

Predicting Technical Problems of Hydropower Engineering Using eXtreme Gradient Boosting

Jing Zhu¹, Yi Chen^{1,*}, Liming Huang², Chunyong She², Yangfeng Wu², Wenyu Zhang¹

¹School of Information, Zhejiang University of Finance and Economics, Hangzhou, China

²Quality & Safety Inspection Center of Hydropower Engineering of Zhejiang Province, Hangzhou, China

Email address:

zhujing765@zufe.edu.cn (Jing Zhu), tifyesung8@zufe.edu.cn (Yi Chen), 981993965@qq.com (Liming Huang), 9261130@qq.com (Chunyong She), 348691814@qq.com (Yangfeng Wu), wyzhang@e.ntu.edu.sg (Wenyu Zhang)

*Corresponding author

To cite this article:

Jing Zhu, Yi Chen, Liming Huang, Chunyong She, Yangfeng Wu, Wenyu Zhang. Predicting Technical Problems of Hydropower Engineering Using eXtreme Gradient Boosting. *Science Journal of Applied Mathematics and Statistics*. Vol. 6, No. 4, 2018, pp. 124-129.

doi: 10.11648/j.sjams.20180604.13

Received: September 14, 2018; **Accepted:** October 16, 2018; **Published:** October 18, 2018

Abstract: Nowadays, water shortage is increasingly severe, which has huge negative influence on daily life. Constructing hydropower engineering is one of the approaches to alleviate such problem. Therefore, it's worth settling technical problems of hydropower engineering timely, which will help people not only make better use of water resources but also get rid of various security risks. To achieve such goal, this study predicts potential technical problems that hydropower engineering might happen. In order to utilize the large amount of data, data mining techniques are used to solve this multi-classification problem. First of all, plenty of data is preprocessed. Particularly, because of the complexity of text data, text mining techniques are applied to transform the unstructured data to structural data. Then, eXtreme Gradient Boosting (XGBoost) is applied to make the classification. To validate efficiency of the model, comparisons are made among XGBoost, Gradient Boosting Decision Tree, Random Forest, Decision Tree, k-Nearest Neighbor and Bernoulli Naïve Bayes from the perspective of accuracy, precision, recall and f-score. The experimental result shows that XGBoost is more suitable to solve this classification problem. This study provides engineering inspectors with helpful suggestions of particular technical problems that need attention, and further enables people to inspect engineering more efficiently and effectively.

Keywords: Data Mining, Hydropower Engineering, Multi-classification Problem, eXtreme Gradient Boosting

1. Introduction

Water resources are indispensable but limited resources needed to be taken advantage of in a rational way. Hydropower engineering is one of the efficient approaches which allow people to make good use of water resources. There are various studies focused on the reasonable construction of hydropower engineering [1-5]. With the development of science and technology, the present era has been an era of information explosion, which provides people with substantial and valuable knowledge. Without exception, relevant departments of hydropower engineering also have accumulated a great amount of data in the process of engineering supervision. Nevertheless, with traditional methods, it costs a lot of time to acquire useful information,

and even it's likely to be drowning in plenty of data without acquisition of any meaningful conclusion. Thus, data mining techniques emerged as helpful solutions to obtaining underlying knowledge from messy data.

There are various methods available to mine data, including text mining, classification, association rules, clustering, outlier analysis and so on. Some researchers have utilized data mining techniques to explore the better ways to construct and operate hydropower engineering [6-9]. This study pays attention to predicting possible technical problems of hydropower engineering with the help of data mining, which has almost not been studied before. eXtreme Gradient Boosting (XGBoost) and other 5 classification models are adopted.

The remainder of this paper is organized as follows. The presentation of related work is in Section 2. Section 3

introduces data preprocessing and modeling method. In Section 4, experimental result is analyzed and Section 5 concludes this study.

2. Related Work

Recently, data mining techniques have been developing rapidly and there are lots of researchers applying data mining in different fields. In professional sports, Valero [10] used data mining to predict Win-Loss outcomes in MLB regular season games. In education, data mining has been utilized as well. Yukselturk *et al.* [11] used data mining techniques to analyze the dropout students' features and Shingari *et al.* [12] did a study to predict students' performance in higher education. What's more, in business, Sun *et al.* [13] predicted the financial distress of companies and Xu *et al.* [14] did an unemployment rate prediction.

