ATLANTIS PRESS

Advances in Engineering Research, volume 140

2nd International Conference on Electrical, Control and Automation Engineering (ECAE 2017)

Leakage Detection in Pipelines Using Decision Tree and Multi-Support Vector Machine

Zhigang Chen[1, 2, *], Xu Xu[1], Xiaolei Du[1], Junling Zhang[1] and Miao Yu[1,2]

[1]School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China

[2]Beijing Engineering Research Center of Monitoring for Construction Safety, Beijing, China

[*]Corresponding author

***Abstract*—In order to solve the problem of leakage detection in the case of complex conditions and limited training samples, a multivariate classification recognition model was built by using Decision Tree and Support Vector Machine, which has advantages of rapid speed and high efficiency in classification and outstanding characteristics in small samples binary classification. The model was trained with a fault feature vector which is a dimensionless value extracted from the pipeline pressure signal characteristic parameters, and then using the model to test the samples. The results show that this method not only can complete the model learning training in the case of small samples, but also has been greatly improved over the neural network method in terms of the recognition performance, and can be effectively applied to leakage detection in pipelines.**

Keywords-leakage detection; decision tree; support vector machine; binary classification

I. INTRODUCTION

Pipeline is one of the important infrastructures in every country. Negative pressure wave will be generated when the pipeline leaks. Due to the special nature of the negative pressure wave, it is often be used to detect the leak in pipelines, but the key problem is to accurately distinguish the cause of the pressure signal fluctuation. However, in many cases, the pressure fluctuation and leakage fluctuation caused by adjusting pumps, regulating valves and other process operations are very similar and it is difficult to accurately identify them. In recent years, artificial intelligence detection methods based on neural networks and pattern recognition have been developed rapidly[1][2]. However, the effectiveness of these algorithms is mostly based on enough training samples. In practice, the number of samples used for testing is small, and in particular the number of fault samples is limited or even missing.

Support Vector Machine (SVM) has the advantages of small training samples, strong generalization ability and easy to get the global optimal solution. It has been widely used in many fields such as electricity, economy, medicine, diagnosis and so on[3]. In recent years, support vector machines have also been used for pipeline leak monitoring. Literrature [4] presents a novel technique support vector machine (SVM) for pipeline leak detection and an SVM classifier was used to classify the signal pattern with few samples. SVM has clearly better advantages than neural network method over small sample set. Later the SVM parameters were optimized by using the Artificial Bee Colony (ABC) with searching power of local and complete area optimal solution, the experiment showed that the ABC-SVM algorithm has a better accuracy and adaptability in leakage recognition than conventional SVM method [5]. In this paper, the SVM classifier was further improved and a Decision-Tree-SVM was present for less calculation and better recognition efficiency.

II. IMPROVED SVM MODEL WITH DECISION-TREE

A. *Theory of Decision-Tree*

Decision tree is a predictive model, which represents a mapping between object attributes and object values [6]. As shown in Figure 1, the decision tree contains three parts, root node, branch node and leaf node, which represent different attributes respectively. Bifurcation path represents a possible attribute value. Decision tree classification process consists of two steps. First, establish a reasonable decision tree model, the second step is the use of the decision tree model in the previous step on the new data by level classification.

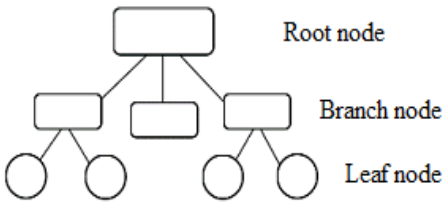

FIGURE I. STRUCTURE OF DECISION TREE

Taking Figure 1 as an example, each node in the tree can complete a classification subtask. In the classification stage, the logic structure of the decision tree is generated by a bottom-up cohesion algorithm or a top-down segmentation algorithm [7]. Because only part of the classifier was used in each level, the classification efficiency and accuracy was relatively high.

B. *Theory of SVM*

The basic idea of SVM is to transform the input space into a high-dimensional space by non-linear transformation, then find the optimal linear classification surface in this new space. The above solution is achieved by defining an appropriate inner product kernel function. SVM is to propose the optimal

classification surface under the condition of linearly separable series. H is the classification line. H1 and H2 are respectively the straight lines parallel to the classification lines and the nearest samples to each classification line, and the distance between them is called classification interval. The optimal classification line is not only the two types of samples can be correctly separated, but also make the classification interval maximum classification line [8].

