is a fixed multiple of the average of all sample points in the cluster and the centroid distance, the point is out of the group. Since the Lymphography dataset does not require sampling to meet our 5% requirement for all outliers, we still use the Lymphography dataset in this section. As shown in Figure 3, the absissa indicates the number of all points marked as outliers after the detection, and the ordinate indicates the recall rate. We can see that when the k-means approach finds all true outliers (recall rate = 100%), a total of 110 outliers are marked, of which 104 points are misjudged and 28 are marked by ODC approach[21]. The ODGA
approach marks 13 sample points as outliers, of which only 7 points are judged to be wrong. From this point of view, the detection accuracy of ODGA is significantly better than the outlier detection approach and ODC approach based on traditional k-means.

Since the k-means initial cluster centers are randomly chosen, the initial centroid of each algorithm changes. We applied ODC, ORC and FindCBLOF to the Lymphography dataset 10 times, the initial cluster centers were different, and they were randomly selected in the dataset. Experiments show that these outliers are basically located in each cluster farther from the centroid. Figure 4 shows that the ORC approach [19] accurately finds 6 actual outliers in 40 points considered to be outliers, and the FindCBLOF approach [20] finds 6 in the middle of 30 points marked as outliers. The true outlier, the ODC approach [21] found 6 true outliers among the 28 points marked. However, our proposed approach only marks 13 data as outliers, and contains all 6 outliers. From this we can conclude that compared with the other three approaches, the proposed approach not only achieves 100% recall (find all sparse class instances), but also minimizes the scope of identifying outliers.

In order to see if our approach can maintain good detection efficiency in the face of different data sets, we tested the proposed approach several times on the WBC, Ionosphere and Parkinson data sets in Table 1, respectively. The results are shown in Figure 5. On the WBC dataset, we found all outliers in only 46 points, and performed equally well on both the Ionosphere dataset and the Parkinson dataset.
Based on the above experimental analysis, the sparse part of the original sample data set is interpolated by genetic algorithm, which solves the problem that the sparse data is easy to be merged in the clustering process and obviously improves the clustering effect, thus improving the accuracy and precision of outlier detection.

6. Conclusion and Future Work

This paper explores an outlier detection approach that is especially suitable for high-dimensional sparse data sets. This approach draws on the law of population reproduction in biology, uses genetic algorithm to insert a large amount of relevant data into the sparse part of the data set to complete the interpolation operation, solves the problem that the sparse data in k-means cluster is easy to be merged, and solves the commonly used problem. The density estimation approach is usually difficult to make a correct judgment on the sparse area of the target class data. Experiments show that the detection approach can detect outliers efficiently.

In future work, we will further explore the mechanisms of detected outlier generation.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No.61462076 and No.61967013).

References

[1] Enderlein G. Hawkins, D. M.: Identification of Outliers. Chapman and Hall, London – New York 1980, 188 $\$ 14, 50[J]. Biometrical Journal, 2010, 29(2):198-198.

[2] Ord K. Outliers in statistical data: v. barnett and t. lewis, 1994, 3rd edition, (john wiley & sons, chichester), 584 pp. £55.00, isbn 0-471-93094-6. [J]. International Journal of Forecasting, 1996, 12(1):175-176.

[3] Aggarwal C C, Yu P S. Outlier detection for high dimensional data[J]. ACM SIGMOD Record, 2001, 30(2):37-46.

[4] L. Chomatek and A. Duraj, "Multiobjective genetic algorithm for outlier’s detection," 2017 IEEE International Conference on INnovations in Intelligent SysTemS and Applications (INISTA), Gdynia, 2017, pp. 379-384. doi: 10.1109/INISTA.2017.8001190.

[5] Steinwart I, Hush D, Scovel C. A Classification Framework for Anomaly Detection[J]. Journal of Machine Learning Research, 2005, 6(1):211-232.

[6] Shawetaylor J. Kernel Approaches for Pattern Analysis[M]// Kernel approaches for pattern analysis =. China Machine Press, 2005.

[7] Roth, Volker. Kernel Fisher Discriminants for Outlier Detection[J]. Neural Computation, 2006, 18(4):942-960.
[8] Roth V. Outlier Detection with One-class Kernel Fisher Discriminants[C]// Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]. DBLP, 2004.

[9] Tarassenko L. Novelty detection for the identification of masses in mammograms[C]// International Conference on Artificial Neural Networks. IET, 1995.

[10] Ester M, Kriegel H P, Xu X. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise[C]// International Conference on Knowledge Discovery & Data Mining. 1996.

[11] Breunig M M, Kriegel H P, Ng R T. LOF: identifying density-based local outliers[C]// Acm Sigmod International Conference on Management of Data. 2000.

[12] Jiang M F, Tseng S S, Su C M. Two-phase clustering process for outlier’s detection[M]. Elsevier Science Inc. 2001.

[13] Hochbaum D S, Shmoys D B. A Best Possible Heuristic for the k-Center Problem[J]. Mathematics of Operations Research, 1985, 10(2):180-184.

[14] Bradley P S, Mangasarian O L. k-Plane Clustering[J]. Journal of Global Optimization, 2000, 16(1):23-32.

[15] Duda R O, Hart P E, Stork D G. Pattern classification /M// Pattern classification. 2001.

[16] Hoffmann H. Kernel PCA for novelty detection[J]. Pattern Recognition, 2007, 40(3):863-874.

[17] Hastie T, Stuetzle W. Principal Curves[J]. Jasa, 1989, 84(406):502-516.

[18] A. Frank and A. Asuncion, “UCI machine learning repository,” 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[19] V. Hautam¨aki, S. Cherednichenko, I. K”ark”ainen, T. Kinnunen, and P. Fr¨anti, “Improving k-means by Outlier Removal,” in Proc. 14th Scandinavian Conference on Image Analysis (SCIA’05), 2005, pp. 978–987.

[20] Z. He, X. Xu, and S. Deng, “Discovering cluster based local outliers,” Pattern Recognition Letters, vol. 2003, pp. 9–10, 2003.

[21] Ahmed, M., & Naser, A. (2013). A novel approach for outlier detection and clustering improvement. 2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA). doi:10.1109/iciea.2013.6566435

[22] V. Hautam¨aki, S. Cherednichenko, I. K”arkk”äinen, T. Kinnunen, and P. Fr¨anti, “Improving k-means by Outlier Removal,” in Proc. 14th Scandinavian Conference on Image Analysis (SCIA’05), 2005, pp. 978–987.

[23] Z. He, X. Xu, and S. Deng, “Discovering cluster based local outliers,” Pattern Recognition Letters, vol. 2003, pp. 9–10, 2003.

[24] S.-y. Jiang and Q.-b. An, “Clustering-based outlier detection approach,” in Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery - Volume 02, ser. FSKD ’08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 429–433. [Online]. Available: http://dx.doi.org/