Detection to water quality for Yangtze River using a machine learning method

Jingyi Li[1,*], Shiwei Chao[2], Xu Zhang[3]

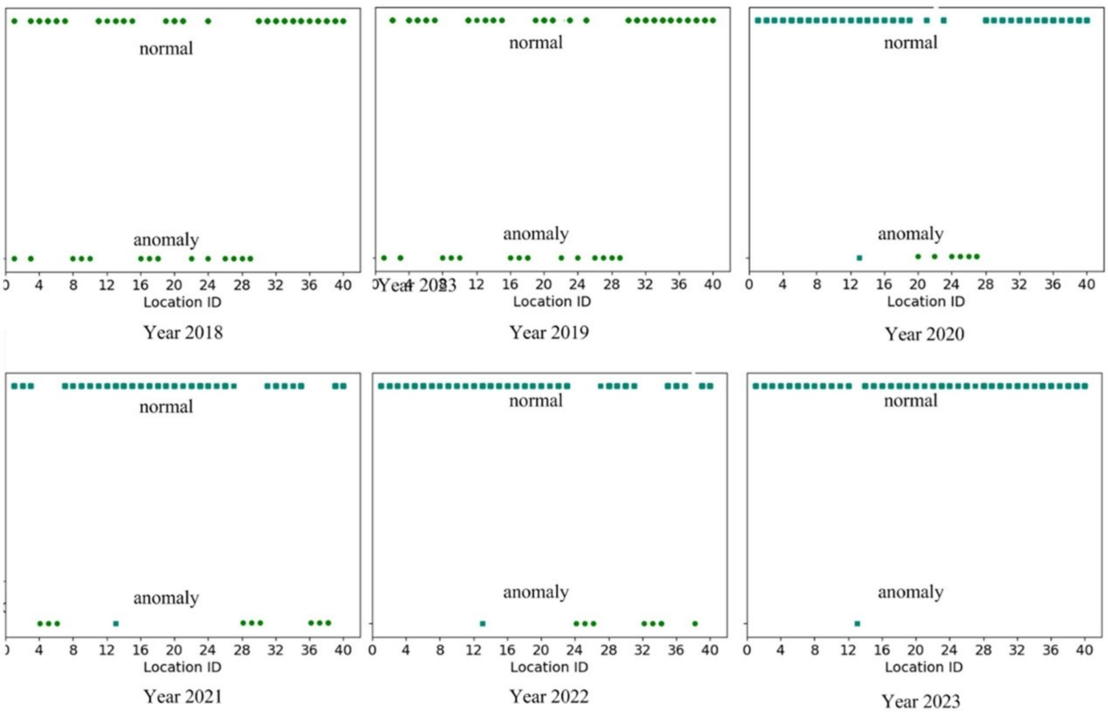
[1]Chongqing College of Mobile Telecommunications, Chongqing, 401520, China

[2]Chongqing Jiangbei International Airport Co.,Ltd. Chongqing, 401120, China

[3]Chongqing College of Mobile Telecommunications, Chongqing, 401520, China

E-mail: ytcquptli@163.com

Graphical Abstract



Abstract.

The Yangtze River, the longest river in China and one of the most important water resources in the world, has been facing significant challenges regarding water pollution in recent years. The issue of water quality safety is related to the national economy and people's livelihood. Water pollution incidents not only damage the local water environment, but also seriously affect the drinking water safety of residents. Traditional chemical methods and other water quality anomaly detection methods are often time-consuming and may cause secondary pollution. This paper proposed a machine learning method used for detecting abnormalities of water quality in the Yangtze River, to provide technical support for ensuring water quality safety. The principle is that using the designed support vector machine separates anomalies and normal values, by doing so, anomalies can be mined. Since there are certain differences between the density of normal data and that of anomalous data, estimating the data density of both can assist promote the ability of the support vector machine to mine anomalies. Then, the probability of water quality anomalies is determined by analyzing the characteristics such as the density of outliers in the sequence. Finally, using the collected the data of water quality from forty different regions of the Yangtze River since 2018 to 2023 as the experimental dataset, and experimental results show that the proposed method can effectively detect the anomaly of water quality of Yangtze River.

