Outlier detection algorithm based on density and distance

Ying Qu\textsuperscript{1,2}, Fei Guo\textsuperscript{1,2}

\textsuperscript{1}School of Economics and Management, Hebei University of Science and Technology, Shijiazhuang, Hebei 050018, China
\textsuperscript{2}Data Science and Intelligent Computing Research Center, Hebei University of Science and Technology, Shijiazhuang, Hebei 050018, China

Corresponding author and e-mail: Fei Guo, 1230881321@cjlu.edu.cn

Abstract. In order to solve the problem that the existing outlier detection algorithm is difficult to detect the one-dimensional integer data set with uneven frequency distribution and uniform distance distribution and low accuracy, the advantages of density outlier detection and distance outlier detection can be combined. An outlier detection algorithm DAD (Density and Distance) based on density and distance was proposed. In order to improve the possibility of outlier sample distance, the algorithm can define the weight distance; introduce the global density outlier factor combined with the weight distance as the relative distance, and use the cutting edge strategy to quickly cut the outliers based on the minimum spanning tree. Then, an artificial data set was used to test the algorithm. Experimental results showed that the algorithm had a good outlier detection accuracy in dealing with data sets with uneven frequency distribution and uniform distance distribution.

1. Introduction

One-dimensional integer data with a single structure is the most basic data, which can be generated at any time in production and living activities and can be seen everywhere, such as student achievement data, student evaluation data, medical health data and traffic detection data. Therefore, the scientific preprocessing of one-dimensional data will directly affect the quality of data sets, and then affect the application effect.

Outlier detection is an important process of data processing, and its purpose is to effectively identify data that deviate from the general level in local or global scope. Usually, the outliers are identified and cleaned in the data preprocessing stage, or the outliers are directly taken as the concerned objects according to the requirements, which lays a good foundation for data mining.

In recent years, researches on outlier detection are mostly aimed at high-dimensional complex data sets, and outlier detection algorithms emerge one after another. However, the outlier detection of one-dimensional data will be affected by the uneven distribution of data frequency and uniform distance distribution, and the true outliers cannot be identified. Different density distribution can introduce global density outlier factor to improve the probability of density outlier of samples; the same density distribution can introduce frequency factor into classical Euclidean distance, which can effectively reduce the influence of uniform distance distribution of data, so it is suitable for outlier data mining in one-dimensional integer data space.
In this paper, an outlier detection algorithm based on density and distance is proposed for one-dimensional integer data sets. The algorithm defines the weight distance, calculates the neighborhood size by referring to the given neighborhood radius method, and calculates the relative distance by combining the weight distance with the global density outlier factor. Based on the minimum spanning tree, the outlier samples are cut quickly by using the maximum edge cutting strategy; Finally, the artificial data set is used to test the algorithm in this paper and verify that it is superior to other outlier detection algorithms in processing one-dimensional integer data sets.

2. Related work
Traditional outlier detection algorithms include statistics-based outlier detection, density-based outlier detection, clustering-based outlier detection and distance-based outlier detection. Statistics-based detection is to build a probability model, in which high probability samples are regarded as normal data and low probability samples are identified as outlier data samples. However, when the given data distribution is uneven, it is difficult to determine the appropriate model [1]. Density-based method considers that outliers are distributed in different density areas, and regards objects in low density areas as outlier data samples, focusing on local detection. Based on clustering method, samples that are not strongly related to other samples are regarded as outlier data samples, but the quality of clusters has great influence on the algorithm. Distance-based method, in which distance is mostly measured by proximity, treats samples far away from most data as outlier data samples, which is global, but is not suitable for large-scale data and cannot handle data with different densities.

