Anomaly Detection for Water Supply Data using Machine Learning Technique

Shu Fang¹, Weize Sun²* and Lei Huang³

¹,²,³College of Electronics and Information Engineering, Shenzhen University, China

*Corresponding author: wsun@szu.edu.cn

Abstract. The advent of the era of big data brings new challenges and opportunities to water data processing. Abnormal detection or authenticity verification of water supply data becomes an urgent problem in natural resource data processing. This article addresses the issue of anomaly detection for a set of measured water supply data that provided by the Ministry of Water Resources of China. The machine learning technique, namely, One-Class support vector machine, is used for anomaly detection, and a detailed analysis of the features of anomalies is also provided. Experiments on three situations, namely, one dimensional situation, three dimensional situation and seven dimensional situation are carried out to analyse the features of anomalies. The experiment results revealed that the cases with entries of large values more than 10 times of median, zero entries together with some large values and ascending or descending sequences tend to have high abnormality scores.

1. Introduction

Water resources are one of the most essential natural and strategic resources that support the development of social economic systems. With the significant increase of social and economic water supply requirement, monitoring the behavior of these users is becoming more and more important. Although a lot of techniques and algorithms had been developed to record the water supply data, methods for judging the authenticity of the data or the quantity of water these users consumed is still unsatisfying. With the development of information technology and the gradual improvement of water management facilities, water data processing, especially anomaly detection of water supply data, has attracted attentions of many scholars [1] [2].

Due to the different measurement and recording methods of different water treatment departments, water supply data may have the problems of inconsistent specifications or sometimes missing values, making some part of the data becomes particularly abnormal. For better management of water supply data, it is necessary to take actions on authenticity checking. The authenticity of water data refers to the process of judging whether the observations are recorded correctly or not. Usually the part of observations deviates very much from other observations and thus called anomalies [3], which makes it difficult for us to obtain valid information correctly from water supply data.

To overcome this problem, researchers have proposed many different methods. In [4], the clustering algorithm named ADWICE of n-dimensional data space is used for real-time anomaly detection in water management system. When the quality indicator changes indicate abnormal conditions, an alarm is generated. Moreover, [5] introduces an urban water network anomaly flow detection method based on functional data analysis (FDA) and the results are very promising, and [6] models the normal system behavior as a hierarchical structure of hidden semi-Markov models and proposes a method based on SCADA for water supply system anomaly detection. Furthermore, a
software tool is introduced in [7] to detect relatively frequent events such as pipe burst and water discoloration in water distribution systems.

The problem of anomaly detection also lies in many other applications, such as urban traffic system anomaly detection [8], cloud computing and fog computing [9][10], network intrusion and wireless sensor network [11][12]. In anomaly detection, one class of classification (OCC) problem is usually accompanied by the case where the negative class usually does not exist or is not correctly sampled [13]. OCC problem is very common in real world applications than multi-classification and many researches has already developed several successful algorithms in various fields in science and engineering, such as text classification [14][15]. Among these, the machine learning method, One-Class support vector machine (OC-SVM), is proved to be very efficient in anomaly detection for high dimensional sample dataset [16]. OC-SVM tries to find the optimal boundary by training the dataset to distinguish normal and abnormal values. There are many researches on this method and its variations, which had been proved to be useful in many applications [17][18]. In this paper, based on water data analysis, we apply OC-SVM for authenticity checking and analyze the characteristics of anomalies in water supply data.

The structure of the remainder of this paper is organized as follows. Section II describes the principle of anomaly detection of OC-SVM and its implementation. Section III gives a briefly introduction to the water supply dataset and the data preprocessing techniques, where the training process is also discussed in details. The experimental results of different cases for abnormal detection is shown in Section IV, together with a detailed analysis of the top abnormal data. Finally, a conclusion is drawn in Section V.