Keywords: anomalous detection, water quality

1 Introduction

Water pollution has become a global issue affecting human beings, plants, and animals. The rapid industrialization and urbanization processes have led to an increased discharge of pollutants into rivers, causing severe damage to ecosystems and threatening public health. The Yangtze River, as a major water source for millions of people, is no exception. As the largest river in China, the water quality of the Yangtze River directly affects the daily life and production activities of coastal residents. With the acceleration of industrialization and urbanization, the problem of water pollution in the Yangtze River is becoming increasingly serious.

To ensure the safety of the water supply, various methods are employed to detect abnormalities in the Yangtze River's water quality. Such as regular monitoring [1], i.e., government agencies and research institutions conduct routine inspections and sampling analyses to assess the water quality. These parameters such as pH, dissolved oxygen, turbidity, and heavy metal concentrations are closely monitored to detect any anomalies [2]. And modern techniques [3][4], such as remote sensing and online monitoring systems, are increasingly being employed to monitor water quality in real-time. These systems provide instant alerts when abnormal conditions are detected, allowing for swift intervention. In addition, local communities living along the Yangtze River also play a crucial role in detecting water quality abnormalities. They can report unusual phenomena, such as discoloration or unusual odors, to authorities, who can then take appropriate action. Therefore, conducting research on abnormal detection of water quality in the Yangtze River has important practical significance.

Contributions. The main contributions of this paper are summarized.

(1) We propose a support vector machine with the data density estimate, which has less dependence on data dimensionality. The proposed method not only suffers from less negative effects caused by data dimensionality, but also effectively detects the anomalies quality.

(2) The execution time of the proposed method is not exponentially impacted by data volume or by data dimensionalities, therefore, we demonstrate that it can be suitable for anomalous detection of large volume water quality.

Motivations. To detect water quality for Yangtze River, we designed a new form of support vector machine model using density estimate and to propose a novel anomaly detection approach based on the proposed model. To accurately detect anomalous section of pollute water, we used the proposed model find the monitor equipment to observe the pollute water region. Then, using the found monitor equipment to detect corresponding anomalous section of pollute water. By doing so, the water quality for Yangtze River can be accurately detected.

This paper is arranged as follows. Section 2 reviews the related work. Section 3 systematically describes the proposed method and the corresponding model. Experimental details and results are illustrated in Section 4. Section 5 draws a conclusion and directs future work.

2. Related work

There exists a rich literature on devising anomalous detection approaches, given that existed literature, we focused on investigating existing works that are mostly related to ours, i.e., support vector machine architectures-based detection approaches, which are regarded as the classification of normal data and anomalous data. Those detection approaches based on support vector machine (SVM) architectures are accustomed to utilize historical data (i.e., training data) to train the detectors and then verify new collections (i.e., testing data) by using the trained detectors [5]. In fact, SVM's approaches are a shallow-architecture approach (compared to deep learning approaches), therefore, in anomalous detection, scholars usually tend to optimize them or combine them with deep learning approaches, thus improving the results of anomalous detection. For instance, Ruff et al. [6] proposed a deep support vector data description (Deep SVDD) model, and the Deep SVDD employed by the [7], which of both obtain superior detection results. However, they need to solve a quadratic programming problem. And the sphere-based one-class support vector machine (S-OC-SVM) proposed by Andrews [8] et al. The [9] proposed a one class-support vector machine (OC-SVM) method for anomalous detection, which is capability to obtain an optimal decision model for the support vector data description, and to avoid under-fitting and over-fitting to a certain extent. Similarly, the [10] designed a quarter-sphere one-class support vector machine (Q-S-OC-SVM) method, which converts the quadratic programming problem into a linear programming problem. Although the detected efficiency of Q-S-OC-SVM is significantly augmented, it has to be retrained once a new testing data is coming. Additionally, also including the OC-SVM in [11] and in [12], which of them obtain advanced detection results. Unlike the models in [8-12], to solve a quadratic programming problem, Deng et al. [13] employed a one-class support Tucker machine to mine the anomalies in high dimensional environments, and experimental results show that it has good adaptability to high dimensional spaces.

