Fuzzy Systems and Data Mining VI
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA200715

Water Quality Data Outlier Detection Method Based on Spatial Series Features

Jianzhuo Yan [a,b], Ya Gao [a,b,1], and Yongchuan Yu [a,b]
[a] *Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China*
[b] *Engineering Research Center of Digital Community, Beijing University of Technology, Beijing 100124, China*

Abstract. Outlier detection is one of the major branch in data mining which has been applied in different fields. Researchers have focused on the outlier detection in time series, but rarely spatial series. In this paper, we propose a new outlier detection method based on k-nearest neighbour (KNN) and Mahalanobis distance, which is first applied to the water field. Experimental results verify that the algorithm has good accuracy and effectiveness in outlier detection for water quality spatial series dataset.

Keywords. outlier detection, spatial series data, Mahalanobis distance, k-nearest neighbour

1. Introduction

One of the basic features of spatial data analysis is that it involves not only spatial attributes such as latitude, longitude and altitude, but also related non-spatial attributes. The spatial relation among the points in the spatial data set is established by an adjacency matrix which can be expressed as the adjacent distance relation. Due to the characteristics of spatial data mining, the detection of spatial outliers needs to find a specific data instance, whose non-spatial attribute values are obviously different from the size of adjacent points. Most of the existing spatial outlier detection algorithms focus on identifying single attribute outliers and may misclassify normal items as outliers, while the real spatial outliers exist in their neighborhood. Notably, many practical applications involve multiple non-spatial attributes that should be incorporated into outlier detection. Because the definition of neighborhood is crucial for determining spatial outliers, in addition, compared with the aggregated distribution of attribute values on all adjacent data, statistical methods are needed to characterize the distribution of attribute values at various locations. Therefore, the detection of spatial outliers is still a great challenge.

In this paper, for the spatial attributes and multivariable characteristics of water quality data, a new outlier detection method based on KNN(k-Nearest Neighbour) and Mahalanobis distance, is presented in this paper, Which first utilize KNN to find the neighboring function points of each data point, and then use the watershed as a

[1] Ya Gao, Corresponding author, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China; E-mail: gy@emails.bjut.edu.cn

comparison function under weight adjustment and the Mahalanobis distance suitable for multivariable as a threshold function to perform outlier detection on water quality data with spatial feature attributes. Results show that the method can detect outliers of water quality data well, with good accuracy and high sensitivity.

2. Related Work

Multivariate time series have been widely used in various fields, such as traffic data analysis [1],[2], livestock [3], finance [4], and others [5][6][7], etc. In recent years, there have been many researches devoted to predicting water quality, for instance, Liu et al.[8][9] forecast the water quality of a station over the next few hours from a data-driven perspective and present a multi-task multi-view learning method to fuse those multiple datasets from different domains into an unified learning model. However, there are few methods to study the outlier detection for water quality spatial data. Liu et al. [10] proposed an outlier detection algorithm for spatial sequence data sets based on the SOM neural network. Martínez Torres et al. put forward to use the depth function to curve the gas emissions with time, and obtain the outliers by comparing the curves instead of vectors [11][12]. Then, Blasi put the function depth function applied to spatial outlier detection of water quality data of different automatic monitoring stations in the Mino River Basin in Spain to detect spatial outliers [13]. Google Earth Engine methodology [14] and improved Z-score test [15] are used to detect spatial data. Ramaswamy et al. proposed a partition-based method that uses the k-th nearest neighbor to determine outliers [16]. Chen et al. demonstrated a neighborhood-based method [17] in 2010, which distinguishes outliers by the sum of all attribute distances of neighbors, but its rules are too inflexible to accurately detect outliers. Yu et al. proposed k local outlier factors and used k-walk similarity to create a new LOF-based metric [18], which improved the accuracy of outlier detection. Araki et al. used LUR and regression kriging to model the spatial distribution of the annual mean [19], and then generated a concentration map based on the predicted method of observation, that can clearly detect the content of PM2.5 and NO2 in air pollutants. However, the water quality data cannot fit the data into a concentration map, this method cannot be well applied to the detection of spatial outliers in water quality data. Harris et al. proposed a geographic weighting algorithm [20] for high-dimensional data. The geographic weighted PCA was used to calculate the score distance, orthogonal distance, and component distance, and the three scores were compared with the theoretical quantile. This method mainly uses PCA to reduce the dimensionality of high-dimensional data, but the correlation and interpretability of data variables in dimensionality reduction will be reduced, so it is not suitable for the detection of water affairs data.

