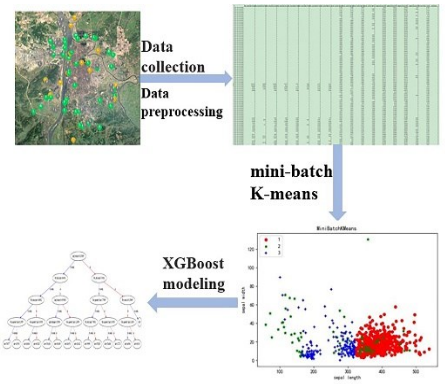
Global NEST Journal, Vol 25, No 4, pp 104-111
Copyright© 2023 Global NEST
Printed in Greece. All rights reserved

Drainage network flow anomaly classification based on XGBoost

Hong C.L., Jie T.Y.*, Peng L.J., Long M.S. and Jun H.
Faculty of Artificial Intelligence and Big Data, Hefei University, Hefei 230031, China
Received: 21/09/2022, Accepted: 14/12/2022, Available online: 24/01/2023
*to whom all correspondence should be addressed: e-mail: 1010423201@qq.com
https://doi.org/10.30955/gnj.004488

Graphical abstract



Abstract

Identifying and classifying anomalies in on-line monitoring systems of drainage systems is important to reduce urban water pollution. In the context of big data, the mini-batch K-means combined with the XGBoost drainage network abnormal flow identification and classification model is proposed to precisely identify and classify abnormalities that occur in real-time updates of online drainage network data while resolving problems with subjectivity and the lack of uniform standards for classification. First, using Mini Batch K-means, the unclassified drainage network data were sorted into four categories: normal drainage, sneaky drainage, rainwater and sewage mixing, and inflow and infiltration. Next, XGBoost performs data modeling to create a model for classification and identification of drainage flow anomalies. To increase the accuracy of the model, the features were ultimately chosen based on the ranking of the importance of the features, and the model parameters were established using grid search and cross-validation. The results showed that the XGBoost Drainage Network Anomaly Classification and Identification Model can accurately identify four drainage network situations with high classification accuracy and good performance. It was also validated through the application of data from the online system monitoring points in Changsha, China, in 2020.

Keywords: traffic anomaly, classification recognition, Mini Batch K-means, XGBoost algorithm

1. Introduction

Due to rapid socioeconomic development, the urban population has proliferated, and water consumption has increased, which, together with the old drainage network in some urban areas and the lack of systematic management, has exacerbated the contradiction between the current situation of drainage networks and water environment management (Lund *et al.*, 2018). The diverse causes of water pollution, the lack of basic data, the lack of positioning means, the inadequate scale of the drainage pipe network, the mixed connection of rain and sewage, inflow and infiltration, and sneaky drainage are the main problems that should be solved in future drainage systems (Qiu *et al.*, 2011; Li *et al.*, 2013; Wang *et al.*, 2021).

Traditional drainage network inspections, the bulk of which call for professionals to go to the monitoring site, are not only time-consuming but also susceptible to various variables, such as the environment and climate. (Jato-Espino *et al.*, 2022). It has real-time detecting technology that can gather information on sewage discharge to enhance the effectiveness of monitoring the drainage network, but the monitoring system needs manual supervision round-the-clock to identify anomalies. On the one hand, the huge amount of data creates a huge workload; on the other hand, manual monitoring relies too much on subjective experience, which inevitably leads to erroneous judgement. The emergence of machine learning has provided a favourable environment for the scientific management of drainage system management issues. The status of the drainage network is monitored using hardware devices, such as cameras to transmit the real-time status of the drainage network; temperature sensors to analyze the water flow temperature of the drainage network to monitor its drainage status; or by collecting sewer water for compound composition analysis to monitor abnormal water flow in the drainage network. The above methods are more theoretically supported by data than traditional manual monitoring but often require brand new electronic equipment and large laboratories to analyze the data, consuming high resources and requiring the cooperation of professional staff. Along with the rise of machine learning, machine learning has been gradually applied to drainage network monitoring, which can be used to simulate and predict possible abnormalities in drainage networks through the collection of data. These data can be adapted to local conditions and modeled based on the data provided by existing inspection equipment, eliminating the problems associated with the reinstallation of equipment (Kwon *et al.*, 2021).

This study aims to solve the problems that drainage network monitoring is affected by various factors and that there is no uniform standard for anomaly classification and to realize anomaly data identification and classification modeling on the existing monitoring data. The data provided by the monitoring system, combined with the unsupervised classification algorithm Mini Batch K-means, were used to classify the collected data into four categories: normal, sneaky drainage, rainwater and sewage mixing, and inflow and infiltration. Then, the XGBoost classification algorithm is used to establish the drainage network flow anomaly identification and classification model.

2. Related work

With the rapid development of modernisation, more and more researchers are using science and technology to transform people's material living conditions. This component of intelligent water services has advanced quickly in line with the overall trend of urban wisdom growth, and more information technology is being applied to the drainage area, eventually replacing low-tech governance with high-tech governance (Yazdanfar *et al.*, 2015; Zhou *et al.*, 2016).