The detected performance of those methods using SVM architectures is easily impacted by support vector machines. The shallow patterns possessed by support vector machines is likely to fail to capture those dependency relations between multiple variables [14]. To make up the disadvantage, Huang et al. [15] proposed a Least-Squares Support Vector Machine (L-S-SVM) model. Through utilizing the least squares, the detected performance of the support vector machine is significant promoted. Consequently, indeed, the above ideas indicate that introducing new forms (such as least squares, or one-class) can assist support vector machines. In addition to support vector machine methods, hypersphere methods are also used for anomalous detection, such as the hypersphere methods implemented in [16-18], whose ascendency does not have to perform matrix inverse operations.

3. Methodology

This section describes the thought of the method and the implement of the corresponding model. Given a sample $x = \{x_1, ..., x_i, ...,\}$, $i \geq 1$, to simplify, assuming that sample *x* does not noise and irrelevant attributes, the support vector machine is formally described as follows

$$\begin{aligned} &\min \ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i, x_j) - \sum_{j=1}^{N}\alpha_i \\ &s.t. \ \sum_{i=1}^{N}\alpha_i y_i = 0 \\ &\forall i = 1, 2, ..., \\ &0 \leq \alpha_i, \alpha_j \end{aligned} \tag{1}$$

Where α_i , α_j is Lagrange multiplier. *N* is the number of *x*. $y_i, y_j \in \{0, +1\}$, where 0 is an anomalous label, and +1 is a normal label. According to Eq. (1), we can obtain the classification decision function, as follows

$$\begin{cases} f(x) = sign\left(\sum_{i=1}^{N}\alpha_i^* y_i K(x, x_i) + b^*\right) \\ b^* = y_j - \sum_{i=1}^{N}\alpha_i^* y_i K(x_i, x_j) \end{cases} \tag{2}$$

$\alpha_i^* > 0$ is a component of α_i . *K* is a kernel function.

The classification ability of support vector machine relies on classification decision function *f*(*x*). For kernel *K* in *f*(*x*), here, we use density estimate to fulfil it, which is a non-parametric method for estimating probability density functions. As follows

$$K(,) = \frac{1}{NB}\sum_{i=1}^{N}\kappa\left(\frac{x - x_i}{B}\right) \tag{3}$$

Where κ is a kernel. 0<*B* is a smoothing parameter, also called bandwidth. For the kernel κ , we chose Gaussian kernel, having that

$$\kappa(x, x_i) = \exp\left(-\frac{\| x - x_i \|^2}{2\sigma^2}\right) \tag{4}$$

σ^2 is variance of *x*. Additionally, we need to think about bandwidth *B* in Eq. (3). The estimation results of kernel functions vary greatly under different bandwidths, as shown in Fig.1. It can be seen that when the bandwidth is not fixed, its variation depends on the estimated position (balloon estimator) or sample point (pointwise estimator), which can generate an adaptive estimation. Given that, we used a certain range to tune bandwidth values according to the scale of the input sample.

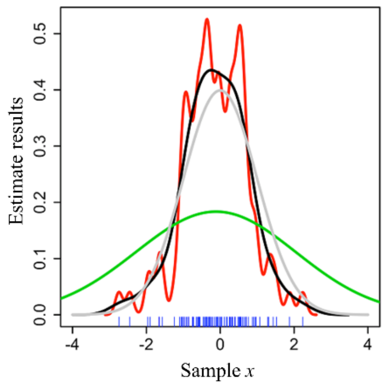


Fig.1 Density estimate with different bandwidths. We generated a random sample of 100 points from a standard normal distribution. Grey curve is true density (standard normal). Red curve is density estimate with bandwidth B=0.05. Black curve is density estimate with bandwidth B=0.337. Green curve is density estimate with bandwidth B=2.

Algorithm 1 is the corresponding algorithm of the proposed method. In algorithm 1, the input and output are dataset x, testing accuracy and predicted labels, respectively. Step 1 initializes model's parameters, and Step 2 fulfils the division of training set and testing set. The model is trained in the procedure between Step 3 and Step 16. For each point in training set x_{train}, we use decision function $f(x)$ in Eq. (2) and the density estimate in Eq. (3) to judge them, as shown in the procedure between Step 4 and Step 7. If the point falls inside the hyper plane learned by the support vector machine (SVM), it is regarded as a normal point and is assigned a normal label +1. Overwise, the point is considered as an abnormal point and is assigned an abnormal label 0, illustrate in Step 8 to Step 11. The training is terminated until all points in training set x_{train} are determined, and then we save the training accuracy and the trained model *SVM-DE*(x_{train}), as shown in Step 12 to Step 16. Thereafter, using testing set x_{test} to verify the trained model *SVM-DE*(x_{train}). Finally, the testing accuracy and predicted labels {+1,…, 0,…, +1,…, 0} are outputted, illustrated in Step 17 to Step 22.