According to the different data dimensions and actual scenes, the outlier detection algorithm is not limited to the classical algorithm, and many improved algorithms have appeared later. Chen Wanghu et al.[2] proposed ODGA outlier algorithm. Through sparse data interpolation processing on the clustered original data, the samples far away from the center of samples in the final result of clustering iteration are identified as outliers. Ma Wenqiang et al.[3] proposed HPOD algorithm. Aiming at the hub phenomenon in high-dimensional space, HPOD takes the ratio of the product of influencing factor and k distance and the sum of 1 to 1 MINUS the product of influencing factor and inverse nearest neighbor data and the sum of 1 as the outlier score of samples, and determines outliers according to the outlier score. Yang Xiaoling et al.[4] proposed RKNMOD algorithm. RKNMOD algorithm introduces relative distance to anti-K nearest neighbor, and uses minimum spanning tree to identify samples with large relative distance as outliers. Fan Ruixuan et al.[5] proposed PKNN algorithm. The average distance of the original k nearest neighbors is taken as the outlier, and the outliers are determined according to the outlier, in which the number of sample nearest neighbors is automatically determined by the algorithm. Li Yi et al.[6] proposed the NGOD algorithm. The algorithm defines the neighborhood granular outlier factor according to the defined granular outlier commitment, and judges outliers according to the outlier factor. Xie Xiong et al.[7] proposed LOLED algorithm. The ratio of local density estimation to the average value of local estimated density in K neighborhood is defined as local outlier degree, and outliers are identified according to local outlier degree. Feng Jiachen et al.[8] proposed FIF algorithm based on isolated forest, which defined the average path length after constructing isolated tree as outliers, and determined outliers according to outliers. Du Xusheng et al.[9] proposed NSD algorithm. The algorithm introduces intercept distance, defines the similarity between density and neighborhood average density as outlier, and the outlier without neighbor is 1. Outliers are identified according to outlier. Liao Liefa et al.[5] proposed an improved algorithm of E-iForest. The path length calculation formulas of three isolation strategies are improved by dimension entropy, and outliers are identified according to outlier scores. Li Changjing et al.[11] proposed an outlier detection algorithm by combining spectral embedding with local density. Yang Xiaoling et al.[12] proposed NRMFOD algorithm. The NRMFOD algorithm is suitable for complex multi-attribute data sets. However, the existing research focuses on the detection of global or local outliers in irregular data sets and multidimensional data sets, ignoring the outlier detection of one-dimensional integer data. According to the characteristics of uneven frequency distribution and uniform distance distribution in one-dimensional integer data sets, this paper proposes an outlier detection algorithm.
DAD based on density and distance. The algorithm defines weight distance, calculates the global density outlier factor of each sample in the sample set by referring to neighborhood density and neighborhood average density, and calculates the relative distance by combining weight distance with global density outlier factor. Based on the minimum spanning tree, the algorithm uses the maximum edge cutting strategy to segment outliers and clusters. Finally, artificial data sets are used to test the algorithm and verify its effectiveness.

3. Detection algorithm based on density and distance

3.1. Detection algorithm based on density and distance

In distance-based outlier detection, K nearest neighbors are often used to measure the distance within data samples, and K data samples with the closest distance are regarded as similar objects, and outliers are identified according to outlier factors. However, for one-dimensional integer data sets, it is easy to ignore the frequency of each data sample by using K nearest neighbor, and because the distance distribution of different data samples is uniform, K nearest neighbor can not reflect the distance between data samples well. According to the outlier factors alone, there may be cases where the outlier factors of data points in dense areas and sparse areas are the same. Therefore, this paper uses the given neighborhood radius method to measure the distance between data samples, and defines the frequency factor to improve the outlier probability of data sets.

The neighborhood size and average density of samples will affect the outlier degree of data samples, that is, the outlier degree of data samples is closely related to the number of samples contained in the neighborhood and the difference between samples. In this paper, the following definitions are introduced according to the characteristics of uneven frequency distribution and uniform distance distribution of one-dimensional integer data.

Definition 1 (Outliers) If \( U = \{x_1, x_2, \ldots, x_n\} \) is a one-dimensional data set, \( Q \) is a sample outlier measurement function, the outlier mapping results of all samples in \( U \) are \( Y \), and the outlier mapping results of most samples in \( U \) are \( Y_1 \), then the outliers are in the set \( O = \{x \in U | x \notin Y_1\} \).