Many academics domestically and abroad have monitored drainage networks using artificial intelligence and other techniques. Direct monitoring of the drainage network is an efficient and intuitive method with a high degree of real-time. Among them, Tatiparthi Sundra R and De Costa Yashika G (Tatiparthi *et al.*, 2021) used monitoring techniques to develop an intelligent monitoring system stroboscopic identification sensor applicable to sewer pipes that can track organic solids in sewer pipes and drainage pipes, among other things, and track the hydraulic and flow velocity in the water pipe network to detect blockage and seepage events in the pipes. This means addressing the traditional need for staff to visit the site of the problem, but also to provide timely feedback on problems that occur within the drainage system, but the need to cooperate with the system to install the stroboscopic identification sensor applied to water pipes. For the Old Town, a lot of material resources are needed to correct drainage monitoring. Jonathan M. AITKEN and MATHEW H. Evans (Aitken *et al.*, 2021), on the other hand, proposed a mobile inspection robotic to better monitor the state of drainage networks using a robot that can be positioned on a mobile basis. The aforementioned techniques, however, do not immediately use the data made available by the current online drainage network inspection systems and need the installation of monitoring equipment for the drainage network. Data and technology cannot be integrated, and data usage was weak.

Monitoring the drainage network can also be done by analyzing the acquired indirect data. Among them, Maryam Beheshti and Sveinung Sgrov (Beheshti and Sægrov, 2018) proposed quantifying the infiltration and inflow of foreign water by analyzing the thermal properties of the sewer network, utilizing light-distributed temperature sensing (DTS) to quantify foreign sewage, and thereby serving to monitor the sewer network regarding infiltration and inflow. V. Bareš, D. Stránský and P. Sýkora *et al.* (Bareš *et al.*, 2012) proposed the use of the COD quality assessment of the sewer infiltration/inflow flux method, which uses changes in chemical oxygen demand (COD) loads in combination with water quality and quantity to monitor waterways experiencing infiltration and inflow problems. Both of these methods also require the installation of relevant equipment for monitoring purposes and cannot be used effectively with existing data.

With the development of machine learning, methods of analytic data modelling are gradually emerging (L'Heureux *et al.*, 2017; Munawar *et al.*, 2021; Gandomi *et al.*, 2022). In order to better manage the stormwater pipe network and avoid urban flooding, Hao Wang and Lixiang Son (Wang and Lixiang, 2020) employed the SVM machine learning technology to estimate the water level of the system. Qiyun Zhu, April Gu (Qiyun *et al.*, 2021), and others then combined UV–visvisible spectroscopy with a derivative neural network algorithm for online monitoring and identification of urban drainage networks, which can accurately identify industrial wastewater, domestic sewage and rainwater mixing. However, it requires the installation of a spectral probe, which cannot be combined with the data provided by existing online monitoring systems for identification and classification. Hybrid neural networks have also been used to analyze the image data collected by CCTVs (Moselhi *et al.*, 1999;2000; Shehab *et al.*,2005) to classify the defects present in drainage networks. However, to better combine the available data, the XGBoost algorithm, which is suitable for analyzing numerical data, needs to be used.

The amount of training data is one of the key factors affecting the accuracy of the model since machine learning models are built by learning from and generalizing training data, and then applying the obtained applicable models to brand-new data for prediction or classification. According to the amount and type of data provided by the existing drainage network inspection system, Mini Batch K-means combined with XGBoost is selected to start from a large amount of monitoring data and parse the data to find the patterns and identify and classify the possible problems at the monitoring points.

3. Methodology

3.1. Mini batch k-means

Unsupervised learning techniques are effective and accurate for determining categories for a set of data when there are no established classification rules and there are vast volumes of data with high latitude and little fluctuation of data attributes. (Barlow *et al.*, 1989; Hastie *et al.*, 2009)

By repeatedly using the unsupervised clustering algorithm, we first classify a large amount of data into two categories, i.e., normal data and abnormal data, and then subdivide the abnormal data into three categories, checking the accuracy of data classification using line graphs and so on for the four categories. Finally, the four categories of data were given labels of normal, sneaky drainage, rainwater and sewage mixing, and inflow and infiltration.

Four unsupervised clustering techniques were chosen to handle the same data to test the accuracy of the unsupervised classification algorithms. The investigations demonstrated that the unsupervised clustering algorithm classification comparison chart displayed in Figure 1 is the same after the same number of unsupervised algorithm clustering, K-means; Mini Batch K-means; Agglomerative clustering; and Birch processing results. Choosing Mini Batch K-means as the data classification algorithm can save time when a large amount of data has to be processed.