Time complexity. The time consumption of Algorithm 1 consists of the running time of SVM and the calculation time of data density. Assuming that the data volume and data dimension of input data are V and D, respectively. The running time $O(SVM)$ of SVM is equal to $O(V*D*\Pi)$, where item Π is the number of SVM. In Algorithm 1, there used a single SVM, therefore, $\Pi=1$, and $O(SVM)=O(V*D)$. the calculation time $O(d)$ of data density is $O(V*D)$, i.e., $O(d)=O(V*D)$. The running time $O(n)$ of Algorithm 1 is $O(V*D)+O(V*D)$, that is, $O(n)=O(V*D)$.

Algorithm 1. SVM-DE.

Input: dataset x.
Output: testing accuracy, predicted labels {+1,…, 0,…, +1,…, 0}.

1 Initialization model's parameters;
2 x is randomly divided into a training set x_{train} and a testing set x_{test};
3 **Foreach** x_i **in** x_{train}**:**
/* training */
4 using decision function $f(x)$ in Eq. (2) to calculate point x_i;
5 $f(x_i) \leftarrow f(x) = sign\left(\sum_{i=1}^{N} \alpha_i^* y_i K(x, x_i) + b^*\right)$;

using Eq. (3) to estimate the data density ;
$data\ density \leftarrow \sum_{i=1}^{N} \kappa(\frac{x - x_i}{B}) / NB$;
If point x_i falls inside the hyper plane of our SVM **then :**
x_i is regarded as a normal point and is assigned a normal label +1 ;
Else:
x_i is considered as an abnormal point and is assigned an abnormal label 0 ;
If training set x_{train} is an empty **then :**
/* all points are determined */
break ;
End Foreach
saving training accuracy;
saving the trained model *SVM-DE*(x_{train});
Foreach x_j **in** x_{train} **:** /* testing */
using *SVM-DE*(x_{train}) to judge point x_j ;
predicting the label of point x_j ;
End Foreach
saving the testing accuracy ;
saving predicted labels {+1,…, 0,…, +1,…, 0} ;

4 Experiments and results analysis

4.1 Experimental settings

We used the monitoring equipment to collect water quality data on forty different regions of the Yangtze River from 2018 to 2023 as the experimental dataset, illustrated in Table 1, with 70% used for model training and the remaining 30% used for model validation. Additionally, Accuracy and F1-score which are regarded as the evaluated metrics in anomalous detection are used to assess detected results. As follows

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{5}$$

$$\text{F1-score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{6}$$

Where TP is that the model correctly predicts the number of in anomalous data. TN is that the model correctly predicts the number of normal data. FP shows that the model predicts normal data as the number of anomalous data. FN shows that the model predicts anomalous data as the number of normal data.

Apart from our model, we also chose the four detection models based on SVM architectures as competitors, i.e., Deep SVDD [6], S-OC-SVM [8], OC-SVM [9] and L-S-SVM [15]. To obtain a fair comparison, the comparative objects are selected based on the same design structures. We implemented corresponding algorithms of the five models (our model and the four competitive models) using Python on Tensorflow framework in Linux Operation System. To test the statistical significance of the difference between them, the Wilcoxon-test was adopted. Additionally, average ranks of the five algorithms are calculated by $(\sum_{i=1}^{N} r_i^j)/N$, where r_i^j is the ranking of *j*-th algorithm on *i*-th dataset.

Table 1 Details of the six datasets.

#	Year	Number of monitoring indicators (data dimension)	Collect data volume	Number of monitoring equipment	Number of monitoring regions
W1	2018	310	6900000	690	40
W2	2019	560	6100000	620	40
W3	2020	280	5600000	590	40
W4	2021	200	7300000	730	40
W5	2022	170	4500000	470	40
W6	2023	260	5400000	510	40

4.2 Result analysis

(I) Comparisons of detection performance

Detected results of on the six datasets are given in Table 2, showing that SVM-DE wins over the four competitors on most datasets. In terms of the metric Accuracy, our SVM-DE obtains the best performance on the four datasets W1-W2 and W4-W5. While for the metric F1-score, our SVM-DE also defeats the four competitive models on the five datasets W1-W5. Fig. 2 displays the detected results of our method on the forty regions.