Definition 2 (Neighborhood) Using the method of given neighborhood radius, the neighborhood of sample \( X \) is defined as the set of samples (including sample \( X \)) contained in a circle with \( X \) as the center and \( K \) radius, which is denoted as \( SKNNk(x) \).

\[
SKNNk(x) = \{x | x, y \in U, |x - y| \leq K\}
\]  

(1)

Definition 3 (Neighborhood density of samples) Neighborhood density is defined as the number of samples per unit area. The neighborhood density of the sample is defined as:

\[
P = \frac{N}{\Pi K^2}
\]

(2)

Where \( n \) is the number of samples (including sample \( x \)) contained in the neighborhood of sample \( x \). Obviously, if \( x \) is located in a data-intensive area, the neighborhood density of sample \( x \) increases with the increase of \( n \) value. If \( x \) is outlier, the smaller the value of \( n \) is, the smaller the neighborhood density of \( x \) is. Neighborhood density of samples can be used to measure local data outliers, which reflects the difference between a data sample and data samples in its neighborhood.

Definition 4 (neighborhood average density of samples) If the number of samples contained in the neighborhood set of sample \( X \) is \( n \), the neighborhood average density of \( X \) is defined as:

\[
\bar{P} = \frac{\sum_{x=1}^{N} P}{N}
\]

(3)

Neighborhood average density can be used to fully reflect the difference between a data sample and other data samples outside its neighborhood.

Definition 5 (global density outlier factor) The global density outlier factor can be defined as:
$$SDOF \left( x \right) = \frac{P}{P} = \frac{N^2}{\prod K^2 \ast \sum P}$$  \hspace{1cm} (4)$$

SDOF \left( x \right) \text{ combines the local neighborhood density and the average neighborhood density of sample } \text{X. The global density outlier SDOF} \left( x \right) \text{ is approximately 1. The denser the sample area is, the greater the average density of the neighborhood is, and the smaller the global density outlier SDOF} \left( x \right) \text{ is. The more sparse the sample area is, the smaller the average density of the neighborhood is, and the larger the SDOF} \left( x \right) \text{ is. SDOF} \left( x \right) \text{ varies with the distribution,}

Choosing a larger global density outlier factor can effectively improve the probability of density outlier of outlier samples.

Definition 6 (Weight Distance) For one-dimensional data set \(U\), the weight distance between samples \(X\) and \(Y\) can be defined as:

$$QD(x,y)=d(x,y) \ast \text{MAX}(FNF(x),FNF(y))$$  \hspace{1cm} (5)$$

For \(X_i \in U\), if the frequency of \(X_i\) in \(U\) is \(n\), then the frequency factor is defined as:

$$FNF(x)=\frac{1}{n}$$  \hspace{1cm} (6)$$

Where \(d (x, y)\) is the Euclidean distance between samples \(X\) and \(y\). Obviously, if \(x\) is located in a sparse region, the frequency \(n\) of \(X\) is smaller than that of dense data points, and the frequency factor is larger. The combination of larger frequency factor and Euclidean distance as weight distance can effectively alleviate the influence of uniform distance distribution of data sets under the same density, and enhance the possibility of distance outliers. Combined with global density outlier factor and weight distance, the distance between one-dimensional data objects can be reflected more objectively.

Definition 7 (relative distance of samples) for one-dimensional data set \(U\), the distance between sample \(X\) and \(Y\) can be defined as:

$$D(x,y) = \text{MAX}(SOF} \left( x \right) , SOF \left( y \right) \ast QD(x,y)$$  \hspace{1cm} (7)$$

Because SDOF \(x\) can vary with different distributions, the sparser the sample is, the larger the SDOF \(x\) is. The frequency distribution of one-dimensional integer data is uneven and uniform distance distribution. Selecting the maximum global density outlier factor can highlight the difference of density among data samples and improve the possibility of outlier of abnormal data. However, in the case of one-dimensional integer data sets, the frequency of a sample in a dense region is the same as that of a sample in a sparse region, or the Euclidean distance of the adjacent data in the dense region and the sparse region is the same as the global density outlier. Therefore, the combination of larger frequency factor and Euclidean distance can effectively solve this problem.

3.2. Algorithm description

The outlier detection algorithm based on density and distance calculates the neighborhood density and average density of the data sample to obtain the global density outlier factor. The larger frequency factor and the Euclidean distance are used as the weight distance, and the larger global density outlier factor is combined with the weight distance to obtain the relative distance between samples. Based on the minimum spanning tree, the maximum cut edge strategy is used to quickly cut outliers. When the number of split nodes is less than the specified proportion, outlier detection can be finished. Therefore, the algorithm steps are as follows:

Input: data set \(U, x_i \in U\), neighborhood radius \(k\) and outlier threshold \(m\%\).