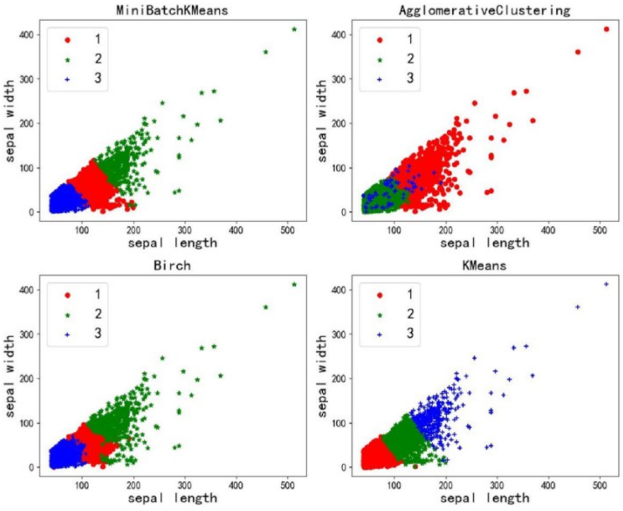


Figure 1. Comparison chart of unsupervised clustering algorithm classification

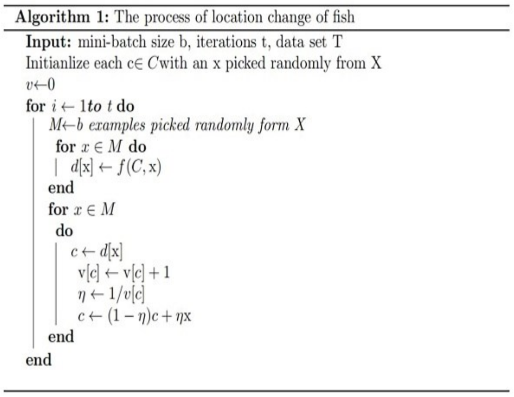
The unsupervised clustering algorithm represents the K-Means clustering algorithm, which serves to automatically group similar samples into one class. Assign each point to the class cluster closest to the initial class cluster centroid, given k initial class cluster centroids. After then, the class cluster centroids are reprogrammed, and the process of assigning points and updating class cluster centroids is repeated recursively until the class cluster centroids no longer change or the specified number of iterations is reached (Jain and Dubes,1988; Bock ,2007; Sinaga and Yang, 2020). When compared to the traditional K-means algorithm, the mini batch K-means algorithm optimizes the same objective function but uses small subsets of data that are randomly selected from the training set as the new training set during training, and most of these small batches reduce the amount of computation required to converge to a local, effectively reducing computation time (Béjar Alonso ,2013; Chavan *et al.*,2015).

Assuming that given a data sample set T contains n objects $T=\{x_1,x_2,x_3,\ldots,x_n\},x_i\in R$, where each object has m-dimensional features, the mini batch K-means algorithm aggregates n objects into the specified k class clusters based on the similarity of m-dimensional features, and each object exists in only one class cluster. The goal of the clustering problem is to find the set of clustering centers C, ($C\in R$) the set achieves minimization of the following objective function.

$$\min\sum_{X\in T}\left\|f(C,X)-X\right\|^2 \quad (1)$$

Among them, $f(C,X)$ returns the center of the nearest cluster, and $|C|=K$, K is the number of clusters we need.

Mini Batch K-means algorithm steps:



```
Algorithm 1: The process of location change of fish
Input: mini-batch size b, iterations t, data set T
Initianlize each c∈ $C$with an x picked randomly from X
$v \leftarrow 0$
for $i \leftarrow 1$to $t$ do
    $M \leftarrow b$ examples picked randomly form X
    for $x \in M$ do
        $d[x] \leftarrow f(C, x)$
    end
    for $x \in M$
     do
        $c \leftarrow d[x]$
          $v[c] \leftarrow v[c] + 1$
          $\eta \leftarrow 1/v[c]$
          $c \leftarrow (1-\eta)c + \eta x$
    end
end
```

3.2. XGBoost classifer

The current mainstream supervised learning algorithms include the random forest algorithm and the support vector machine algorithm. Small sample size, nonlinearity, and high-level pattern identification are all advantages of the support vector machine algorithm. The random forest technique combines many decision trees, and the decision trees vote to determine the final classification result, making it appropriate for multidimensional data and quick data processing. However, the number of aberrant data samples in the flow monitoring data provided by the drainage network's live monitoring system is significantly smaller than the number of normal data samples, which is a typical nonequilibrium dataset, and the gap between the data is not obvious. Therefore, a reasonable supervised learning algorithm is chosen to solve the problems of unbalanced samples and a small number of partial samples. XGBoost (eXtreme Gradient Boosting) is an open-source machine learning project developed by Tianqi Chen's team (Qiu *et al.*, 2016; Sodhi, Pinky *et al.*, 2019; Arora, 2020; Nti *et al.*, 2021) in 2016 that is widely used in the training of classification and regression models. The GBDT algorithm has been improved and optimized to effectively handle sparse data (Chen and Guestrin , 2016; Candice, 2021).

