Detection of water pipeline leakage based on random forest

Zhaozhao Chi1,2, Yunfei Li1,4, Weihao Wang1, Caishun Xu1,3 and Rui Yuan1

1Sciences research department of Hefei Institute of public safety, Tsinghua University, Hefei, Anhui, 230601, China
2Hefei Kdlian Safety Technology Company, Limited, Hefei, Anhui, 230601, China
3School of Electric and Information Engineering, Anhui Jianzhu University, Hefei, Anhui, 230601, China
4Corresponding author’s e-mail: uc_fly@163.com

Abstract. Pipeline leakage is a great concern for the transportation industries and researchers have been devoted in leakage detection for a long time. Machine learning is developed for leakage recognition recently and it can help to achieve the leakage detection. However, the effect is limited by feature complexity and noise. As a machine learning method, Random Forest (RF) is good at handling with high-dimensional data and predicts well even when the signal is interrupted by noise. As a result, RF was applied to better deal with the leakage detection. Researches herein have compared the RF classifier and other well-developed machine learning methods in respects of the classification accuracy and calculation time. The result indicated that the RF classifier outperformed Support Vector Machine (SVM), Artificial Neural Network (ANN), k-Nearest Neighbors (k-NN) and Decision Tree (DT) classifiers, with the classification accuracy of 88.33%.

Nomenclatures

| Abbreviation | Description                      |
|--------------|----------------------------------|
| DT           | Decision Tree                    |
| ANN          | Artificial Neural Networks       |
| k-NN         | k-Nearest Neighbors              |
| SVM          | Support Vector Machines          |
| RF           | Random Forest                    |
| Mtry         | the number of input variables tried at each split of RF |
| NoT          | Number of decision trees in RF   |

1. Introduction

The oil and gas pipelines have increased considerably recent years, and many accidents occurred in pipeline transportation. Pipeline leakages pose a critical problem in pipeline transportation, especially long-distance transportation, and research on theories and methods of leakage detection and location are vital in pipeline transport. [1]

With the development of big data analytics, leak detection research mainly relies on machine learning these years. [2] The Support Vector Machine (SVM) [3], fuzzy SVM [4] and least squares
SVM [5] have been used for leakage recognition. However, the classification accuracy influenced by the quality of data largely. This paper aims at selecting a better machine learning method to deal with this problem.

The well-known instances of machine learning include SVM, Decision Tree (DT), [6] Artificial Neural Networks (ANN), [7, 8] k-Nearest Neighbors (k-NN), [8] Partial Least Squares [9] and Linear Discriminant Analysis [10]. Nonlinear SVM, k-NN and ANN perform well and are good at modeling the relationship between leakage signal and leak states. Nevertheless, ANN and k-NN need feature dimension reduction when dealing with high-dimensional data. [8] Nonlinear SVM can cope with the data with high dimension but is not steady when facing with many irrelevant features, thus feature selection is also needed. As linear methods, Partial Least Squares and Linear Discriminant Analysis is not suitable for complicated signals in leakage detection. [10]

DT has many merits compared with other methods. It can handle with high-dimensional data; good at ignoring irrelevant features; handles complicated signal of leakage. DT building algorithms has been developed integrally. [6] However, DT has a drawback that it usually attains a bad performance. [6]

Because of the advantages of DT, many efforts have been devoted in improving its prediction accuracy. Several tree-based algorithms were developed as a result. [11] One way to enhance the performance of DT method is to use ensembles of trees. [12] Therefore, the ensemble method, Random Forest (RF), is presented in this article. [13] The result revealed that RF doing well in the classification situation.

2. Principle of Random Forest
The RF was first proposed by Breiman. [13] The principle of RF is illustrated in this part.

