2014

Decision Tree Classification Model In Water Supply Network

Sun Jilong

Wang Ronghe

Ping Junhui

Cai Liang

Xiao Chaohong

DECISION TREE CLASSIFICATION MODEL IN WATER SUPPLY NETWORK

JILONG SUN (1), RONGHE WANG (1), JUNHUI PING (1), CHAOHONG XIAO (1), SI LI (1), LIANG CAI (1)
(1):Graduate School at Shenzhen, Tsinghua University, the University Town, Shenzhen, 518055, China

ABSTRACT

With the service life of water supply network (WSN) growth, the growing phenomenon of aging pipe network have become exceedingly serious. Urban water supply network is a hidden underground asset, therefore, it is difficult for monitoring staff to make a direct classification towards the faults of pipe network by means of the modern detecting technology. In this paper, based on the basic property data (e.g. diameter, material, pressure, distance to pump, distance to tank, load, etc.) of water supply network (WSN), decision tree algorithm (C4.5) can be carried out to classify the specific situation of water supply pipeline. Part of the historical data is to establish a decision tree classification model, then the remaining historical data is to validate this established model. Adopting statistical method is to assess the decision tree model including Basic Statistical Method (BSM), Receiver Operating Characteristic (ROC)). These methods can be successfully used to assess the accuracy of this established classification model of water pipe network. The purpose of classification model is to classify the specific condition of water pipe network. It is important to maintain the pipeline according to the classification results including asset unserviceable(AU)、near perfect condition(NPC) and serious deterioration(SD). Finally, this research focuses on pipe classification which plays a significant role in maintaining water supply networks in the future.

INTRODUCTION

The municipal infrastructure construction in urban construction is an essential part, however the maintenance costs with pipeline in water supply networks(WSN) are rapidly rising so that the total cost which accounts for 56% of the total investment in water supply network [1].Most importantly, it is exceedingly difficult for the construction and maintenance of water supply network with less information about the operational state of water supply network and the hysteresis of getting information[2].

In an attempt to alleviate the failures in water supply network, we adopt decision method to predict the specific situations according to service lift of pipeline, in some extent, this method can weaken the failure reasons in water supply pipeline [3]. This advantage in classification can be well widely used. One key task in the decision tree approach is to identify pipes at risk of failure and performing inspection, maintenance and repair prior to pipe failures [4]. Failure mode of supply water network pipes can be contributed to two reasons including structural and hydraulic failures, which are associated with two different deterioration processes. The structural deterioration is a more complex process since it is subjected to multiple attacks such as chemical corrosion and mechanical stress. Structural failure is often observable as a collapse event which causes traffic disruption and in the worse situation, human loss. The hydraulic deterioration is a simpler process since it involves the build-up of sediments [5].

Based on spatial historical data in water supply network and method of decision tree, this

paper attempts to carry out a classification analysis towards historical failure data and makes a classification for network area in terms of historical data. Meanwhile, the paper intends to carry out statistical analysis by ROC curves in order to assess the precision of the experimental method in this test. Finally, establishing this method for the purpose of laying a good foundation for classifying the supply water network in future.

BASIC THEORY OF EXPERIMENT

The algorithm theory

Quinlan [6] proposed ID3 algorithm which had lay foundation for the algorithm of decision tree. All the methods of decision tree like C4.5、CART、CHAID and QUEST etc., which have derived from ID3 algorithm, so it was landmark in the process of the development of decision tree.

Based on the evolution of ID3 algorithm, Quinlan. J.R [7] had proposed C4.5 algorithm. Compared to ID3 algorithm, most importantly it has lower computational complexity and higher computational efficiency. It is important for the improvement of ID3 algorithm through ratio of information gain to select node attributes. The experiment shows that ratio of information gain is better than method of information gain.

Description of C.4.5 algorithm

C4.5 algorithm is able to handle continuous attribute value. The standard of selecting node is based on the maximum the ratio of information gain, the specific algorithm steps are as follows: Building a tree begins with a root node that represents the entire, given dataset and recursively split the data into smaller subsets by testing for a given property at each node. The sub-tree represents the partitions of the original dataset that satisfy specified attribute value test. This splitting process for given dataset typically continues until the subsets are single category, all instances in the subset are pure, in which the tree splitting is stopped [8].