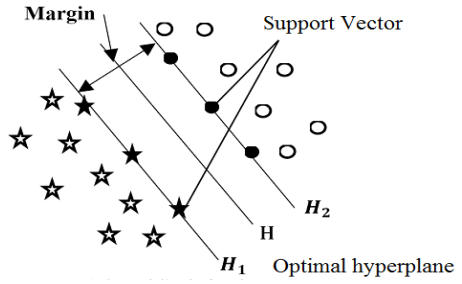

FIGURE II. THE OPTIMAL HYPERPLANE OF DIAGRAM

Consider a planar classification task, Figure 2 for instance. There are two kinds of training samples: ○ and ☆. The set of vectors should be optimally separated by a hyperplane without error. The maximal distance between the hyperplane separating the two classes and the closet data points to the hyperplane is defined as margin. Then the error bound of machine learning is minimized by maximizing the margin to have a better generalization performance. Hence, a separating hyperplane in canonical form must satisfy following constraints:

$$y_i(w \cdot x_i + b) \geq 1 \; i = 1,2,...,n \tag{1}$$

Where, x_i is a set of training samples, $y \in \{-1,1\}$ is corresponding label. w is the normal vector of the hyperplane. In most conditions, such a hyperplane does not exist. So we need to relax the constraints of Equation (1) by introducing slack variable $\varepsilon_i \geq 0$, $i = 1,2,...,n$

$$y_i(w \cdot x_i + b) + \varepsilon_i \geq 1 \; i = 1,2,...,n \tag{2}$$

To maximize the margin, the task is therefore:

Minimize $\phi(w) = \frac{1}{2}\|w\|^2$ subject to Function (2), and the learning task can be reduced to minimization of the primal Lagrangian:

$$L_p(w,b,\alpha) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{n} \alpha_i [y_i(w \cdot x_i + b) - 1] \tag{3}$$

Where α_i are Lagrangian multipliers, hence $\alpha_i \geq 0$. Now we need minimize $\sum_i \alpha_i \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j$, subject to

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \quad , \quad 0 \leq \alpha_i \leq C \; i = 1,2,...,n \tag{4}$$

C is the punishment coefficient to control the trade-off between training error and generalization ability. The decision function now is:

$$f(x) = \text{sgn}(\sum_i y_i \alpha_i (x_i \cdot x_j) + b) \tag{5}$$

Choose different forms of kernel function, you can get different support vector. There are four commonly used kernel functions, the most commonly one is Gauss radial basis function [9].

C. Multi-Support Vector Machine

(1) ONE-VERSUS-REST

The one-versus-rest support vector machine algorithm is the earliest used method when dealing with multi-valued classification problem [9]. The principle of the method is to distinguish each category from all the other categories in turn using a two-class support vector machine classifier. For an n-type problem, the number of support vector machines needed to be training is n in the one-versus-rest method needs, that is, n classification hyperplanes is used to classify.

The classification process of one-to-many method is clear, but there are some shortcomings as follows. First, when using the one-versus-rest approach, the number of positive samples per classifier is generally much smaller than the number of negative samples, which will greatly reduce the accuracy of classification. Secondly, Each SVM training needs all the training samples to be trained, and the computational efficiency is low.

(2) ONE-VERSUS-ONE

For an n-type problem, a support vector machine is needed to be constructed for each category of the n-type samples in the one-versus-one method, so the total number of support vector machines is n (n-1) / 2. Although the number of support vector machines that need to be trained is more than one-versus-rest method, however, only two types of samples were required to be trained for training each support vector machine. The training speed of this method is faster than that of the one-versus-rest method.

Literature [10] compares several Multi-SVM classification algorithms mentioned above and finds that the one-versus-one method has a better classification effect, but its computational cost is large. The one-versus-rest method is less in computation, but its classification effect in general.

D. MSVM Model with Decision-Tree

Support vector machines have good generalization performance in the case of small training samples, but multiple classifiers are needed to be constructed multiple type problems, and the time of training and diagnosis is long. In this paper, a multi-class fault identification model with combination of decision tree and support vector machine was established for pipeline leakage detection.

Support Vector Machine method based Decision Tree is to decompose the multi-classification problem into a series of binary classification problems and these binary classification problems are distributed on each node of the decision tree. When modeling, according to different attributes the decision tree root nodes and branch nodes are divided into several sub-sets step by step, until all the leaf nodes are obtained. In attribute sub-set according to the actual selection of one-versus-one or one-versus-rest support vector machine classification model. When attribute sub-set by the time, according to the actual situation choose one-versus-one or one-versus-rest support vector machine classification model. Taking division of 6 categories as an example, Figure 3 is a schematic diagram of the classification of one of the decision trees, and the 6 types of input samples are hierarchically divided into their respective categories.

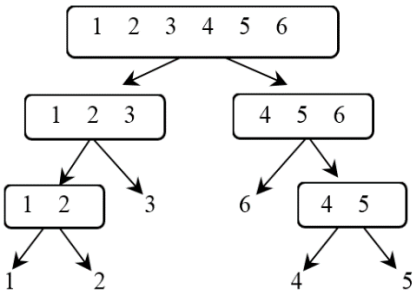

FIGURE III. THE SCHEMATIC DIAGRAM OF DT-SVM

As can be seen from Figure 3, the SVM based on decision tree comprehensively considers the advantages of less one-versus-rest classification model vector machines, higher accuracy of one-versus-one multi-class classification and high classification efficiency of decision trees.