2.1. Ensemble of trees
Ensemble of B trees \{T_1(X), ..., T_B(X)\}, in which X= \{x_1, ..., x_p\} is the properties related to a leak state and is a p-dimensional vector, form the RF method. Accordingly, the ensemble of B trees generates B outputs \{\hat{Y}_1 = T_1(X), ..., \hat{Y}_B = T_B(X)\} in which \hat{Y}_b, b=1, ..., B. \hat{Y}_b is known as the prediction of the leak state by the b-th tree. The final prediction \hat{Y}, is determined by ensemble of all trees’ outputs.

2.2. Training procedure
Suppose the training data has n numbers and the form is \(D = \{(X_i, Y_i), \ldots, (X_n, Y_n)\}\), in which \(X_i\) is a vector of features and \(Y_i\) is the corresponding leak state (e.g., leak/leak-free), the training algorithm is revealed as follows.

1. Select a bootstrap sample (i.e., choose the sample randomly in \(D\), with replacement) from the training data.
2. For each bootstrap sample, a tree is generated with the following procedure: select the best segmentation in the randomly selected subset of \(M_{xy}\) (rather than all) features at each node. \(M_{xy}\) is substantially the only tuning parameter in the algorithm. The tree is grown to its maximum size (i.e., until no further splitting is possible) and is not trimmed back.
3. Repeat (1) and (2) steps until (a sufficiently large number) B such trees are grown.

The RF algorithm is the same as Bagging When \(M_{xy}=p\), that is, when the optimal segmentation of each node is selected among all features. [14] The tree growth algorithm used in RF is C4.5 [15] although other options can also be considered.

2.3. Testing procedure
1. The testing procedure starts from the root node of the developed tree, then decides which node to go, if the value < \(th\), choose the left node, otherwise the vice. \(th\) is the threshold of the current node. The step continues until it arrives at the end of the tree, a certain leaf node, then the predicted result of a tree is obtained.
(2) Repeat step (1) for B trees, then B numbers of predicted result is outputted. The final prediction \( \hat{Y} \) is the ensemble of predicted results in all trees.

The RF belongs to bagging integrated algorithm, and Bootstrap is adopted. It can be found that about 1/3 samples of Bootstrap will not be used in the growth of the b-th tree. The out-of-bag data is applied for validation of the b-th tree.

3. Experimental equipment

The pipeline leakage can result in acoustic signal, and the signal will propagate along the pipe. The leakage detection system in Sciences research department of Hefei Institute of public safety is developed to acquire the acoustic signal, consisting of sensor, processing card and PC, as is shown in Figure 1.

In Figure 1 (c), the water flows out of the tank and along the pipe into the tank. The leakage orifice size is 1 mm, the pressure of the pipe is 0.2 MPa. The experiment was carried out on sunny days to eliminate the white noise of the rainy day.

In Figure 1 (d), the sensor was located 5 m away from the leakage point, attached to the iron pipe. The sensor acquires the acoustic signal at the sample rate of 4096 Hz. 130 times of experiments were
carried out when leakage happened, with the duration of 1 s each time. As for the signal without leak, the experiment time and frequency were the same. The signal attained by sensor is the voltage signal in fact, the processing card transfer the signal into digital signal, then sending to PC for leakage recognition.

4. Data processing procedures
The signal attained by acoustic sensor was composed of leak signal and noise. The median filter was used to process the noise in this experiment. Median filter substitutes the value of a point with the median value of a digital sequence, thus guarantee the surrounding values as similar as the true value, then removing pulsed noise. The unsteady water flow would cause the pulsed noise and it influences the signal most. As a result, the median filter was applied herein. The size of the filter is set to 8 empirically.

In Figure 2, the signal before and after median filter was revealed, of which the leak happened in the middle of the sampling.