XGBoost is an additive model based on the idea of boosting integration using forward distribution for greedy learning. The basic idea is as follows:

3.2.1. Additive model

The primary purpose of integrated learning is to train many models iteratively and merge these models in a certain way to build a powerful integrated model. XGBoost's purpose is to continuously add new functions to fit the residuals of the previous prediction, to lower the residuals by continuous training or to stop after iterating to a predefined threshold, and to finally total the outcomes of each model to achieve model construction. Known training dataset $T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$, loss function $l(y_i,\hat{y}_i)$, regularization term $\Omega(f_k)$. Assuming that K trees have been trained, the final predicted value for the ith sample is:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \mathrm{k} \le n \tag{2}$$

Objective function:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

where $\hat{y}_i^{(k)}$ denotes the sum of the prediction results of the ith sample in the previous k models. $f_k(x_i)$ denotes the prediction result of the ith sample in the kth model. Because neither the prediction results of the first k-1 nor the regular terms of the first k-1 affect the objective function, the objective function at the kth tree can be optimized as:

$$Obj_k = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(k-1)} + f_k(x_i)) + \sum_{j=1}^{k-1} \Omega(f_k) \tag{4}$$

3.2.2. Expansion of the optimized approximate objective function using the second-order Taylor formula

XGBoost expands the optimized approximate objective function using the second-order Taylor formula, which can unify the loss function derivative form to support user-defined loss functions, while the second-order information itself can make the gradient converge faster and more accurately, with the first-order derivative guiding the gradient direction and the second-order derivative guiding how the gradient direction changes. According to the Taylor second-order expansion formula, Obj_k can be optimized as follows:

$$Obj_k = \sum_{i=1}^{n} \left[g_i f_k(x_i) + \frac{1}{2} h_i f_k^{\,2}(x_i) \right] + \Omega(f_k) \tag{5}$$

Among them $g_i = \partial_{\hat{y}^{(k-1)}} l(y_i, \hat{y}^{(k-1)})$, $h_i = \partial^2_{\hat{y}^{(k-1)}} l(y_i, \hat{y}^{(k-1)})$

3.2.3. Using regular terms

XGBoost adds a regular term to the objective function, which effectively controls the model's complexity. In terms of the bias-variance tradeoff, the regular term reduces the variance of the model, making the learned model simpler and preventing overfitting. XGBoost multiplies the weights of the leaf nodes by this factor after one iteration, mainly to weaken the influence of each tree and allow more learning space later. Define a tree of complexity Ω , which has two components:

$$\Omega(f_x) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2 \tag{6}$$

The parameterization of the leaf node weight vector and the number of leaf nodes. Bringing in the objective function, the final objective function of XGBoost can be obtained by merging the coefficients:

$$L^{(t)} = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T \tag{7}$$

3.2.4. Weight setting parameter class_weight

Because of the data provided by the online inspection system of the drainage network, after data cleaning, the four types of data that can be used are unevenly distributed and are typical of unbalanced sample data. There are 5642 pieces of sneaky drainage data, 682 pieces of rainfall and sewage mixing data, 1727 pieces of inflow and infiltration data, and 3563 pieces of normal data. When performing classification, when the sample sizes of different categories vary greatly, it may affect the classification results. To solve the problem of uneven samples of training data, the data are corrected. The SKlearn function is used for adjustment. The data volume and weights of each category are multiplied together to generate the same proportion by assigning various weights to distinct categories of data. The results of data volume and data weight adjustment are shown in Table 1.

Table 1. Data volume and weighting table

Number of samples	Weight	Result
5642	0.51462247	2903.49998
682	4.25733138	2903.5
1727	1.68123914	2903.45366
3563	0.81490317	2903.49999

3.2.5. Base learner

XGBoost supports three basic learners, gbtree, gblinear, and dart, which can be chosen via the booster parameter. Because this project involves constructing a multiclassification model, gbtree is picked as the basis learner. On the one hand, the decision tree is computationally simple and unaffected by intermediate missing values. On the other hand, it has a strong interpretation and can be more intuitive to comprehend the model operation, while the combination of decision trees can improve the tree's prediction effect. The tree structure of XGBoost for this project is shown in Figure 2.

4. Experimental results and discussion

In this paper, a combination of the mini-batch K-means algorithm and XGBoost is used to classify and model anomalies in the data, and the model-building process is shown in Figure 3. First, we preprocess the collected data to eliminate default and abnormal values. Then, for the first time, we use the mini-batch K-means algorithm to classify all of the data into abnormal and normal, and then classify the abnormal data into three categories: sneaky drainage, rainwater and sewage mixing, and inflow and infiltration, so that all of the data is classified into four categories. Simultaneously, in order to avoid misclassification, normal and abnormal data are reviewed using graphs and other techniques to assure classification accuracy. Finally, the classified data are labeled 0, 1, 2, and

3, which are used as samples for learning by the XGBoost algorithm and building classification models by XGBoost.