2. Anomaly detection with OC-SVM

In probability, data have their own hidden distribution pattern, which can be characterized by a probabilistic model. Representation of probabilistic model in p-dimensional space learned from raw input data is a tightly constraint boundary shape. Through this boundary shape, different samples can be clearly classified. It is well known that support vector machine (SVM) finds an optimal hyper-plane to deal with multi-class classification problem. In OC-SVM, an optimal hyper-sphere is required to discriminate one class from other probable classes. As shown in [19], the strategy of OC-SVM is to map the data into a feature space using a kernel function, then find an optimal hyper-sphere as small as possible while including most of data. This can be considered as an optimization problem and the objective function is expressed as follows:

\[
\min_{R \in \mathbb{R}, \xi \in \mathbb{R}^\ell, \epsilon \in \mathbb{F}} \frac{R^2}{\nu} + \frac{1}{\nu} \ell \sum \xi_i s.t. \quad \|\Phi(x_i) - \epsilon\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \ldots, \ell.
\]

(1)

where \( R \) is the radius of the ball, \( \nu \in (0, 1) \) determines the upper limit of the anomalies and the lower bound of support vector samples, which controls the radius of the hyper-sphere. When \( \nu \) is sufficiently small, the size of the “ball” should be cut down. \( \chi = [x_1, x_2, \ldots, x_\ell] \) is the training dataset. \( \Phi \) is the feature map from input dataset to feature space \( F \), and \( \epsilon \) is the center of the “ball”. \( \xi_i \) is a nonzero slack variable penalization. After the using Lagrange multipliers, we can get the decision function as follows:

\[
f(x) = R^2 - \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) + 2 \sum_i \alpha_i k(x_i, x) - k(x, x), \quad i, j = 1, 2, \ldots, \ell.
\]

(2)

where \( 0 \leq \alpha_i, \alpha_j \leq \frac{1}{\nu} \) is the Gaussian RBF kernel, and \( k(x_i, x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma^2} \). This kernel function measures the similarity score between one sample and the others in the dataset. Data with higher score are more likely to be the target, or the normal data in this model.

OC-SVM is an unsupervised anomaly detection method and usually used in the case where the number of positive samples are much larger than that of the negative samples. In our experiment, we proposed to calculate the score to measure the similarity of each case to full dataset and use this score to analyze the top abnormal cases.
3. Dataset and preprocessing
The dataset was collected from 204 different sources and each source contains water quantity data of 738 days. There are totally 150,546 entries of daily water quantity data for the dataset. The distributions of water quantity values for the 204 water sources are significantly different from one and the others. For example, the water source “4408400001001” showed the least water quantity with a mean of 51.48835 and a standard deviation of 91.02929, whereas the water source “4403030001001” showed the highest water quantity with a mean of 4679111.52846 and a standard deviation of 1664651.68455. To this reason, an extremely large value for one water source (abnormal) might be quite normal for another. Therefore, a normalization step is required for the analysis. Here the data is normalized based on the median value of each water source. Let \( \mu_i \) be the median water quantity value of the \( i \)-th water source. For any given water quantity value \( x \) from the source, \( x \) is normalized into \( y = \frac{x}{\mu_i} (i = 1, 2, ..., 204) \). Based on this normalization, a water quantity value that is equal to the median of the source will be transformed to 1, a water quantity value that is zero will be still 0, and a water quantity value that is \( N \) times larger than the median will be transformed to a value of \( 0 \). It is notified by the domain experts that an entry of water quantity might be abnormal if it is zero or larger than five times of the median, however, the validity of this logic is required to check.

![Figure 1. Distribution of normalized water quantity values sorted based on the normality scores.](image1)

![Figure 2. Distribution of normalized water quantity values sorted based on the normality scores.](image2)

4. Experimental results
For the water supply dataset, experiments are conducted by using the normalized water quantity values as the features (one dimensional situation), the set of continuous three days water quantity values as the features (three dimensional situation), and the set of continuous seven days (a week) water quantity values as the features (seven dimensional situation).