Furthermore, data mining techniques also become powerful tools in the field of hydropower engineering. Cobaner *et al.* [15] build an ANN model to predict the possibility of hydropower plant installation to an existing irrigation dam and Su *et al.* [16] used it to predict the service life of hydropower engineering. Jiang *et al.* [17] used the technology of data mining to calculate the failure probability of hydraulic structures for rural hydropower. Shi [18] used PCA-ANN model to predict the elevation of water conservancy projects.

Based on the previous study, it can be concluded that many researchers have attached great importance to data mining in various fields, however, few paid attentions to the specific technical problems of hydropower engineering. As a consequence, this study focuses on this particular respect, intending to assist engineering inspectors to inspect and solve problems in time.

3. Data Preprocessing and Modeling Method

The dataset used in this study comes from a supervision and management center of hydropower engineering in China. This section will detail how to preprocess the dataset and build a classification model.

3.1. Data Preprocessing

The raw dataset is composed of several documents with partial same features so that data merging is conducted at first. The merged data has 4472 records, containing 16 features and 1 label- "technical problem of hydropower engineering". Since it's not convenient to build classification models directly with such data, data preprocessing is done as following steps.

(1) Text mining

The feature- "the description of fact" is text data, which is unstructured data. It's of great difficulty to make use of such data directly when predicting. There possibly exists misspelling, terminologies, abbreviations and other linguistic structures in text, what's more, context is important when

acquiring information from text. Nevertheless, text data provides engineering inspectors with plenty of information, which means that it's not advisable to simply ignore the feature. Therefore, in order to take full advantage of raw data, the text undergoes processing and is transformed to feature vectors, based on which classification models are applied.

Each record of "the description of fact" is seen as a document in text mining, and the collection of the whole documents is called the corpus. All the words in the corpus rank in order of the occurrence. Thus, every word owns its corresponding sequence number.

In the processing, the grammar, word order, sentence structure and punctuation are ignored [19]. Every single word in a given document helps predict which class the document belongs to more or less. However, stop words are eliminated, such as "about", "before" and "either", owing to the fact that such words are too common in every document to mean anything. Even if stop words are deleted, there are still too many words, which will hinder computers from operating well if all the remaining words are applied to make prediction. Then, for each document, term frequency-inverse document frequency (TFIDF) are utilized to calculate the weight of every word [19]. The TFIDF value of a word t in a given document d is defined as equation (1), additionally, equation (2) and equation (3) define TF and IDF respectively [19]. TF(t, d) measures the frequency of a word t in the given document d . IDF(t) measures the weight of a document d in the whole corpus for a given word.

$$\text{TFIDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t) \quad (1)$$

$$\text{TF}(t, d) = \frac{\text{number of feature } t \text{ in the document } d}{\text{number of all the features in the document } d} \quad (2)$$

$$\text{IDF}(t) = 1 + \log\left(\frac{\text{total number of documents}}{\text{number of documents containing } t}\right) \quad (3)$$

For every word, its weight is its TFIDF value. The higher TFIDF value is, the more significant the word is when classifying. To be more specific, the more documents containing word t are, the more common word t is, furthermore, the word can't differentiate the class of documents well because many classes share a same word.

For convenience of classification, 50 new features (from feature 1 to feature 50) are created for the whole corpus. For a given document, values of these features are the sequence numbers of important words sorting by weight. More specifically, sequence number of the word with the highest weight is the value of feature 1, while sequence number of the word with the lowest weight is the value of feature 50. If a document doesn't have 50 words in all, the missing value of remaining features will be filled by -1.

(2) Deletion of unimportant features

In the dataset, some features don't affect technical problems so that they are removed in order to acquire more satisfactory experimental result, such as "the name of the hydropower engineering". Additionally, feature- "the description of fact" has been preprocessed in the first step and 50 new features have been created, therefore, the very feature can be

abandoned now.