![Figure 2. The signal before and after median filter.](image)

| Features in time domain | Expressions | Features in frequency domain | Expressions |
|-------------------------|-------------|-----------------------------|-------------|
| Mean                    | \( \mu_s = \frac{1}{N} \sum_{i=1}^{N} \text{abs}(x_i) \) | Skewness \( \frac{1}{N} \sum_{i=1}^{N} \left( \frac{f_i - \mu_f}{\sigma_f} \right)^3 \) |
| Standard deviation (STD) | \( \left[ \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_s)^2 \right]^{1/2} \) | Kurtosis \( \frac{1}{N} \sum_{i=1}^{N} \left( \frac{f_i - \mu_f}{\sigma_f} \right)^4 \) |
| Energy                  | \( \frac{1}{N} \sum_{i=1}^{N} (x_i)^2 \) | Spectral centroid \( \frac{\sum_{i=1}^{N} f_i \cdot p_i}{\sum_{i=1}^{N} p_i} \) |
| Peak                    | \( \text{max}[\text{abs}(x_i)] \) | Correlation factor \( \frac{\sum_{i=1}^{N} \cos(2\pi f_i)p_i}{\sum_{i=1}^{N} p_i} \) |
RF method, as an ensemble of trees, generates the relationship between signal features and leakage states in a “black box”. Therefore, the features should be calculated at first. Several features were calculated from the filtered signal for the preparation of RF classifier, which is shown in Table 1.

5. Result and discussion
The feature of the signal can be different under leak and leak-free state. Machine learning method can be used to recognize this difference. RF is applied for machine learning in this article. As is known, there are 130 times of experiments in each state, including leak and leak-free. Select 100 groups in each state and combine them, the data was defined as the training set. The other 30 groups of the data were selected as testing set. [18] The RF was accomplished by Matlab.

For RF classifiers, the number of input variables tried at each split $M_{try}$, affect the performance a lot, usually set to the square root of the number of input features. [19]

Additionally, number of decision trees (NoT) is of vital importance in the optimization of the RF. [20] Usually, NoT is determined with the default value of 10 and 100. With the change of NoT, the test data accuracy was revealed in Figure 3.

![Figure 3. The trend of model accuracy with the change of NoT.](image)

Figure 3 demonstrated that test data accuracy is the highest when NoT is 10. The recognition accuracy of the leak state reached 88.33%.

In order to compare RF with other machine learning methods, ANN, k-NN, DT and SVM are also applied for leak state classification. The result was shown in Table 2. ANN, k-NN, DT and SVM are implemented by Prtool. The parameter of ANN goes as follows: number of neurons in the hidden layer is 78, standard deviation of weights in an input layer is 1. As for k-NN, number of the nearest neighbors is optional. When using SVM model, radial basis function was chose as kernel function. The one-against-one method, which constructs $k(k-1)/2$ SVM models, is applied for multiple classification recognition. The penalty factor $C$ and the kernel parameter $G$ were optimized by cross-validation, settled as $9.766 \times 10^{-4}$. In DT method, there is no pruning, and the computation of a decision tree classifier out of a dataset $A$ using the Fisher criterion.

### Table 2. Comparison among several machine learning methods.

| Methods   | ANN  | k-NN | SVM  | DT   | RF    |
|-----------|------|------|------|------|-------|
| Accuracy  | 0.8667 | 0.8503 | 0.8503 | 0.8333 | 0.8833 |
| Training time/s | 10.4015 | 0.0286 | 0.0641 | 0.0010 | 0.0580 |
| Testing time/s | 0.0192 | 0.0612 | 0.0184 | 0.0172 | 0.0298 |
As is shown in Table 2, the accuracy of RF is 1.66% higher than ANN, 3.3% higher than k-NN and SVM, 5% higher than DT. Considering calculating time, ANN cost the most in training, DT cost the least. It is inferred that ANN have more complicated construction, and DT is much simpler than RF. There’s little difference between testing time. In conclusion, RF have the best accuracy and it won’t take much time, which is the most suitable for the leakage detection in this experiment.