To facilitate the following discussion, we define a case as an element of the \( N \) dimensional feature set, therefore, a case contains \( N \) values, each of which is called an entry. All cases are used in both the training dataset and the evaluation dataset. In the evaluation procedure, for each case, a score that measures the similarity of the case to the class (the full dataset) is calculated. The inverse of this score measures the abnormal levels, which is termed as the abnormality score. We sort the dataset based on the abnormality scores, and then analyze the top abnormal cases in this article.
4.1. One dimensional situation

Let the discussion starts from the one dimensional situation. For the top nine most abnormal cases, the normalized water quantity values are all over 200, meaning that they are all extremely large values that are more than 200 times of the medians. There are 109 entries over 100 in the dataset and they are all listed within the top 200 abnormal cases. The top 742 cases are those with the entries over 10. It is also interesting to note that the number of entries over 10 is exactly 742 in the dataset. Table 1 shows the normalized water quantity distribution over the top 200 abnormal cases. It can be seen that all the 109 entries over 100 are within the top 200 abnormal cases. The 21 entries between 50 and 100 are all listed within the top 100 abnormal entries.

Figure 1 and figure 2 illustrate the distribution of the normalized water quantity values ordered by the abnormality scores. The left end refers to the cases with the highest abnormality scores, while the right end refers to the cases with the lowest abnormality scores. The vertical coordinate illustrates the normalized water quantity values. From figure 1, we can see that the extremely large values (over a hundred) concentrate on the left end. By setting the maximum vertical coordinate to 10, we get the figure 2, from which, it can be observed that the points associated with water quantity values larger than 2 are located in the left section, followed by points with zero values. Then points from around 2 gradually decrease while the points from 0 gradually increases to approximately 1. For the large section on the right part of the chart, there are points with the normalized water quantity values around 1.

Table 1. (Normalized) water quantity distribution over the top 200 abnormal entries.

|                | Top 50 | Top 51-100 | Top 101-150 | Top 151-200 |
|----------------|--------|------------|-------------|-------------|
| [100, +∞)     | 32     | 27         | 42          | 8           |
| [50, 100)     | 16     | 5          | 0           | 0           |
| [10, 50)      | 2      | 18         | 8           | 42          |

Table 2. Three dimensional cases

|                | Overall | Top 114 | Top 115-323 |
|----------------|---------|---------|-------------|
| Number of cases that contain 0 | 7574(5.03%) | 23(20.18%) | 11(5.29%)   |
| Number of cases that contain entries > 100 | 151(0.10%) | 67(58.77%) | 75(36.6%)   |

4.2. Three dimensional situation

For the three dimensional situation, the top 114 abnormal cases have the same abnormality score, while the top 323 abnormal cases contain at least one entry larger than 10 (ten times of the median). Table 2 summaries some observation of the results, showing that the most abnormal cases tend to have zero entries or entries larger than 100. For the top 114 abnormal cases, 67 out of the 114 cases contains at least an entry over 100, the subset of top 115 to top 323 abnormal cases contains 75 such cases. This has indicated that the 142 out of the 151 cases with at least an entry over 100 are within the top 323 cases. For the top 114 cases, there are 23 cases containing at least an entry of zero. The percentage of such cases in this group is 20.18%, much higher than 5.03%, which is the percentage of cases containing at least one zero in the full dataset. We have also observed that the top 10 most normal cases are all with entries between 0.95 and 1.

4.3. Seven dimensional situation

In this seven dimensional situation. There are 62 cases with all entries over 100 and they are all ranked within the top 454 cases, and totally 233 cases with at least an entry larger than 100, which are all listed in the top 1424 (0.95%) abnormal cases. The top 319 abnormal cases have the same abnormality score. Table 3 shows statistical information for these cases compared to that of the full dataset. It can be observed from table 3 that all the 319 cases contain at least some entries larger than 10 (ten times of the median). Around half of them are cases that contain at least some entries over 100. In particular, there are 78 cases with all entries over ten, and 41 of them with entries all over 100. In the full dataset,
there are around 8.57% cases containing some zero entries, whereas in the top 319 abnormal cases there are 34.48% (110) such cases.