(3) Addition of necessary features

In the first step, 50 features are created after text mining, which will not be repeated in detail. The focus of this step is the features which don't influence the technical problems independently. Nevertheless, combing such features will present some interesting results. For instance, "time interval" is such a combined feature subtracting "commence time of hydropower engineering" from "recording time of the technical problem".

(4) Transformation of literal data

It's of great difficulty for classification models to process literal data directly so that all the literal data is transformed to numeric data. For example, for feature- "type of the hydropower engineering", 1 represents "seawall" while 2 represents "river channel".

For the sake of clearness and conciseness, Table 1 presents part of selected features, while Table 2 illustrates part of candidate problem category of this multi-classification problem.

Table 1. Part of selected features.

Feature	Description	Range
TypResUnits01	types of responsible units	regulatory unit-1, detection unit-2, design unit-3, construction unit-4, project legal person-5
TypEngin	type of the hydropower engineering	seawall-1, river channel-2, agriculture-3, reservoir-4, reclamation-5, small hydropower-6, water diversion-7, others-8
TypSupVision	type of the supervision agency	provincial level-1, prefecture level-2, county level-3
TotalInv	total investment of the hydropower engineering	from 1 to 1382390, unit: ten thousand yuan
HydroGrad	grade of the hydropower engineering	6 grades altogether
EnginNature	nature of the hydropower engineering	reinforcement-1, reconstructure-2, extension-3, new-4
EnginStatus	status of the hydropower engineering	initial period-1, peak period-2, later period-3, completion-4, acceptance-5, suspension-6
TimeInter	time interval between recording time of the technical problem and commence time of the hydropower engineering	from -24 to 43083, unit: day

Table 2. Part of candidate problem category.

Problem	Description
1	earth rock excavation doesn't meet requirements
2	drainage of foundation pit is not timely
3	bolt-shotcrete support doesn't comply with standards
4	the construction of pine piles is not in line with regulations
5	the construction of bored piles is not in accordance with standards
46	the setting of safety warning signs doesn't meet the standards
47	safety measures for blasting are not satisfactory
48	safety measures for welding and cutting are not in place
49	protective equipment wearing is not up to requirements
50	safety measures for working over water are not conform to regulars

3.2. Modeling Method

To verify the performance of different classification models, the preprocessed dataset is divided into two sub-datasets, in other words, training data and testing data. Training data is used for building models, while testing data is for detecting the performance of classifiers. In the process of division, there exists a special situation needed careful consideration. Regardless of accountability units, different records of a same hydropower project may share the same technical problem. If the dataset is randomly divided without any special handling, there might be strong similarity between the records in training data and ones in testing data, which influences performance of classifiers. Thus, a definite ID is set up for every record in the dataset, records of the same hydropower project sharing a same ID. Then the preprocessed dataset is divided randomly on the premise of allocating records with same ID to the same sub dataset, 70% of data being training

data and 30% being testing data.

This study employs eXtreme Gradient Boosting (XGBoost) to predict technical problems, which makes some improvements on the basis of Gradient Boosting Decision Tree (GBDT), assembling many regression trees also [20]. Other than Random Forest (RF), trees in the XGBoost correlates with each other. The objective function of XGBoost is defined as

$$\text{Obj} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (4)$$

where i represents the i th record to be predicted, n is the number of all the records to be predicted, y_i represents the actual value of the i th record, \hat{y}_i represents the predicted value, and $l(y_i, \hat{y}_i)$ is the loss function of the i th record. The loss function represents the forecast error which is expected to be as small as possible. Moreover, the loss function undergoes Taylor expansion, which is one of advantages over GBDT. What's more, K is the number of the whole regression trees, k

represents the k th tree, and $\Omega(f_k)$ describes complexity of the k th tree. The smaller $\Omega(f_k)$ is, the less complex the model is, and further the more excellent generalization ability the model owns, which can prevent the model from overfitting effectively.

4. Experiment

This section reveals the process of experiments. Values of important parameters are firstly decided, and then performance of XGBoost with other 5 classifiers is compared. The experiments are conducted by Python Version 3.6 on a PC with a 2.60 GHz Intel CORE i5 processor. The PC has 4 GB of RAM, running the Microsoft Windows 7 operating system.

