FSDM 2020
November 13-16, 2020
Online Conference

Water Quality Data Outlier Detection Method based on Spatial Series Features

FSDM3412

Jianzhuo Yan, Ya Gao, Yongchuan Yu
Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

Introduction

A new outlier detection method based on k-nearest neighbour (KNN) and Mahalanobis distance is proposed by us, which is first applied to the water field. Firstly, utilize KNN to find the neighboring function points of each data point, and then use the watershed as a 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.

Methods

Outlier Detection Algorithm based on Mahalanobis distance and KNN

- ➢Step 1: For each spatial point x_i, calculate k nearest neighbor sets 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: Calculate the μ^* and Σ^* of the comparison functions h(x_1),h(x_2),h(x_3),…, h(x_n).
- ➢Step 5: To check whether the distance reaches the requirement of the abnormal value, a predetermined threshold is needed.

Graphics / Images

1 classification indicators of the water quality data of the eleven sewage treatment plant

- 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.
- Select four variable indicators as ammonia nitrogen (NH4N)、Nitrate (NO3N)dissolved oxygen (DO) and permanganate index (CODMn).

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

2 Experimental comparison

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).

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

Conclusions

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.

Thanks for your time and suggestions