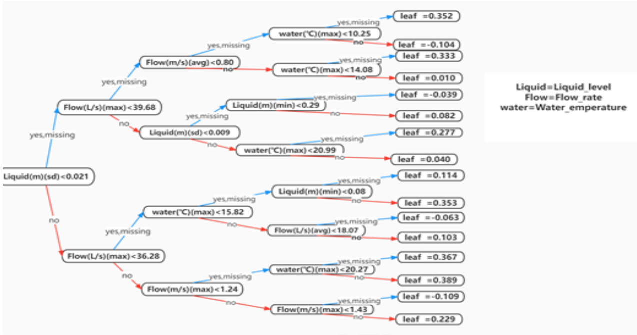


Figure 2. XGBoost tree structure

4.1. Data sources

The flow monitoring data provided by the drainage network's online monitoring system are typical high-latitude data with explicit variances, requiring considerable preprocessing before machine learning can be conducted (García *et al.*, 2016; Alexandropoulos *et al.*, 2019). The experimental data for this project were collected from several monitoring stations across Changsha City, Hunan Province, China. The drainage network's online monitoring system uses a Doppler flow meter to monitor the data, and the monitoring stations transmit data every five minutes, yielding 105,120 monitoring data in a year cycle. The data characteristics are flow rate (L/s) (maximum, minimum, mean, standard deviation); flow rate (m/s) (maximum, minimum, mean, standard deviation); level (m) (maximum, minimum, mean, standard deviation); water temperature (°C) (maximum, minimum, mean, standard deviation); Yellow Sea level; and rainfall (mm). To make the training data more generalized, a total of 16 monitoring point data points were selected and processed for a total of 11614 training data points. The data features are introduced as shown in Table 2.

Table 2: Data characteristics introduction table

Feature Name	Unit	Type
Flow rate (max)	L/s	float64
Flow rate (min)	L/s	float64
Flow rate (avg)	L/s	float64
Flow rate (sd)	L/s	float64
Flow rate (max)	m/s	float64
Flow rate (min)	m/s	float64
Flow rate (avg)	m/s	float64
Flow rate (sd)	m/s	float64
Liquid level(max)	m	float64
Liquid level(min)	m	float64
Liquid level(sd)	m	float64
Liquid level(avg)	m	float64
Water temperature(max)	℃	float64
Water temperature(min)	℃	float64
Water temperature(ad)	℃	float64
Water temperature(avg)	℃	float64
Yellow Sea level	m	float64
Weather	*10 mm	int64
Label		int64

Abbreviation: where max denotes the maximum value of data within five minutes, min denotes the minimum value of data within five minutes, avg denotes the mean value of data within five minutes, and sd denotes the standard deviation of data within five minutes.

4.1.1. Abnormal data processing

The monitoring device is located in the harsh environment of the sewer pipe and is vulnerable to a number of circumstances that might lead to mistakes in the obtained data, such as negative flow rates and periods of zero data gathering. To assure the correctness of the succeeding training and testing data sets, mistakes in the data are eliminated before analysis.

4.1.2. Feature filtering

The class of features is filtered by the variance of the feature itself (Jovic *et al.*, 2015). When a feature's variance is modest, it suggests that the characteristic changes little in essence and has no substantial influence on sample discrimination. Therefore, the varianceThreshold function is used to preferentially eliminate features with a variance of 0.

4.1.3. Feature selection

Feature selection is an effective step in machine learning data preprocessing that can effectively reduce data redundancy, remove feature data with minimal impact on data classification, improve model learning accuracy, and reduce model computation complexity while improving model understandability and interpretability (Anukrishna and Vince, 2017; Jie *et al.*, 2018). The features of this experiment are sorted by XGBoost's feature_imprtances_ function, and the top 12 features are selected for comprehensive training and testing. The data importance ranking and importance scores are shown in Figure 3. Furthermore, several qualities in XGBoost are used to evaluate the value of features to give an overall perspective of feature importance. Where weight is determined by the number of times the feature is utilized, as shown in Figure 4; gain is determined by the Gini index, as shown in Figure 5; and cover is determined by the average of the second-order derivatives of an indicator covering the sample, as shown in Figures 6 and 7.