4.1. Parameters Selection

Eta (i.e., `learning_rate`) and `max_depth` are significant parameters which require special adjustments. If values of these two parameters are not suitable, the classification model is likely to fall into overfitting problem, which has serious impacts on the universality of the classification model. When deciding the value of a certain parameter, other parameters are kept as the defaults or the values that have been adjusted to the best. The specific method of defining eta is as follows: testing the accuracy rates of XGBoost with eta changing from 0 to 1 and increasing by 0.1, and the method to define `max_depth` is similar. Figure 1 and Figure 2 give an account of the results of parameters selection. From the two figures, eta is set as 0.1 and `max_depth` is set as 2.

4.2. Performance Comparison

For the purpose of evaluating the efficiency of XGBoost, other 5 classifiers are also utilized to predict in order to make comparison, which are Gradient Boosting Decision Tree (GBDT), Random Forest (RF), Decision Tree (DT), k-Nearest

Neighbor (KNN) and Bernoulli Naïve Bayes (Bernoulli NB), respectively. Evaluation indexes are accuracy, precision, recall and f-score, which are introduced subsequently as next 4 equations. In addition, Table 3 explains variables of equations (5)-(8).

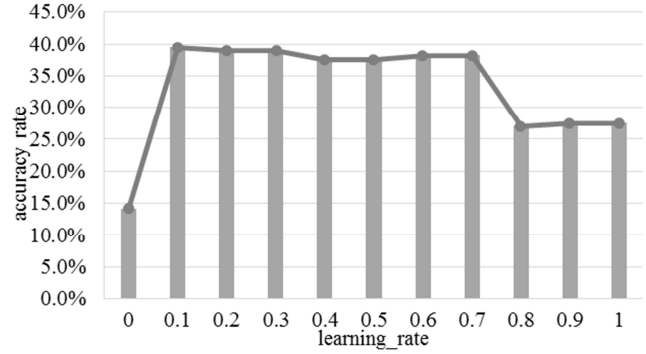


Figure 1. XGBoost accuracy with change of eta.

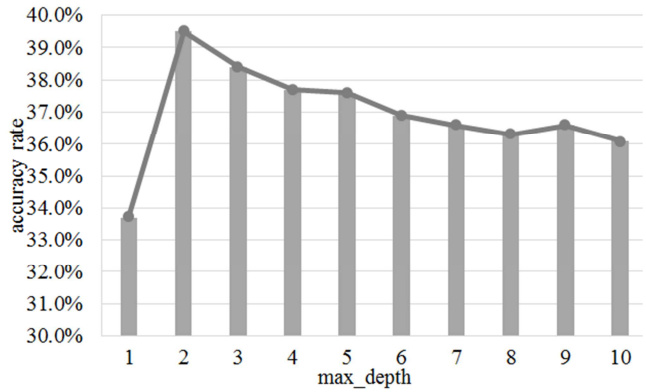


Figure 2. XGBoost accuracy with change of max_depth.

Table 3. Confusion Matrix.

		True	
		positive	negative
Predicted	positive	True Positive (TP)	False Positive (FP)
	negative	False Negative (FN)	True Negative (TN)

$$\text{accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (5)$$

$$\text{precision} = \frac{TP}{TP+FP} \quad (6)$$

$$\text{recall} = \frac{TP}{TP+FN} \quad (7)$$

$$\text{f-score} = \frac{2}{1/\text{precision} + 1/\text{recall}} \quad (8)$$

Figure 3 illustrates different accuracy rates of 6 classification models, while Figure 4 shows precision, recall, and f-score of these classifiers. From these two figures, it can be easily concluded that XGBoost is more appropriate in this multi-class problem.