Average ranks of each algorithm are given in the first row in Table 2. Though observing and analyzing, we find that SVM-DE obtaining the best average ranks is statistically better than the four comparison algorithms at the 95% confidence level. Moreover, there are no differences between the five algorithms for these detection results.

Fig. 3 unveils the relations among the detection ability of our SVM-DE, data volume and data dimensionality. Through comparing Fig. 3 (a) with Fig. 3 (b), in terms of SVM-DE, it can be seen that the correlation between the detection capabilities and data dimensionality are weaker than that between the detection capabilities and data volume, where the former is 0.3324 and the latter is 0.6226 (weak correlation). This also further indicates that SVM-DE has less dependence on data dimensionality in the process of anomalous detection. In summary, that is why our model defeats the four competitive models.

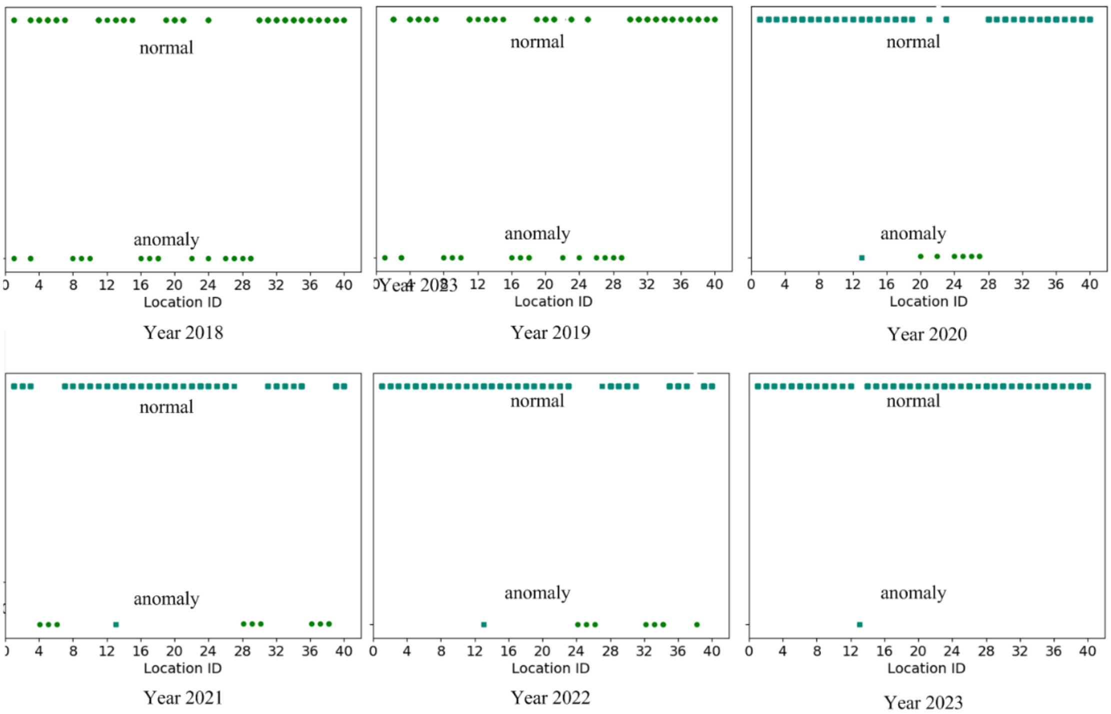


Fig. 2 Detected results of our method on forty different regions.