3. Methodology

3.1. Mahalanobis Distance

Mahalanobis distance is a statistical distance proposed by Indian statistician Mahalanobis in 1936 [21]. This method considers the correlation between feature quantities and is not affected by the feature quantity scale [22]. Mahalanobis distance provides a suitable method to identify points far away from other points in multi-

dimensional space, it is widely used in discriminant analysis, clustering, and principle analysis. When dealing with multivariate data, it has many advantages over Euclidean distance. It is not affected by the dimension that the Mahalanobis distance between two points has nothing to do with the unit of measurement of the original data. It takes into account the connections between various characteristics, and excludes the interference of the correlation between variables.

If the density function of the multivariate normal random variable $(x_1, x_2, \cdots x_n)$ is as follows:

$$f(x_1, x_2, \cdots, x_n) = \frac{|\Sigma|^{-1/2} e^{-1/2(x-\mu)'\Sigma^{-1}(x-\mu)}}{(2\pi)^{-n/2}} \tag{1}$$

Where μ is the mean vector of the sample population and Σ is the covariance matrix of the sample population. Mahalanobis distance is defined as Eq.(2).

$$d(M) = \sqrt{(X-\mu)'\Sigma^{-1}(X-\mu)} \tag{2}$$

Where X is a feature vector composed of all feature quantities of sample point x.

If $\Sigma = LL'$, transform Eq.(2) as follows:

$$d(M) = \sqrt{(L^{-1}(X-\mu))'L^{-1}(X-\mu))} \tag{3}$$

where u has the same function as the Euclidean distance, which is the average of each feature; L has the function of normalizing and decorrelating each feature. From a geometric point of view, it can be regarded as the Euclidean distance after converting the original feature quantity into standardized uncorrelated data through L, in other words, the Euclidean distance is the equal weight operation of each feature quantity, and the Mahalanobis distance adjusts the sample feature weight by L Value to obtain a better distance. The Mahalanobis distance is based on the overall distribution characteristics of various samples. The feature quantity consistent with the overall correlation of the sample, the Mahalanobis distance is given a smaller weight, and the calculated distance is small, otherwise, a larger weight is given, and the calculated distance is larger.

3.2. K-Nearest Neighbour Algorithm

KNN algorithm is one of the most basic and simplest algorithms in machine learning algorithms. It can be used for both classification and regression. The idea of the KNN algorithm is very simple: for any n-dimensional input vector, corresponding to a point in the feature space, the output is the category label or a predicted value corresponding to the feature vector. The following are the steps of the KNN algorithm:

Step 1: Calculate the distance between the test data and each training data.

Step 2: Sort according to the increasing relationship of distance.

Step 3: Pick K points with the smallest distance.

Step 4: Determine the frequency of the first K points in the category.

Step 5: Return the most frequent category of the first K points as the predicted classification of test data.

3.3. Outlier Detection Algorithm based on Mahalanobis distance and KNN

Assuming that x for spatial object data points has $q(\geqq 1)$ attribute values, use a to represent the vector of these q values at x. The attribute value of a given set of spatial points X is $X=\{X_1, X_2, X_3,\ldots, X_n\}$, with p $(\geqq 1)$ spatial points, and the attribute function $f(x)$ is the attribute vector of each spatial point a. $\mathrm{NNk}(x_i)$ represents the k nearest neighbors of x_i, where $i=1,2,3,\ldots,n$, $k=k(x_i)$, and the neighborhood function $g(x)$ is the return attribute function of the function of the k nearest neighbors of $f(x)$. To detect spatial outliers, all attribute values of x should be compared with the corresponding quantities from the nearest neighbors of x.

We choose g as the vector of neighboring points, and each component represents the degree of influence of the neighboring points on the point. Then we calculate the difference between f and g, let $h(x)=f(x)-g(x)$, where $h(x)$ is the comparison function. Next, check the Mahalanobis distance from each point to that point, and the point whose distance is greater than the predetermined threshold will be returned as an outlier.