Output: outlier set \(p\).

Begin

\(x \in U, y \in U,\)

Step 1, For each data point \(x\)

Step 2, calculating the neighborhood SKNNk(x) of \(x\);

Step 3, calculate the number \(n\) of samples (including \(x\)) in the neighborhood of \(x\);

Step 4, calculating the neighborhood density of \(x\);
5

Step 5, calculating the neighborhood average density of x;
Step 6, calculating the global density outlier factor of x;
Step 7, End for
Step 8, For each data point x
Step 9, Calculate the frequency factor FNF(x) of x;
Step 10, calculating d(x,y);
Step 11, calculating QD(x,y);
Step 12, calculating D(x,y);
Step 13, End for
Step 14:, Generate the minimum spanning tree according to the D matrix
Step 15, cuts the largest edge to form sets T1 and T2
Step 16, If then
Step 17, P = T1
Step 18, End if
End

When calculating the neighborhood density of each data, the algorithm in this paper needs to determine the K value and the frequency of each data sample in the neighborhood, and then calculate the global density outlier factor and weight distance of all data. X and y are different, and their global density outlier factors and weight distance are different. Combining the larger frequency factor with Euclidean distance as the weight distance, and then combining with the larger global density outlier factor, the relative distance between samples is calculated. The time complexity of obtaining the relative distance is T(n^2), where n is the number in the one-dimensional integer data set u. Using the relative distance between the minimum spanning trees, the outlier clusters or outliers are obtained by using the maximum edge cutting strategy. Therefore, the time complexity of the algorithm is T(n^2).

4. Experiment and result analysis

4.1. Experimental data sources

Since there is no recognized public data set for outlier detection, data sets are usually sampled and outlier thresholds are set as required. In this paper, two data sets of students' evaluation of teaching with different class sizes are selected as test data sets for experiments. The normal value of students' evaluation data is between (85,100), and the abnormal data is outside this range. Data1 data set is the teaching evaluation data of students with class size of 24, and Data2 data set is the teaching evaluation data set of students with class size of 105. The frequency distribution of each data in the test data set is shown in the figure, and sparse outliers can be clearly seen.

Figure 1. Data1 data set.
4.2. Experimental process

In order to test the effectiveness of the algorithm, the experiments were compared with the DAD algorithm, Z-score method and LOF algorithm, among which Z-score method based on statistics is suitable for data sets that do not conform to normal distribution, and LOF method based on density is often used to detect local outliers. The evaluation index adopts outlier detection accuracy, that is, the ratio of the number of real outliers to the number of detected outliers. The greater the accuracy, the greater the accuracy.

For Data1 and Data2 data sets, the DAD algorithm is used to set the neighborhood radius k to 2, and calculate the relative distance of the sorted data points, so as to generate a three-dimensional line chart of the relative distance between the data, and then increase the value of k to 8 in turn. The experimental results are shown in figs. 3 and 4.
Figure 4. Three-dimensional line chart of data 2 relative distance.

Figure 3 and figure 4 are three-dimensional line charts of the relative distance when the neighborhood radius \( k \) of Data1 data set and Data2 data set take different values. The \( x \)-axis takes the midpoint value of the sorted data series (for example, when \( K=2 \), the relative distance between 97 and 98, then the abscissa takes 97.5), the \( y \)-axis takes the neighborhood radius \( k \), and the \( z \)-axis takes the relative distance between the data series. The location of the peak is a region with a large relative distance. It can be seen from the two figures that, regardless of the difference in the number of peaks for the Data1 dataset with smaller sample size or the Data2 dataset with slightly larger sample size, it can be seen from the figures that the peaks are mostly gathered before 85, and the trend of relative distance is roughly the same when different values of \( k \) are taken, with little influence. Therefore, the value of \( k \) has a low influence on the accuracy of DAR algorithm, and DAD algorithm is more suitable for outlier detection of one-dimensional integer data sets. Cutting from the peak position, that is, the maximum relative distance, can effectively segment outlier data.