It has been known that each case contains entries for continuously seven days water quantity values. The cases shown in table 4 are ascending sequences that mean the value increases day by day, whereas the cases shown in table 5 are descending sequences that mean the value decreases day by day. To avoid the problem of division by zero, $\mu_i$ will be set to a small positive value when it is equal to zero. In the dataset, four cases are ascending sequences, and fourteen cases are descending sequences. They all ranked within the top 5% most abnormal cases. In summary, we can get the following conclusion based on the above analysis:

- the cases with entries of large values (over 10 times of median) tend to have high abnormality scores.
- the cases with zero entries together with some large values tend to have high abnormality scores.
- the cases with ascending or descending sequences tend to have high abnormality scores.

### Table 3. Seven dimensional cases

|                         | Overall (150546 cases) | Top 319 abnormal cases |
|-------------------------|------------------------|------------------------|
| Number of cases that contain 0 entries | 12903 (8.57%) | 110 (34.48%) |
| Number of cases that contain entries > 10 | 1422 (0.94%) | 319 (100%) |
| Number of cases that contain entries > 100 | 233 (0.15%) | 150 (47.02%) |
| Number of cases with all entries > 10 | 426 (0.28%) | 78 (24.45%) |
| Number of cases with all entries > 100 | 62 (0.04%) | 41 (12.85%) |

### Table 4. Ascending sequence

| day1 | day2 | day3 | day4 | day5 | day6 | day7 | Abnormality rank |
|------|------|------|------|------|------|------|------------------|
| 0.90 | 1.01 | 1.12 | 7.94 | 12.33 | 12.48 | 12.50 | 784 (top 0.52%) |
| 0.38 | 0.46 | 0.92 | 1.05 | 1.29  | 4.08  | 11.99 | 1413 (top 0.94%) |
| 1.31 | 1.62 | 2.08 | 2.77 | 3.19  | 4.04  | 5.38  | 1870 (top 1.24%) |
| 0.76 | 0.77 | 0.82 | 0.86 | 1.60  | 4.35  | 4.37  | 3160 (top 2.1%)  |

### Table 5. Descending sequence

| day1 | day2 | day3 | day4 | day5 | day6 | day7 | Abnormality rank |
|------|------|------|------|------|------|------|------------------|
| 385.40 | 175.91 | 169.34 | 156.93 | 149.64 | 130.66 | 129.19 | 319 (top 0.21%) |
| 388.46 | 173.08 | 170.77 | 156.15 | 144.62 | 135.38 | 124.61 | 1413 (top 0.94%) |
| 27.11 | 25.72 | 25.37 | 24.18 | 23.97 | 22.36 | 21.17 | 677 (top 0.45%) |
| 12.13 | 12.10 | 11.79 | 11.67 | 11.36 | 10.38 | 0.44 | 1064 (top 0.71%) |
| 12.50 | 12.40 | 12.10 | 12.00 | 10.79 | 1.16 | 0.91 | 1127 (top 0.75%) |
| 18.70 | 1.00 | 1.04 | 0.98 | 0.92 | 0.91 | 0.90 | 2010 (top 1.34%) |
| 7.63 | 1.35 | 1.33 | 1.32 | 1.30 | 1.27 | 1.23 | 4238 (top 2.82%) |
| 7.58 | 1.33 | 1.33 | 1.29 | 1.29 | 1.24 | 1.24 | 4275 (top 2.84%) |
| 7.05 | 1.21 | 1.20 | 1.15 | 1.09 | 1.07 | 1.04 | 5071 (top 3.37%) |
| 6.53 | 1.06 | 1.06 | 1.04 | 1.00 | 0.98 | 0.96 | 5747 (top 3.82%) |
| 6.06 | 1.15 | 1.15 | 1.14 | 1.11 | 1.02 | 1.00 | 5998 (top 3.98%) |
| 4.81 | 1.25 | 1.22 | 1.20 | 1.16 | 1.14 | 1.14 | 6694 (top 4.45%) |
| 4.64 | 1.22 | 1.19 | 1.16 | 1.12 | 1.10 | 1.10 | 6984 (top 4.64%) |

### 5. Experimental results

In this paper, we discussed the problem of authenticity checking of water supply data and applied the OC-SVM to solve this problem. The water supply dataset was firstly discussed and then the preprocessing step was carried out. The one-dimensional, three-dimensional and seven-dimensional situations were discussed, and the abnormal data analysis was performed respectively. Generally speaking, the cases with entries of large values, zero entries as well as ascending or descending sequences tend to have high abnormality scores, which means that they are more likely to be abnormal data.
Acknowledgment
The work described in this paper was supported by the National Natural Science Foundation of China (NSFC) under Grant No. U1501253.