Input: an attribute-valued dataset D

1: Tree = {}
2: ***if*** D is “pure” OR other stopping criteria met then
3: Terminate
4: ***end if***
5: ***for*** all attribute a∈D do
6: Compute information-theoretic criteria if we split on a
7: ***end for***
8: abest = Best attribute according to above computed criteria
9: Tree = Create a decision node that tests abest in the root
10: Dv = Induced sub-datasets from D based on abest
11: ***for*** all Dv do
12: Treev = C4.5 (Dv)
13: Attach Treev to the corresponding branch of Tree
14: ***end for***
15: ***return*** Tree

EVALUATION STANDARD EXPERIMETAL PERFORMANCE

In this part, after discussing the experimental algorithm, then we will provide several methods with precision evaluation of C4.5 algorithm. It consist of two methods: Receiver Operating Characteristic Curves (ROC) and Statistical Method(SM).Establishing method is essential for assessing the test method.

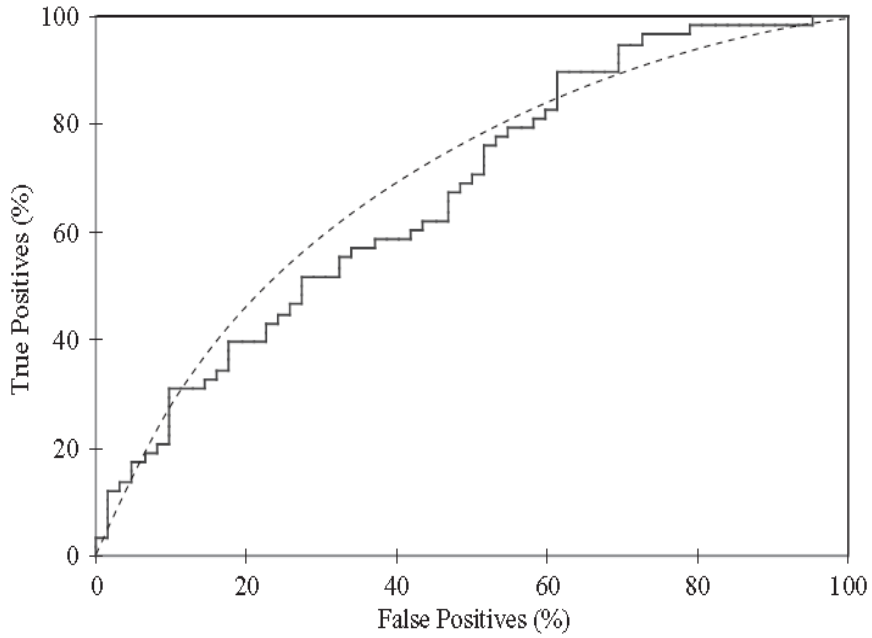

Figure 1. A sample ROC

Figure 1 is ROC curve [8]—the jagged line, which is widely used to balance the tradeoff between true-rate and false-rate over a noisy channel. ROC curves defines the performance of classification experiment. In Figure 1, it shows that vertical axis represents the true positive rate, however, the horizontal axis represents the true negative rate. The true rate is the number of positives included in this sample (TP Rate =100 × TP/ (TP + FN)).The true negative is the number of negatives included in this sample (FP Rate =100 × FP/ (FP + TN)). Generally，adopting ROC area is to assess effectively, when the ROC area is greater than 0.5, the result is effective. If ROC area is equal to 1, the establishing method is quietly perfect, however this situation cannot meet the reality.

TEST CASE

Description of basic information

The training data is from WUXI, the fault data of water supply network can be divided into the following situations: valve damage, faucet damage, pipeline losses, water tank damage and bursting pipes etc., however, this paper does not distinguish these damages; Instead, all circumstances of damage are considered as the fault point. The two-dimensional coordinates (X, Y) of recording the fault point is used to record the data of damaged point. In total, there are 20 fault points of pipe network in experiment, the recorded GIS data in fault points is extracted as a two-dimensional coordinate data, that's means, this coordinates in the chart are not the actual coordinates. All parameters explain: Material(Ductile iron，pipe (DIP)，Cast iron pipe (CIP)，Concrete pipe(CP))、Distance to pump(D-P，m)、Distance to Tank(D-T, m). Class((Near perfect condition(NPC)，Serious deterioration(SD)，Asset unserviceable(AU))

Experimental data

In Table 1, it can provide the data type which is used in the blow test.S In this table, it consists ID、pipe Diameter、water Pressure、pipe material、Distance to pump、Distance to Tank、Pipe Age、pipe Load and pipe Class.