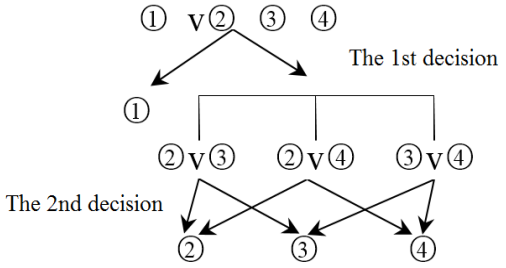

FIGURE IV. THE CLASSIFICATION DIAGRAM OF DT-SVM

The main object of this paper is leakage identification of pressure pipes. Based on the above principles, a multi-classification SVM fault diagnosis model based on decision tree is constructed. Take the normal working condition, small hole leakage (the amount of leakage is not less than 5%), stopping pump and regulating valve as four kinds of fault, the classification model shown in Figure 4. Since the equipment is in normal operation under most conditions, it is relatively easy to get the samples of the normal operation of the pipeline during the actual test. At the same time, it is relatively easy to distinguish the normal operation of the pipeline from other faults. The main purpose of the first layer is to exclude non-fault samples from all samples, so the one-versus-rest classifier is used to quickly identify non-fault samples. one-versus-one algorithm is adopted in the second-level decision-making to identify three types of faults one by one. In this case, only three one-to-one classifiers are built. This model only needs to construct 4 support vector machines, which is reduced by 2 SVMs compared with the one-to-one method (4 * (4-1) / 2 = 6) in MSVM Model with Decision-Tree method. In theory, the training and testing time will be reduced and the diagnostic efficiency will increase.

III. EXPERIMENTAL RESEARCH

A. Experiment System and Feature Extraction

The study and tests presented here are based on data of a pilot water pressurized pipe which is located in mechatronics lab. The length of the test pipeline is totally 220 m. In the experimental system four conditions were simulated which are normal fluctuations, small leaks, pump operation and valve operation conditions respectively. For each type of pipeline, a typical fault or condition was tested several times, and multiple sets of eigenvectors that can reflect the fault rule are taken as samples of such fault training.

The main tested parameter is pressure wave of inner water. In order to obtain more abundant parameter signals, parameters were acquired at high speed with OPC mode and four synchronical channels [8], the sampling frequency was 50Hz, the sampling total sections were 40 and every section had 1024 integer data.

Various characteristic parameters of pipeline pressure signal have different emphases on the ability of expressing fault information. In this paper, the peak coefficient, kurtosis index, skewness index, effective value and standard deviation are selected in reference [14] to describe the waveform characteristics of the signal. Because of the dimensional difference between these time-frequency domain parameters, normalization is performed prior to modeling to transform it into data for [0, 1].

B. Model Training

Four types of samples of pipeline operation parameters are selected, each type consists of five groups, a total of 20 groups of signal samples, according to Figure 4 to establish a pipeline fault identification model for learning and training. After analysis this paper selected radial basis function, the training steps are as follows:

(1) Transforming signal data into a recognizable format required by the Libsvm package;

(2) The training samples are scaled and mapped to [-1, 1]

(3) Model parameter training (penalty factor C and kernel function parameters);

(4) Using the C and the model parameters obtained in the step (3), the scaled training sample in the step (2) is trained by using the model;

(5) Put in the test sample into the trained model to test the classification results;

The SVM classifier is trained by using the sample data in turn to obtain the optimal classification function, and finally the radial basis parameter of the penalty function C = 2 and the kernel function is obtained.

C. Testing

In order to test the above classifier recognition effect, eight sets of known samples to be tested are used to verify the classifier to verify the generalization ability and accuracy of the classifier. Table I shows the output of different SVM decision functions to be diagnosed. In the first column of data, the normal fluctuation 1 quickly identified from the leak or operations 2, 3, 4, the result is positive then judged as positive samples, that is, normal fluctuation, the recognition ended; The negative is classified as the other three group type, leak or operation, you need to carry out 1-to-1 classification, as shown in the 2nd, 3rd, 4th column.