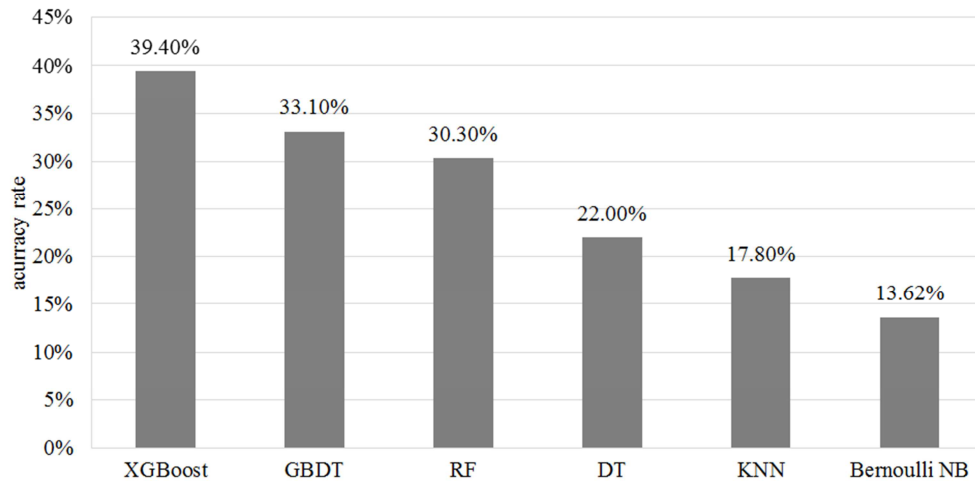


Figure 3. Comparison of accuracy rates of different classifiers.

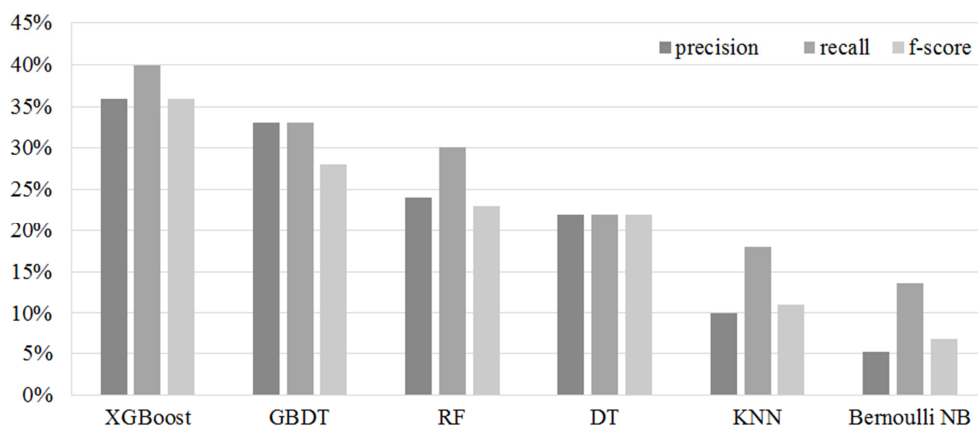


Figure 4. The precision, recall and f1-score in different model.

5. Conclusion

Hydropower engineering is significant for utilizing limited water resources, which consumes large amount of human resources, material resources and financial resources in the construction process. If technical problems of hydropower engineering can't be resolved effectively and efficiently, not only extra investment will be wasted, but also safety accidents will happen. Therefore, it's of great significance to predict technical problems, which assists the concerned people to inspect the engineering more pertinently. In this study, firstly, necessary preparations are conducted so as to build more accurate classification models. Particularly, in order to make better use of data, text mining techniques are applied, which improves the performance of classifiers considerably. Then, XGBoost is used to make prediction in comparison with other 5 classifiers. The experimental result shows that XGBoost outperforms other classification models whatever the evaluation index is. Based on the prediction results, people concerned could attach more attentions to potential problems and inspect engineering accordingly. This study supplies engineering inspectors with support for management of hydropower engineering.

At the same time, there still exists space for improvement. The

data used in this study is limited, and it's possible that the very classification models utilized are not suitable for other data. Hence, it's advisable to apply more data to validate the experimental result. What's more, to further enhance the accuracy rates of models, heuristic algorithms can be put into use.