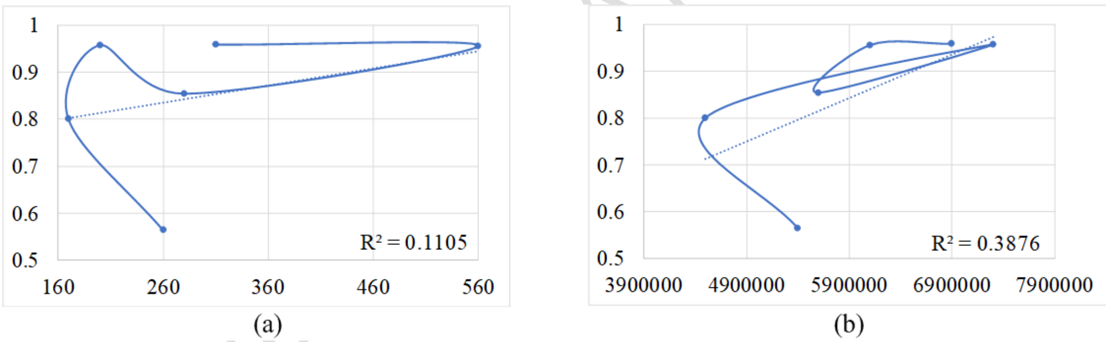


Fig. 3 Correlations among detection ability, data volume and data dimensionality. (a) displays the correlations of detection ability and data dimensionality. (b) displays the correlations among detection ability and data volume.

Table 2. Average values of 10 cross-validation on the six datasets. Best results are highlighted in bold font. Average ranks are given in the first row. *p*-value is given in the last row, and the sign '*' shows significant at *p*=0.05 level.

Method	SVM-DE	Deep SVDD [6]	L-S-SVM [15]	S-OC-SVM [8]	OC-SVM [9]
	metrics	Accuracy,	{F1-score}		
Average ranks	1.333	1.833	1.500	1.667	1.833

W1	**0.958, [0.979]**	0.899, [0.912]	0.902, [0.901]	0.900, [0.921]	0.878, [0.918]
W2	**0.955, [0.977]**	0.942, [0.953]	0.909, [0.945]	0.932, [0.927]	0.911, [0.877]
W3	0.854, **[0.921}**	**0.901**, [0.906]	0.887, [0.913]	0.827, [0.823]	**0.901**, [0.905]
W4	**0.957, [0.978]**	0.912, [0.907]	0.888, [0.955]	0.933, [0.919]	0.900, [0.802]
W5	**0.800**, **[0.889]**	0.771, [0.806]	0.797, [0.883]	0.727, [0.803]	0.701, [0.855]
W6	0.565, [0.722]	0.777, [0.799]	**0.807**, [0.779]	0.801, **[0.863]**	0.711, [0.835]
p=0.05	*	*	*	*	*

Table 3. Average time (second) of 10 cross-validation on the six datasets. Best efficiency is highlighted in bold font.

#	SVM-DE	Deep SVDD [6]	OC-SVM [9]	L-S-SVM [15]	S-OC-SVM [8]
W1	**1.539**	5.119	7.721	4.369	9.420
W2	473.547	**81.662**	80.889	82.889	117.818
W3	2.084	3.658	**1.658**	3.074	5.772
W4	**0.125**	1.505	0.788	1.103	1.552
W5	74.471	155.341	88.116	**72.438**	122.549
W6	**2.160**	7.166	11.558	25.769	14.286

(II) Comparisons of efficiency

This section main discusses the running efficiency of our algorithm and the four comparison algorithms. Here, we first analyzed the effects of data volume and data dimensionality on the running efficiency of our algorithm. We find that data volume has more effects on the execution time than data dimensionality does, of which the correlation for the former is 0.7076 (i.e., strong correlation) and that for the latter is 0.0804 (weak correlation). Table 3 gives the execution time of the five algorithms, showing that our algorithm wins the four competitive algorithms on part datasets. The execution time of our algorithm is not exponentially increased as data volume augments. These mean that detection efficiency of our algorithm is not exponentially impacted by data volume or data dimensionality. For our algorithm, the density estimate of the data needs to spend time cost, especially on large-scale datasets, hence, this is main consumption of our algorithm. As for the four competitive algorithms, the execution time is related to the complexity of their architectures.

5. Conclusion

To detect the water quality of Yangtze River, this paper proposes a novel support vector machine method with data density estimate. The critical thought is that we constructed a support vector machine to separate anomalous data and normal data, thus fulfilling anomalous detection. To accurately detect anomalous data, through estimating the data density, the support vector machine obtains the advanced detection results. Since the density between normal data and anomalous data shows certain differences, estimating data density is effectively in anomalous detection. Experimental results show that the proposed method defeats the competitors in detection accuracy of water quality. Results also indicate that the detection efficiency of the proposed method is not exponentially impacted by data volume or data dimensionality, which means that it is suitable for anomalous detection of large volume water quality. In future work, we will explore more intelligent detect methods used for anomalous detection to water quality.