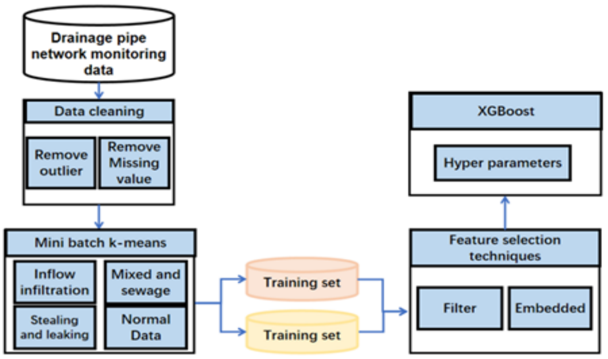


Figure 3. Model flow chart

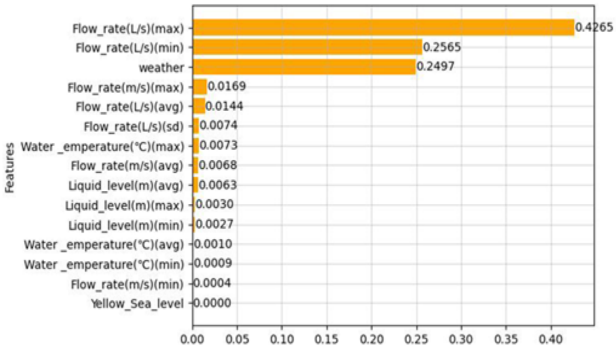


Figure 4. Feature importance ranking

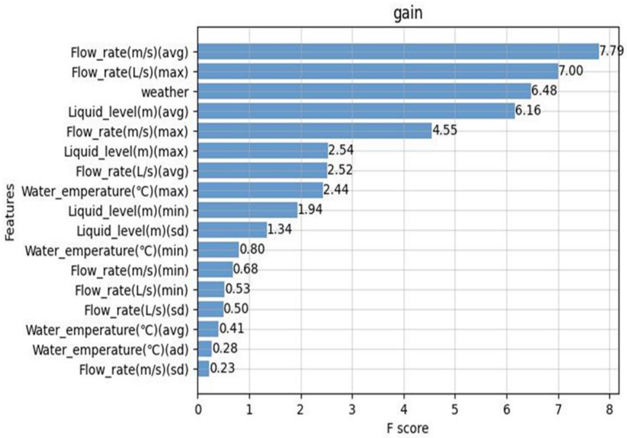


Figure 5. Gain feature importance ranking

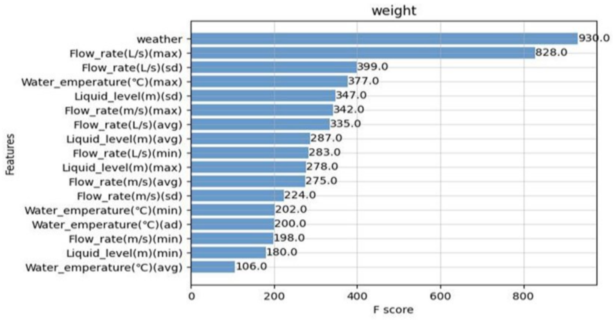


Figure 6. Weight feature importance ranking

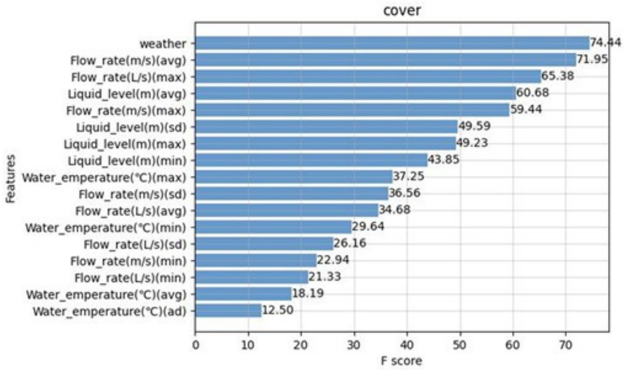


Figure 7. Cover feature importance ranking

4.2. Training

Train the XGBoost algorithm on all of the processed data. First, 20% of the data samples from normal drainage, sneaky drainage, rainwater and sewage mixing, and inflow and infiltration were utilized as the test set for the XGBoost classifier, while the remaining 80% were used as the training set. Improve model accuracy by adjusting the XGBoost parameters. A training subset and a validation subset are created from the training set, and the effect of parameter tuning is validated using a triple cross-validation approach. In the training subset, the XGBoost algorithm is used to train the classification model and to select the optimal Booster parameter combination. The experiments use CV grid search to determine the optimal parameters and verify the model effect by validating themselves. The parameters are set as shown in Table 3 below, and the experiment proves that the model has the highest accuracy when the parameters are set to the following table.

Table 3: XGBoost model parameters table

Parameter name	Parameter value
learning_rate	0.2
n_estimators	320
colsample_bytree	0.8
max_depth	4
random_state	3
subsample	0.7
reg_lambda	0.25
reg_alpha	0
objective	multi:softmax

The confusion matrix, as shown in Figure 8, depicts the many sorts of expected and true values. With the data training provided by the drainage network, XGBoost has a high accuracy rate.

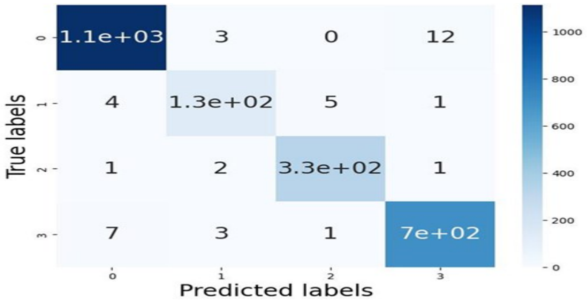


Figure 8. Model training confusion matrix

Additionally, to test the effectiveness of XGBoost on drainage network flow anomaly identification and classification, the SVM model, AdaBoost model, RF model, and XGBoost model were compared, and the comparison results are shown in Figure 9. It is demonstrated that XGBoost performs superiorly on the dataset provided by the drainage network monitoring system when the same training and test sets are selected and both default parameters are used.