Table 1. Test sample of data

ID	Diameter	Pressure	Material	D-P	D-T	Age	Load	Class
1	4	Low	CIP	1020	9940	40	High	NPC
2	8	Medium	DIP	1500	8430	50	High	AU
...	...	...	...	...	...	...	...	...

RESULT AND DISCUSSTION

In this test, Figure 5 shows that decision tree is four levels. Tree root is Load. Tree order is from Load(Low)、(D-P、D-T and Diameter) to (AU、MD、NPC、MD、NPC). The rules: Load (Low)-D-P (<3950)-AU is two test examples. Another rules: Load (Low)-D-P (>3950)-MD is two test examples. Load (Medium)-D-T (<5620)-NPC is three examples. Load (medium)-D-T (>5620)-MD is three examples. Load (High)-Diameter (<10)-NPC is six examples. Load (High)-Diameter (>10)-MD is four examples. Finally, the experimental precision degree is 60%.

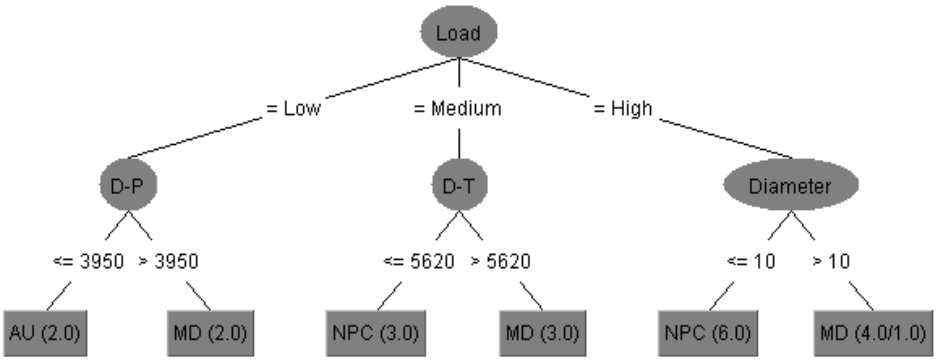

Figure 2. Decision view

From Figure 2, we can evidently get conclusion, the correct classification proportion of all samples is 60%. Total number of Instances is 20. Correctly Classified Instances 60%, Kappa Statistic is 0.322. Mean absolute error is to establish all rules in this sample test. Precision rate in this test reach to 74.44%.

Table 2. Result of statistical analysis in three levels

Number of Instances	Correctly Classified Instances	Kappa statistic	Mean absolute error
20	60%	0.322	25.56%

In table 3，TP (NPC)rate , TP (MD)rate and TP (AU)rate are 0.8, 0.625 and 0 separately. Most importantly, assessment parameter is ROC Area. Generally, ROC Area is grater than 0.7, experimental level is well, on the contrary, ROC Area is less than 0.5, experimental level is bad. So in this experiment, ROC Area is 0.755(NPC), 0.651(MD) and 0.889(AU). The test can meet precision need. The result shows large percent of TP with small FP Rate. The value of F-Measure is (NPC)0.778、(MD)0.556 and (AU)0 separately.

Table 3. Analysis of statistical result in four levels

TP Rate	FP Rate	F-Measure	ROC Area	Class
0.8	0.1	0.778	0.755	NPC
0.625	0.414	0.556	0.651	MD
0	0.111	0	0.889	AU

Figure 3. Confusion matrix in three levels

Secondary diagonal of this graph in Figure 3, the line represents the purity of classification in the test. Congruent relationship is obvious, so the establishing rules can be applied into classification of faults points in water supply network.