For samples $h(x_1)$, $h(x_2)$, $h(x_3),\ldots, h(x_n)$ and n spatially referenced objects $x_1, x_2, x_3,\ldots, x_n$. MCD is the sample mean μ_J^* and covariance matrix $\sum_J^*$ under sample s, is defined as the following formula:

$$MCD = (\mu_J^*, \sum_J^*) \tag{4}$$

$$\mathrm{J}=\{\left|\sum_J^*\right| \leq \left|\sum_M^*\right|, \quad \forall M\ s.t.\ |\mathrm{M}| = s\}\} \tag{5}$$

$$\mu_J^*=\frac{1}{s}\sum_{i\in J} h(x_i) \tag{6}$$

$$\sum_M^*=\frac{1}{s}\sum_{i\in J}[h(x_i)-\mu_J^*][h(x_i)-\mu_J^*]^T \tag{7}$$

The specific algorithm steps are as follows:

Step 1: For each spatial point x_i, calculate k nearest neighbor sets $\mathrm{NNk}(x_i)$.

Step 2: For each spatial point x_i, calculate the attribute function $f(x_i)$ and the neighborhood function $g(x_i)$.

Step 3: Calculate the comparison function $h(x_i)=f(x_i)-g(x_i)$.

Step 4: Assuming that the distribution of $h(x)$ is $Nq(\mu, \Sigma)$, it means that the q-dimensional vector $h(x)$ follows a multivariate normal distribution with mean vector μ and covariance matrix Σ. Use μ^* and Σ^* to simulate the real parameter mean vector μ and covariance matrix Σ. Then, calculate the μ^* and Σ^* of the comparison functions $h(x_1),h(x_2),h(x_3),\ldots, h(x_n)$.

Step 5: To check whether the distance reaches the requirement of the abnormal value, a predetermined threshold is needed.

If $\mathrm{d(x)} = [(h(x)-\mu_J^*)^T(\Sigma^*)^{-1}(h(x)-\mu_J^*)]^{\frac{1}{2}}$, $\frac{c(m-1+1)}{qm}d^2(\mathrm{x})$ is an approximate distribution, which is an F distribution with q and $(m\text{-}q+1)$ degrees of freedom, and the parameters m and c can be calculated from the asymptotic formula. Therefore, $\frac{c(m-q+1)}{qm}d^2(x) > F_{q,m-q+1}(a)$, which is the probability that $h(x)$ satisfies a. $F_{q,m-q+1}(a)$ is a percentage of the F distribution with q and $m\text{-}q+1$ degrees of freedom.

Step 6: Calculate the $d(x)$ in each spatial point x_i, where $d(x_i) = [(h(x_i) - \mu_J^*)^T(\Sigma^*)^{-1}(h(x_i) - \mu_J^*)]^{\frac{1}{2}}$.

Step 7: If $d^2((x_i) > \frac{qm}{c(m-q+1)}F_{q,m-q+1}(a)$, return as an outlier.

3.4. Evaluation of Performance

To evaluate the performance of the method we proposed, the following evaluation indicators are applied, while the calculation formulas are as follows:

(1) Accuracy

Accuracy rate measures the overall effectiveness of the classification model, which is the proportion of the number of correctly classified samples in the total number of samples. The formula can be obtained as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

(2) Sensitivity

Sensitivity represents the proportion of pairs among all positive examples, and measures the classifier's recognition ability of positive examples. The formula is as follows:

$$Sensitivity = \frac{TP}{TP+FN} \tag{9}$$

(3) Specificity

Specificity refers to the ratio of pairs among all negative cases, and measures the classifier's ability to recognize negative cases. The formula is as follows:

$$Specific \quad t \quad y \quad = \frac{TN}{TN+FP} \tag{10}$$

(4) F value

F value is the comprehensive evaluation standard of precision rate and recall rate.The formula has the following form:

$$F\ value = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{11}$$

Where Precision and Recall can be given by Eqs. (12) and (13).

$$Precision = \frac{TP}{TP+FP} \tag{12}$$

$$Recall = \frac{TP}{TP+FN} \tag{13}$$

Where, TP represents the case that the sample which belongs to a positive class is also recognized as a positive class, while FP represents the case where the sample that belongs to a negative class is identified as a positive class; TN represents the case that

the negative class is identified as the negative class, while FN represents the case that the positive class is identified as negative.