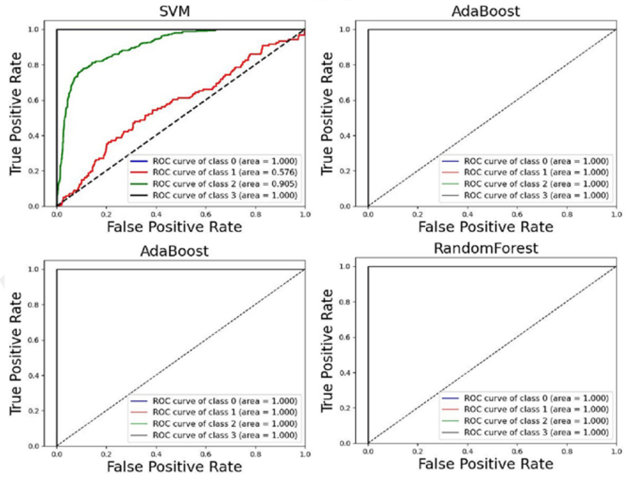


Figure 9. Algorithm Comparison Chart

4.3. *Model validation*

To test the efficacy and accuracy of XGBoost's drainage network flow anomaly classification in drainage network monitoring data species, data from other monitoring stations that were not involved in the model training were used initially. Four monitoring points with perfect data were selected, and the validation results are shown in Table 4 below. To see the model validation results more intuitively, the confusion matrix is shown in the form of Figure 10 below.

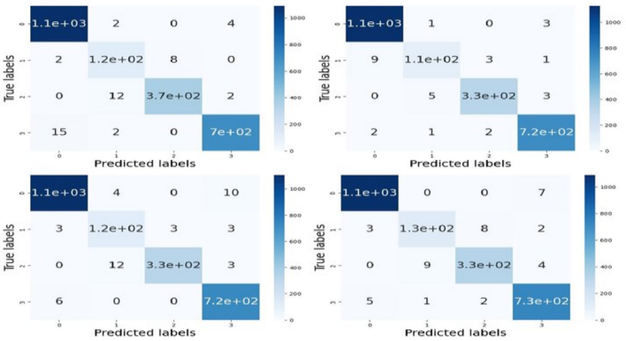


Figure 10. Model Prediction Confusion Matrix

Table 4: Model validation accuracy table

Monitoring number	Accuracy
1	83.571%
2	81.107%
3	81.765%
4	89.455%

5. Conclusion

An XGBoost-based online drainage network flow data anomaly identification and classification model is proposed. The model uses the mini-batch K-means algorithm as an unsupervised classification algorithm for monitoring data anomaly calibration and then uses XGBoost to model the calibrated data. The modeling was done with data provided by an inspection site in Changsha, China, and the model was tested with data provided by other sites that did not participate in the training. At the early stage of model training, the same modeled data were compared on the SVM model, AdaBoost model, and RF model, and the results showed that XGBoost can effectively identify and classify drainage pipe flow anomalies and that the algorithm has good performance. This project combines the mini-batch K-means algorithm, XGBoost algorithm, and Doppler flow meter monitoring data for the first time and classifies the flow data into four specific categories: normal, sneaky drainage, rainwater and sewage mixing, and inflow and infiltration. The analysis using big data can not only serve as the basis for the classification of abnormal data in the online drainage network but also cooperate with the online system to alarm abnormal conditions in real time, helping the monitoring personnel solve the problems of sneaky drainage, rain and sewage mixing, inflow and infiltration. It is crucial for urban development and sustainable urban development that drainage network monitoring systems are built scientifically, using contemporary methods of effective management, to assist relevant government officials in fully exploiting the role that drainage network information plays in the construction, operation, and management of urbanization.

Due to the limitations of this study, some future work is necessary. First off, as this model only works with numerical characteristics, all data must first be transformed into numerical features in order to use it. Second, additional research is required for other areas with significant variations in water consumption and precipitation statistics. Future research can gather data with significant regional variations for comparative analysis to improve the model's generalizability.

Acknowledgements

This research was supported by the Anhui Province Key Research and Development Program (202004107020009, 202004h07020023).

References

Aitken J.M., Evans M.H., Worley R., Edwards S.S., Zhang R., Dodd T., Mihaylova L.S. and Anderson S.R. (2021). Simultaneous Localization and Mapping for Inspection Robots in Water and Sewer Pipe Networks: A Review. IEEE Access, 9, 140173–40198.

Alan J & Karla.B. and Bogunovic N. (2015). A review of feature selection methods with applications. 1200–1205.

Alexandropoulos S., Kotsiantis S. and Vrahatis M. (2019). Data preprocessing in predictive data mining. *The Knowledge Engineering Review*, **34**, E1.

Alonso J.B. (2013). K-means vs Mini Batch K-means: a comparison.

Amir G & Fang C. and Laith A. (2022). Machine Learning Technologies for Big Data Analytics. Electronics. 11. 421.

Anukrishna P.R. and Paul V. (2017). A review on feature selection for high dimensional data. 2017 International Conference on Inventive Systems and Control (ICISC), 1–4.

Bareš V., Stránský D. and Sýkora P. (2012). Evaluation of sewer infiltration/inflow using COD mass flux method: case study in Prague. Water science and technology : a journal of the International Association on Water Pollution Research, 66(3), 673–680.

Barlow H. (1989). Unsupervised Learning. Neural Computation, 1(3), 295–311.

Beheshti M. and Sægrov S. (2018). Quantification Assessment of Extraneous Water Infiltration and Inflow by Analysis of the Thermal Behavior of the Sewer Network. Water.

Bock H. (2007). Clustering Methods: A History of k-Means Algorithms.

Cai J., Luo J., Wang S. and Yang S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70–79.

Candice B & Anna C. and Gonzalo M.M. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*. **54**.