Contributions

Jingyi Li proposed the method and wrote the paper. Shiwei Chao and Xu Zhang performed the source codes. Jingyi Li, Shiwei Chao and Xu Zhang designed the experiments and analyzed the results.

Declaration

There are no conflict interests of the authors.

Data availability

The data can be allowed be used.

References

[1] Ziwen Yu, Yifu Sheng, Hao Li, et al. Water Quality Classification Evaluation based on Water Quality Monitoring Data. 2023 11th International Conference on Information Systems and Computing Technology,2023, IEEE, pp.1-7.

[2] Manikannan Govindasamy, K. Jayanthi, S. Rajagopan. IoT Product on Smart Water Quality Monitoring System (Iot Wq-Kit) for Puducherry Union Territory. 2023 Second International Conference on Advances in Computational Intelligence and Communication, 2023, IEEE, PP. 1-10.

[3] Saira Varghese, Sreela Sreedhar, Janaki S Nair, et al. Portable Frmaework For Real time Water Quality Assessment and Analytics. 2023 4th International Conference for Emerging Technology, 2023, IEEE, pp.1-14.

[4] Boyu Zhang, Shanlin Sun, Yishan Su, et al. Surface Water Quality Monitoring System Based on Autonomous Underwater Vehicles.2023 3rd International Conference on Electrical Engineering and Control Science, 2023, IEEE, pp.1-17.

[5] Yan Qi, Kui W, Peng J. Efficient anomaly detection for high-Dimensional sensing data with one-class support vector machine. IEEE Transactions on Knowledge and Data Engineering. 2021;35(1):404-417.

[6] Ruff L, Vandermeulen R, Goernitz N, Deecke L, Siddiqui, Binder, Muller, Kloft. Deep one-class classification. in International conference on machine learning. 2018:4393-4402.

[7] Ying K, Hao W, Zhihua Z, Yaxing L, Jin M. DL-Based Anomaly Detection at the Physical-Layer of

Cognitive Radio by Deep Support Vector Data Description. IEEE Transactions on Cognitive Communications and Networking. 2022; 8(4):1689-1705.

[8] Andrews J T, Morton E J, Griffin D L. Detecting anomalous data using auto-encoders. International Journal of Machine Learning and Computing. 2016; 6(1):21.

[9] Z. H, C. W, Gh. Li. Outlier detection in wireless sensor networks using model selection-based support vector data descriptions. Sensors. 2018; 18(12):4328.

[10] Rajasegarar S, Leckie C, Palaniswami M, Bezdek. Quarter sphere based distributed anomaly detection in wireless sensor networks. In IEEE International Conference on Communications, 2007.

[11] Waqas Rasheed, Tong Boon Tang. Anomaly Detection of Moderate Traumatic Brain Injury Using Auto-Regularized Multi-Instance One-Class SVM[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2020, 28(1):83-93.

[12] Yonghyeok Ji, Hyeongcheol Lee. Event-Based Anomaly Detection Using a One-Class SVM for a Hybrid Electric Vehicle[J]. IEEE Transactions on Vehicular Technology, 2022, 71(6):6032-6043.

[13] X. D, P. J, X. P, C. M. An intelligent outlier detection method with one class support tucker machine and genetic algorithm toward big sensor data in internet of things. IEEE Transactions on Industrial Electronics. 2019; 66(6):4672-4683.

[14] Bengio Y, Y. L. Scaling learning algorithms towards AI. Large Scale Kernel Machines. 2007:1-41

[15] Huang, C.H, G. Z, X. L, G. X, L. S, K. D. Data Processing Method of Multibeam Bathymetry Based on Sparse Weighted LS-SVM Machine Algorithm. IEEE Journal of Oceanic Engineering. 2020; 45(4):1538-1551.

[16] Xu Y. Maximum margin of twin spheres support vector machine for imbalanced data classification. IEEE Transactions on Cybernetics. 2017; 47(6):1540-1550.

[17] Mei B, Xu Y. Multi-task least squares twin support vector machine for classification. Neurocomputing. 2019; 338:26-33.

[18] Qing A, Anna W. A novel feature weighted twin-hypersphere support vector machine for pattern recognition. 2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS), IEEE. 2018: 1-10.