6. Conclusion
Machine learning has been applied in leak detection in recent years, and it turned out that this method could solve the problem in some degree. The RF classifier is an ensemble of decision trees, integrating the result of B trees, thus is more convincing in the classification result. RF also has excellent performance in classification tasks. Compared with other machine learning methods, RF was used in this article in order to attain a more suitable leak detection model. The result showed that RF exceeded ANN, k-NN, DT, SVM in leakage recognition. The accuracy of the RF classifier reaches 88.33% in this experiment. More experiments will be carried out to test practicality of this method in the future.

Acknowledgments
This work was supported by Anhui Provincial Central guidance of local science and technology development special funding projects (201907d07050007).

References
[1] Liu C W, Meng L Y, Wang W C and Zhang F 2014 Study on leak-acoustics generation mechanism for natural gas pipelines J. Loss Prev Process Ind 32 174-181
[2] Diao X, Jiang J C, Shen G D, et al. 2020 An improved variational mode decomposition method based on particle swarm optimization for leak detection of liquid pipelines Mech Syst Signal Pr 143 106787
[3] Sun J D, Xiao Q Y, Wen J T and Zhang Y 2016 Natural gas pipeline leak aperture identification and location based on local mean decomposition analysis Measurement 79 147-57
[4] Xu Q Q, Zhang L B and Liang W 2013 Acoustic detection technology for gas pipeline leakage Process Saf Environ Prot 91 253-261
[5] Jin H, Zhang L B, Liang W and Ding Q K 2014 Integrated leakage detection and localization model for gas pipelines based on the acoustic wave method J Loss Prev Process Ind 27 74-88
[6] Rusinko A III, Farmen M W, Lambert C G, Brown P L and Young S S 1999 Analysis of a large structure/biological activity data set using recursive partitioning J. Chem Inf Comput Sci 39 1017-1026
[7] Doniger S, Hofmann T and Yeh J 2002 Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms J. Comput Biol 9 849-864
[8] Kauffman G W and Jurs P C, 2001 QSAR and k-nearest neighbor classification analysis of selective cylooxygenase-2 inhibitors using topologically based numerical descriptors J. Chem Inf Comput Sci 41 1553-1560
[9] Sheridan R P, Nachbar R B and Bush B L, 1994 Extending the trend vector: the trend matrix and sample-based partial least squares J. Comput-Aided Mol Des 8 323-340
[10] Bakken G A and Jurs P C, 2000 Classification of multidrug-resistance reversal agents using structure-based descriptors and linear discriminant analysis J. Med Chem 43 4534-4541
[11] Tong W, Hong H, Fang H, Xie Q and Perkins R, 2003 Decision forest: combining the predictions of multiple independent decision tree models J. Chem Inf Comput Sci 43 525-531
[12] Dietterich T G, 2002 Ensemble learning In The Handbook of Brain Theory and Neural Networks 2nd ed, Arbib M A Ed; The MIT Press: Cambridge
[13] Breiman L, 2001 Random forests. Machine Learning 45 5-32
[14] Breiman L, 1996 Bagging predictors. Machine Learning 24 123-140
[15] Breiman L, Friedman J H, Olshen R A and Stone C J, 1984 Classification and Regression Trees \textit{Chapman & Hall/CRC: Boca Raton}

[16] Li S Z, Song Y J and Zhou G Q, 2018 Leak detection of water distribution pipeline subject to failure of socket joint based on acoustic emission and pattern recognition \textit{Measurement 115} 39-44

[17] Rui Z, Ru Q Y, Zheng H C, et al, 2019 Deep learning and its applications to machine health monitoring \textit{Mech Syst Signal Pr 115} 213-237

[18] Aymon L, 2019 Leak Detection using Random Forest and Pressure Simulation \textit{2019 6th Swiss Conf. on Data Science} (Bern:Switzerland) 109-110

[19] Liaw A, Wiener M, Classification and regression by random forest 2002 \textit{Rnews 2} 18-22

[20] Jaehoon C, Hyunseung K, Addis L G and Dongil S, 2018 Deep neural network and random forest classifier for source tracking of chemical leaks using fence monitoring data \textit{J. Loss Prevent Proc 56} 548-558