4. Results and Discussion

In this paper, the water quality data of the eleven sewage treatment plant stations in Beijing is selected as this experiment's data, coming from Beijing Municipal Water Affairs Bureau. These stations include Tuanjiehu, Gaobeidian, Qijiahuozi, Zhangfang, Songlin Gate, Longtan Gate, Sanjiadian, Xiahui, Zhangjiafen, Shidu, Hot Springs. We select four variable indicators as ammonia nitrogen (NH_4N), Nitrate (NO_3N) dissolved oxygen (DO) and permanganate index (CODMn). The time interval of data is from January 1, 2010, to January 1, 2011, a total of 1760 data. Each sewage treatment plant has 160 pieces of data at the same time. The outlier samples with labels are shown in Figure 1.

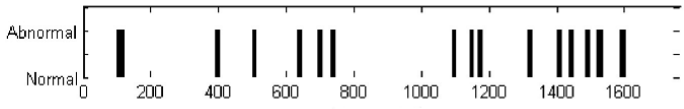

Figure 1. Normal value and outliers of real data.

This experimental algorithm is implemented by the matlab2018 platform. Since four attributes are used in the experiment, the parameter q = 4. Through the implementation of the FAST MCD algorithm in the Matlab toolbox LIBRA to generate a robust mean and covariance matrix estimate, calculate the parameters m and c, m = 352.365 and c = 0.469. In this algorithm, we take the parameters of each sewage treatment plant watershed as -1 to 1, indicating the correlation between the plants in the watershed.

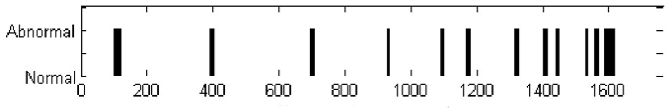

Figure 2. Normal value and outliers detected by the algorithm

The final detected outliers are shown in Figure 2. It can be seen that the algorithm detects most of the outliers in real data, and the detected error outliers are very few, which can meet the requirements of water quality data outlier detection algorithm. Table 1 illustrates the specific indicators of outlier detection of each sewage treatment plant.

Table 1. Comparison of classification indicators of each plant

Sewage treatment plants	Accuracy	Sensitivity	Specificity	F value
Tuanjiehu	0.8232	0.902	0.9862	0.9267
Gaobeidian	0.8597	0.9789	0.8898	0.9783
Qijiahuozi	0.8864	0.9932	0.9168	0.9219
Zhangfang	0.8794	0.9168	0.9378	0.8453
Songlin Gate	0.8657	0.8465	0.9634	0.9267
Longtan Gate	0.9425	0.9689	0.9265	0.9436
Sanjiadian	0.9347	0.8668	0.9767	0.9523

Xiahui	0.9689	0.9399	0.8864	0.9094
Zhangjiafen	0.8986	0.9648	0.9263	0.9812
Shidu	0.9279	0.8367	0.9459	0.9098
Hot Spring	0.9386	0.9124	0.9398	0.9764
Average value	0.902	0.9206	0.936	0.934

As shown in Table 1, the accuracy, sensitivity, specificity, and F value of each factory are almost above 0.9, and only a few indicators are around 0.85. Whether it is the index of each factory or the average index, it meets the classification index requirements for detecting outliers, indicating that the model can detect spatial outliers well and has a good effect.

In order to prove the effectiveness and better performance of our method, we compare the proposed algorithm with the K-nearest neighbor algorithm based on Manhattan distance(MD-KNN) and the weighted distance based outlier detection (WDBOD). As can be seen from the performance indicators in Table 2, compared with the MD-KNN and WDBOD, our proposed algorithm shows better results that precision increased from 0.8426, 0.8524 to 0.9949, recall increased from 0.8302, 0.8357 to 0.9021, and F value increased from 0.8399, 0.8440 to 0.9462.

Table 2. Comparison of classification indicators for outlier detection

Method	Precision	Recall	F value
MD-KNN	0.8426	0.8302	0.8399
WDBOD	0.8524	0.8357	0.8440
Our method	0.9949	0.9021	0.9462

5. Conclusion

Outlier detection for spatial series dataset is a very meaningful and challenging task. Facing the multidimensional dataset containing outliers, we propose an algorithm based on KNN and Mahalanobis distance, results illustrate that the method gets better results as compare to the existing technique, providing a new idea for detecting spatial outliers. At the same time, it provides good research value for data outlier detection with similar data features in other fields. In future research, we will collect more water quality spatial data and optimize the method to improve performance in outlier detection.

References

[1] Yu JC, Ju PH, Jia DH, Yue W. Zhang X. Spatial–temporal traffic outlier detection by coupling road level of service. IET Intelligent Transport Systems. 2019 June; 13(6):1016-1022.