References
[1] M Beaufils and J M Follin 2011 Geospatial data fusion for water resources assessment The 14th AGILE Int. Conf. on Geographic Information Science (The Netherlands: Utrecht) pp 18-21
[2] A V Dijk and L J Renzullo 2010 Water resource monitoring systems and the role of satellite observations Hydrology and Earth System Sciences Discussions 7 6305-6349
[3] D M Hawkins 1980 Identification of outliers Springer chapter 1 pp 1–2
[4] M Raciti, J Cucurull and S Nadjm-Tehrani 2012 Anomaly detection in water management systems Critical infrastructure protection 7130 98-119
[5] L Millan-Roures, I Epifanio and V Martinez 2018 Detection of anomalies in water networks by functional data analysis Mathematical Problems in Engineering 2018 13
[6] Z Zohrevand, U Glasser, H Y Shahir, M A Tayebi and R Costanzo 2016 Hidden Markov based anomaly detection for water supply systems IEEE Int. Conf. on Big Data (USA: Washington DC) pp 1551–1560
[7] M Bakker, T Lapikas, B H Tangena and J H G Vreeburg 2012 Monitoring water supply systems for anomaly detection and response IWC Int. Water Conf. New Developments in IT and Water (The Netherlands: Amsterdam)
[8] X Kong, X Song, F Xia, H G, J W and A.Tolba 2018 LoTAD: Long-term traffic anomaly detection based on crowdsourced bus trajectory data World Wide Web 21 825-847
[9] A Nurunnabi, G West and D Belton 2015 Outlier detection and robust normal-curvature estimation in mobile laser scanning 3D point cloud data Pattern Recognition 48 1404-1419
[10] N Moustafa, K-K R Choo, I Radwan and S Camtepe 2019 Outlier Dirichlet Mixture Mechanism: Adversarial Statistical Learning for Anomaly Detection in the Fog IEEE Transactions on Information Forensics and Security PP 1-1
[11] T Bajtos, A Gajdos, L Kleinova, K Lucivjanska and P Sokol 2018 Network Intrusion Detection with Threat Agent Profiling Security and Communication Networks 2018 17
[12] A D Paola, S Gaglio, G L Re, F Milazzo and M Ortolani 2015 Adaptive Distributed Outlier Detection for WSNs IEEE Transactions on Cybernetics 45 902-913
[13] S S Khan and M G Madden 2009 A Survey of Recent Trends in One Class Classification Irish Conf. on Artificial Intelligence and Cognitive Science (Ireland: Dublin) pp 188-197
[14] B Liu, W S Lee, P S Yu and X L 2002 Partially supervised classification of text documents the Nineteenth Int. Conf. on Machine Learning (California: San Francisco) pp 387-394
[15] H Yu, J Han and K C C Chang 2004 PEBL: Web page classification without negative examples IEEE Transactions on Knowledge and Data Engineering 16 78-81
[16] B Scholkopf, J C Platt, J Shawe-Taylor, A J Smola and R C Williamson 2001 Estimating the support of a high-dimensional distribution Neural computation 13 1443-1471
[17] J T A Andrews, E J Morton, and L D Griffin 2016 Detecting anomalous data using auto-encoders International Journal of Machine Learning and Computing 6 21
[18] M Gutoski, N M R Aquino, M Ribeiro, A E Lazzaretti and H S Lopes 2017 Detection of video anomalies using convolutional autoencoders and one-class support vector machines XIII Brazilian Congress on Computational Intelligence (RJ: Rio das Ostras)
[19] Y Chen, X S Zhou and T S Huang 2001 One-class SVM for learning in image retrieval Proc. of 2001 Int. Conf. on Image Processing (Greece: Thessaloniki) pp 34-37