Chavan M, Patil A, Dalvi L, *et al.* (2015). Mini batch K-Means clustering on large dataset[J]. *International Journal of Science, Engineering and Technology.Research*, **4**: 1356–1358.

García S., Ramírez-Gallego S., Luengo J., Benítez J. M. and Herrera F. (2016). Big data preprocessing: methods and prospects. Big Data Analytics, 1(1).

Hastie T., Tibshirani R. and Friedman J. (2009). Unsupervised Learning. In: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY.

Heureux L.A. Grolinger K. Elyamany H.F. and Capretz M.A.M. (2017). Machine Learning With Big Data: Challenges and Approaches in IEEE Access, **5**, 7776–7797.

Hongrong Q., Jianzhong L., Guohui Z., and Luexuan L. (2011). Study of problems and corrective actions of urban drainage network. 2011 International Conference on Electric Technology and Civil Engineering (ICETCE), 1561–1564.

Jain K.A. and Dubes C.R. (1988). Algorithms for Clustering Data， Englewood Cliffs， NJ， USA : Prentice-Hall.

Jato-Espino D., Toro-Huertas E. I. and Güereca L.P. (2022). Lifecycle sustainability assessment for the comparison of traditional and sustainable drainage systems. *The Science of the total environment*, **817**, 152959.

Kwon S.H. and Kim J.H. (2021). Machine learning and urban drainage systems: State-of-the-art review. Water (Switzerland), 13(24), [3545].

Li Q.S. and Wang T.C. (2013). Construction and Sustainable Development of Pipe Networks of Underground Drainage System. *Applied Mechanics and Materials*, 438–439, 1076 – 1079.

Moselhi O. and Shehab-Eldeen T. (1999). Automated detection of surface defects in water and sewer pipes. Automation in Construction, **8**, 581–588.

Moselhi O. and Shehab-Eldeen T. (2000). Classification of Defects in Sewer Pipes Using Neural Networks. *Journal of Infrastructure Systems*, **6,** 97–104.

Munawar H.S., Hammad A.W. and Waller S.T. (2021). A review on flood management technologies related to image processing and machine learning. Automation in Construction.

Nadia L & Falk Anne Katrine & Morten B & Henrik M. and Peter M. (2018). Model predictive control of urban drainage systems: A review and perspective towards smart real-time water management. *Critical Reviews in Environmental Science and Technology*. **48**. 1–61.

Nti I.K., Aning J., Bbk A., Frimpong K., Appiah A.Y. and Nyarko-Boateng O. (2021). A Comparative Empirical Analysis of 21 Machine Learning Algorithms for Real–World Applications in Diverse Domains.

Qiu Junfei amd Wu, Qihui & Guoru D and Yuhua X & Shuo F. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*. 2016.

Sinaga K.P. and Yang M. (2020). Unsupervised K-Means Clustering Algorithm. IEEE Access, 8, 80716–80727.

Sodhi P., Awasthi N. and Sharma V. (2019). Introduction to Machine Learning and Its Basic Application in Python. *SSRN Electronic Journal*.

Tariq S. and Osama M. (2005). Automated Detection and Classification of Infiltration in Sewer Pipes. *Journal of Infrastructure Systems* – J INFRASTRUCT SYST. **11**. 10.1061/(ASCE)1076–0342(2005)11:3(165).

Tatiparthi S.R., De Costa Y.G., Whittaker C.N., Hu S., Yuan Z., Zhong R.Y. and Zhuang W. (2021). Development of radio-frequency identification (RFID) sensors suitable for smart-monitoring applications in sewer systems. *Water research*, **198**, 117107 .

Tianqi C. and Carlos G. (2016). XGBoost: A Scalable Tree Boosting System. 785–794.

Wang H. and Song L. (2020). Water Level Prediction of Rainwater Pipe Network Using an SVM-Based Machine Learning Method. *International Journal of Pattern Recognition and Artificial Intelligence*., **34**, 2051002:1–2051002:18.

Wang J., Liu G., Wang J., Xu X., Shao Y., Zhang Q., Liu Y., Qi L. and Wang H. (2021). Current status, existent problems, and coping strategy of urban drainage pipeline network in China. *Environmental Science and Pollution Research*, **28**, 43035 – 43049.

Yazdanfar Z. and Sharma A.K. (2015). Urban drainage system planning and design--challenges with climate change and urbanization: a review. Water science and technology : *a journal of the International Association on Water Pollution Research*, **72** 2, 165–79 .

Yojna A. (2020). A review of machine learning techniques over big data case studies. *International Journal of Innovative Research in Computer Science & Technology*. **8**. 10.21276/ijircst.2020.8.3.34.

Zhou Q., Ren Y., Xu M., Han N. and Wang H. (2016). Adaptation to urbanization impacts on drainage in the city of Hohhot, China. Water science and technology : *a journal of the International Association on Water Pollution Research*, 73 1, 167–75 .

Zhu Q., Gu A., Li D. *et al.* (2021). Online recognition of drainage type based on UV-vis spectra and derivative neural network algorithm. Front. *Environmental Science and Engineering*. **15**, 136.