Table 1. Comparison of outlier detection accuracy of each comparison algorithm.

|              | Z-score method | LOF algorithm | NDARD algorithm |
|--------------|----------------|---------------|-----------------|
| Data1        | 3/6            | 3/3           | 3/3             |
| Data2        | 8/31           | 8/11          | 8/8             |

It can be seen from Table 1 that Z-score method and LOF algorithm are not as efficient as outlier detection algorithm based on neighborhood density and relative distance for multi-frequency ordered one-dimensional integer data. The algorithm proposed in this paper can effectively detect outliers and has better detection effect.

5. Conclusions

In this paper, an outlier detection algorithm DAD based on density and distance was proposed for one-dimensional integer data sets, the given neighborhood radius method was used to measure the neighborhood size, and the local neighborhood density which can reflect the difference between data samples and the average neighborhood density which can reflect the difference between data samples can be combined and taken as the global density outlier factor. A larger frequency factor and the classical Euclidean distance can be selected as the weight distance to improve the possibility of distance outlier of outlier samples. The larger global density outlier factor was combined with the weight distance to calculate the relative distance, which could improve the outlier possibility of outlier data samples; Then, the maximum edge cutting strategy was used to cut the outlier clusters or outliers quickly. Then, the artificial data set was used to detect and verify that the outlier detection algorithm
can effectively alleviate the influence of uneven frequency distribution and uniform distance distribution in one-dimensional integer data set based on density and distance.

Acknowledgement
This work was supported by the Hebei Higher Education Reform Research and Practice Project "Construction and Practice of the Big Data Thinking Ability Training System for Economic Management Talents in the Context of New Business" (No. 2019GJJG188) and the Hebei Innovation Funding Project "Hebei Province Based on Information Sharing" Supported by the research on the development mechanism of military-civilian integration of science and technology collaborative innovation" (No. CXZZSS2021096).

References
[1] Ding Tianyi, Zhang Min, Fang Shengliang. An outlier detection algorithm based on similarity pruning [J]. Miniature Microcomputer System, 2018, 39(08):1680-1684.
[2] Chen Wanghu, Tian Zhen, Zhang Lizhi, Liang Xiaoyan, Gao Yaqiong. outlier detection method for high-dimensional sparse data based on interpolation [J]. computer engineering and science, 2020, 42(06):966-972.
[3] Ma Wenqi, Zhao Xujun, Zhang Jifu, Rao Yuanqi. An outlier detection algorithm for hub phenomenon [J]. Miniature Microcomputer System, 2020, 41(05):919-924.
[4] Yang Xiaoling, Feng Shan, Yuan Zhong. Outlier detection of anti-K nearest neighbor tree based on relative distance [J]. Acta Electronica Sinica, 2020, 48(05):937-945.
[5] Fan Rui-Xuan, Jiang Gao-Xia, Wang Wenjian. A personalized outlier detection algorithm for K nearest neighbors [J]. Miniature Microcomputer System, 2020, 41(04):752-757.
[6] Li Yi, Hu Jiancheng. A neighborhood granular outlier detection for mixed attribute data [J]. Miniature Microcomputer System, 2020, 41(04):855-860.
[7] Xie Xiong, Tang Yu. Local outlier detection algorithm based on local estimated density [J]. Miniature Microcomputer System, 2020, 41(02):387-392.
[8] Feng Jiachen, Cai Jianghui, Yang Haifeng. A fast outlier detection algorithm for improved isolated forest [J]. Miniature Microcomputer System, 2019, 40(11):2418-2423.
[9] Du Xusheng, Yu Jiong, Chen Jiaying, Wang Yuefei, Pu Yonglin, Ye Lele. An outlier detection algorithm based on density difference measurement of neighborhood system [J]. Computer Applied Research, 2020, 37(07):1969-1973.
[10] Liao Lie-fa, Bin Luo. Outlier detection algorithm based on dimension entropy [J]. Computer Engineering and Design, 2019, 40(04):983-988.
[11] Li Changjie, Zhao Shuliang, Chi Yunxian. An outlier detection algorithm based on spectral embedding and local density [J]. Computer Science, 2019, 46(03):260-266.
[12] Yang Xiaoling, Zhang Xianyong. Outlier detection based on neighborhood rough membership function [J]. Computer Engineering and Design, 2019, 40(02):533-539.
[13] Ma Fumin, Chen Jingwen, Zhang Tengfei. Fast reduction algorithm of neighborhood multi-granularity rough sets based on double granulation criterion [J]. Control and Decision, 2017, 32(06):1121-1127.