TABLE I. RESULT OF TEST SAMPLES

NO.	2V234	2V3	2V4	3V4
1	0.8502	/	/	/
2	0.9122	/	/	/
3	-1.0000	1.0000	1.0000	1.0000
4	-1.0000	-0.9343	0.0758	1.0000
5	-1.0000	0.2174	-0.4027	-1.0000
6	-0.1405	0.5323	1.0000	0.2031
7	-1.0000	-0.2641	1.0000	1.0000
8	-1.0000	1.0000	-0.3510	-1.0000

According to the output membership of each independent SVM in the decision structure, the ownership of the sample to be diagnosed is judged. When one of the SVM *i*, *j* for the faulty sample *x* is judged as the *i*th type of fault, the number of votes in class *i* increases by 1, otherwise, the number of votes in class *j* increases by one. According to this diagnostic decision rule, we have carried out some experiments and analysis with multiple samples. The final diagnosis results coincide with the type of sample fault set, which shows the correctness of the method. If there is multiple fault types or more fault samples, we can promote it by this method.

D. Compared with the Conventional Method

In order to further compare the classification results based on decision tree SVM with traditional neural networks and conventional multivariate support vector machines, 20 groups of samples were used to experiment on different models. The classification results are shown in Table II below.

TABLE II. COMPARISON OF RECOGNITION RESULTS

Name	Samples	Number of correct	Time/s	Accuracy
DT--SVM	20	19	1.69	95%
SVM	20	19	2.55	95%
BP（δ=0.01）	20	16	11	80%
BP（δ=0.05）	20	15	15	75%

As can be seen from Table II, the corresponding decision tree construction based on actual fault support vector machine recognition effect and conventional multivariate support vector machine the same method, are obviously due to the traditional neural network method, but the use of decision tree support vector machine classification recognition time than conventional Support vector machine method is shortened by about 35%.

IV. CONCLUSION

In this paper, leak detection identification model based on decision tree and support vector machine is designed combining the advantages of decision tree decision-making efficiency and one-versus-one and one-versus-rest multi-valued classification in SVM. The model has better recognition ability and classification effect in the case of small sample and leakage, which is obviously better than the traditional neural network method. The recognition effect is equivalent to one-to-one multivariate support vector machine, but it is shorter than the traditional support vector machine in learning training and testing. With the increase of the number of categories, compared with the conventional SVM, MSVM with tree decision is more obvious and more efficient.

ACKNOWLEDGMENTS

This research was supported by the Projects of the National Natural Science Foundation of China (No.51004005) and the Beijing Excellent Talent Training (No. 2013D005017000013) and the Ministry of Housing and Urban-Rural Development of the People's Republic of China (No. 2016-K4-081) and Beijing Municipal Commission of Education Science and Technology Plan (SQKM 201710016014).

REFERENCES

[1] Zhi-gang Chen, Xinrong Zhong, YidongXie, “Leakage Diagnosis Method for Pipelines Based on Multi-weight Neural Network”, Applied Mechanics and Materials, vol. 697, pp. 429-433, 2005.

[2] Fang Wang, Weiguo Lin, Zheng Liu,Shuochen Wu, Xiaobo Qiu, “Pipeline Leak Detection by Using Time-Domain Statistical Features”, IEEE SENSORS JOURNAL, vol. 17(19), pp. 6431-6441, 2017

[3] Songrong Luo, Junsheng Cheng, and HungLinh Ao, “Application of LCD-SVD Technique and CRO-SVM Method to Fault Diagnosis for Roller Bearing”, Shock and Vibration, pp. 1–8, 2015.

[4] Xue-yuan Wang, Zhi-gang Chen, Xin-rong Zhong, “Research on Leak Detection of Water Pipeline Base on PSO-SVM”, 2016 International Conference on Applied Mechanics, Electronics and Mechatronics Engineering (AMEME2016), pp. 227-233, June 2016.

[5] Santosh Kumar Mandala, Felix T.S.Chan, M.K.Tiwari, "Leak detection of pipeline: An integrated approach of rough set theory and artificial bee colony trained SVM", Expert Systems with Applications, vol. 39(3), pp. 3071-3080, 2012.

[6] Sugumaran V, Ramachandran KI, "Automatic rule learning using decision tree for fuzzy classifier in fault diagnosis of roller bearing", Mech Syst Signal Process, vol. 21(5), pp. 2237–2247, 2007.

[7] Krishnakumari, A. Elayaperumal, M. Saravanan, C. Arvindan, "Fault diagnostics of spur gear using decision tree and fuzzy classifier", The International Journal of Advanced Manufacturing Technology, vol. 89(9–12), pp. 3487–3494, 2017.

[8] Vapnik V N., "The nature of statistical learning theory", New York: Spring-Verlag, 1999.

[9] YongtaoZheng, YuhuLiu, "An Analysis of Multi-class Support Vector Machines", Computer Engineering and Applications, vol. 41 (23), pp.190-192, 2005.

[10] YanWang, Huan-huanChen,Yi Shen, "Multi-class support vector machine based on directed acyclic graph", ELECTRIC MACHINES AND CONTROL,vol. 15(4), pp. 85-89, 2011.