[2] Pu J, Wang Y, Liu X, Zhang X. STLP-OD: Spatial and Temporal Label Propagation for Traffic Outlier Detection. IEEE Access. 2019 May; 7:63036-63044.

[3] Ismail ZH, Chun AKK. Shapiai Razak MI. Efficient Herd – Outlier Detection in Livestock Monitoring System Based on Density – Based Spatial Clustering. IEEE Access. 2019; 7:175062-175070.

[4] Cheng SH, Chen SM, Jian WS. A Novel Fuzzy time series forecasting based on fuzzy logical relationships and similarity measures. Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics; 2015 Oct 9-12; Kowloon, China; c2015. p. 2250-2254.

[5] Heydari G, Vali MA, Gharaveisi AA. Chaotic time series prediction via artificial neural square fuzzy inference system. Expert Systems with Applications. 2016 Aug; 55:461-468.

[6] Heydari A, Tavakoli M, Salim N. Detection of fake opinions using time series. Expert Systems with Applications. 2016 Oct; 58:83-92

[7] Lingras P, Haider F, Triff M. Granular meta-clustering based on hierarchical, network, and temporal connections. Granular Computing. 2016; 1(1):71-92.

[8] Liu Y, Zheng Y, Liang YX, Liu SM. Urban water quality prediction based on multi-task multi-view learning. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence; 2016 July 9-15; New York, USA; c2015.p.2576-2582.
[9] Liu Y, Liang Y, Ouyang K, Liu S, Rosenblum D, Zheng Y. Predicting Urban Water Quality with Ubiquitous Data - A Data-driven Approach. IEEE Transactions on Big Data. 2020; p.1-1.
[10] Liu Y, Lu H. Outlier detection algorithm based on SOM neural network for spatial series dataset. Proceedings of the 2018 Tenth International Conference on Advanced Computational Intelligence (ICACI); 2018 Mar 29-31; Xiamen, China; c2018. p. 162-168.
[11] Martínez Torres J, Nieto PJG. Alejano L. Detection of outliers in gas emissions from urban areas using functional data analysis. Journal of Hazardous Materials. 2011; 186(1):144-149.
[12] Martínez J, García-Nieto PJ, Piñeiro JI, Iglesias C, Taboada J. Air quality parameters outliers detection using functional data analysis in the Langreo urban area (Northern Spain). Applied Mathematics and Computation. 2014 Aug; 241:1-10.
[13] Blasi JIPD, Torres JM, Nieto PJG. Analysis and detection of functional outliers in water quality parameters from different automated monitoring stations in the Nalón River Basin (Northern Spain). Environ Sci Pollut Res Int. 2015; 22(1):387-396.
[14] Peter BG, Messina JP. Errors in Time-Series Remote Sensing and an Open Access Application for Detecting and Visualizing Spatial Data Outliers Using Google Earth Engine. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2019 Apr; 12(4):1165-1174.
[15] Aggarwal V, Gupta V, Singh P, Sharma K, Sharma N. Detection of Spatial Outlier by Using Improved Z-Score Test. Proceeding of the 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI); 2019 Apr 23-25; Tirunelveli, India; c2019. p.788-790.
[16] Ramaswamy, Sridhar, Rastogi, Rajeev, Shim. Efficient Algorithms for Mining Outliers from Large Data Sets. 2000.
[17] Chen Y, Miao D, Zhang H. Neighborhood outlier detection. Expert Systems with Applications. 2010; 37(12):8745-8749.
[18] Yu JX, Qian W, Lu H. Finding centric local outliers in categorical/numerical spaces. Knowledge & Information Systems. 2006; 9(3):309-338.
[19] Araki S, Shimadera H, Yamamoto K. Effect of spatial outliers on the regression modelling of air pollutant concentrations: A case study in Japan. Atmospheric Environment. 2017 Mar; 153:83-99.
[20] Harris P, Brunsdon C, Charlton M. Multivariate Spatial Outlier Detection Using Robust Geographically Weighted Methods. Mathematical Geosciences. 2014 Sept; 46(1):1-31.
[21] Mahalanobis PC. On the generalized distance in statistics. Proceedings of the National Institute of Sciences of India. 1936; 12(1):49-55.
[22] Ömer NG, Dogan GE. Power-quality event analysis using higher order cumulants and quadratic classifiers. IEEE Transactions on Power Delivery. 2006 Mar; 21(2):883-889.