References

- [1] Kowalczykjuško, A., Mazur, A., Grzywna, A., et al. (2017). Evaluation of the possibilities of using water-damming devices on the Tyśmienica River to build small hydropower plants. *Journal of Water and Land Development*, 35(1), 113-119.
- [2] Qin, P., & Cheng, C. (2017). Prediction of seawall settlement based on a combined LS-ARIMA model. *Mathematical Problems in Engineering*, 2017, Article ID: 7840569.
- [3] Sojka, M., Jaskula, J., Wicher-Dysarz, J., et al. (2016). Assessment of dam construction impact on hydrological regime changes in lowland river - a case of study: the Stare Miasto reservoir located on the Powa River. *Journal of Water and Land Development*, 30(1), 119-125.
- [4] Sadaoui, M., Ludwig, W., Bourrin, F., et al. (2018). The impact of reservoir construction on riverine sediment and carbon fluxes to the Mediterranean Sea. *Progress in Oceanography*, 163, 94-111.

- [5] Yaeger, M. A., Massey, J. H., Reba, M. L., et al. (2018). Trends in the construction of on-farm irrigation reservoirs in response to aquifer decline in eastern Arkansas: implications for conjunctive water resource management. *Agricultural Water Management*, 208, 373-383.
- [6] Ghimire, B. S., & Jangareddy, M. (2013). Optimal reservoir operation for hydropower production using particle swarm optimization and sustainability analysis of hydropower. *ISH Journal of Hydraulic Engineering*, 19(3), 196-210.
- [7] Naumann, S., Schwanenberg, D., Karimanzira, D., et al. (2015). Short-term management of hydropower reservoirs under meteorological uncertainty by means of multi-stage optimization. *AT - Automatisierungstechnik*, 63(7), 535-542.
- [8] Su, H., Li, X., Yang, B., et al. (2018). Wavelet support vector machine-based prediction model of dam deformation. *Mechanical Systems and Signal Processing*, 110, 412-427.
- [9] Zhong, D., Du, R., Cui, B., et al. (2018). Real-time spreading thickness monitoring of high-core rockfill dam based on k - Nearest Neighbor algorithm. *Transactions of Tianjin University*, 24(3), 282-289.
- [10] Valero, C. S. (2016). Predicting win-loss outcomes in MLB regular season games-a comparative study using data mining methods. *International Journal of Computer Science in Sport*, 15(2), 91-112.
- [11] Yukselturk, E., Ozekes, S., & Turel, Y. K. (2014). Predicting dropout student: an application of data mining methods in an online education program. *European Journal of Open, Distance and E-Learning*, 17(1), 118-133.
- [12] Shingari, I., Kumar, D., & Khetan, M. (2017). A review of applications of data mining techniques for prediction of students' performance in higher education. *Journal of Statistics and Management Systems*, 20(4), 713-722.
- [13] Sun, J., & Li, H. (2008). Data mining method for listed companies' financial distress prediction. *Knowledge-Based Systems*, 21(1), 1-5.
- [14] Xu, W., Li, Z., Cheng, C., et al. (2013). Data mining for unemployment rate prediction using search engine query data. *Service Oriented Computing and Applications*, 7(1), 33-42.
- [15] Cobaner, M., Haktanir, T., & Kisi, O. (2008). Prediction of hydropower energy using ANN for the feasibility of hydropower plant installation to an existing irrigation dam. *Water Resources Management*, 22(6), 757-774.
- [16] Su, H., Hu, J., Yang, M., et al. (2015). Assessment and prediction for service life of water resources and hydropower engineering. *Natural Hazards*, 75(3), 3005-3019.
- [17] Jiang, C., Sheng, J., Zhang, G., et al. (2012). Calculation of failure probability of hydraulic structures for rural hydropower. *Procedia Engineering*, 28, 161-164.
- [18] Shi, L. L. (2014). Prediction model for mark-up of water conservancy projects based on PCA-ANN. *Journal of Economics of Water Resources*, 32(3), 52-55.
- [19] Fawcett, T., & Provost, F. (2015). *Data Science for Business*. USA: O'Reilly Media, Inc.
- [20] Chen, T. Q., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 13-August 17, San Francisco, USA, pp. 785-794.