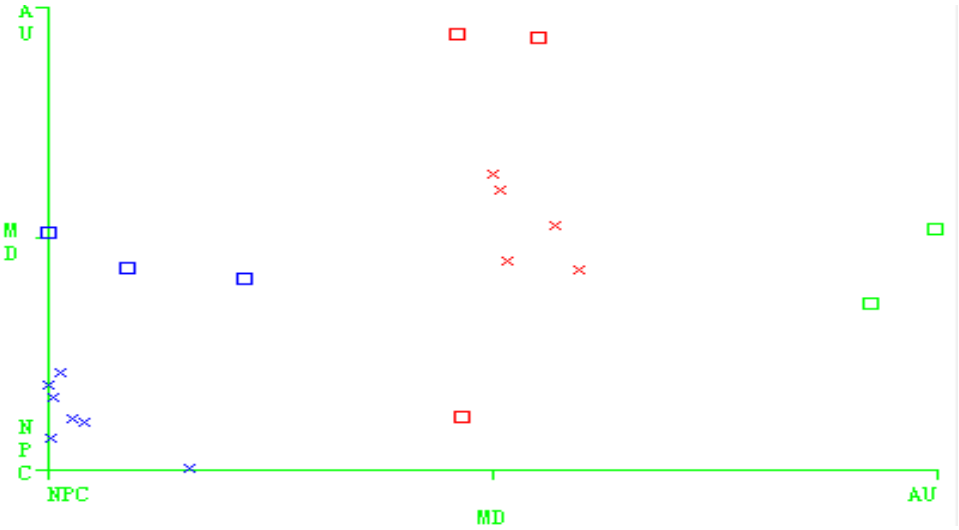

Figure 4. X：Class (Num.) Y: Predicted Class (Num.)

In Figure 4, it is easy to get conclusion, 8 examples cannot classify correctly in this test. Total number of square symbol in Figure 4 is 8. These 8 samples classify incorrectly.AU-MD is two samples, MD-AU is two samples, MD-NPC is one sample and NPC-MD is three examples. So the number of correct classification is 12. The successful classification percentage is 60%.

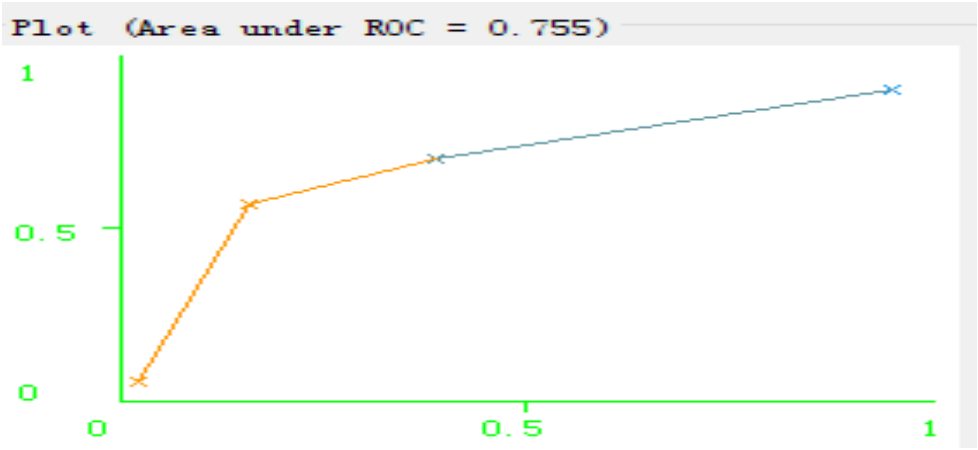

Figure 5. ROC curve for NPC(X: FP, Y: TP)

In Figure 5, when TP rate is equal to 0.4, FP rate is 0.1, so Near Perfect Condition of pipe can be well selected, that means：Category of NPC can be classified. Besides, the ROC area reaches to 0.755, this show that classification is effective in this test.

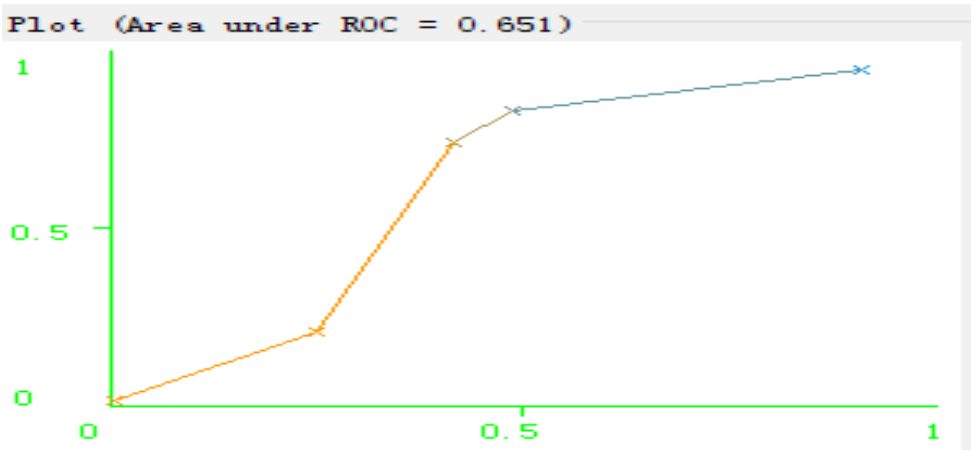

Figure 6. ROC curve for SD(X: FP, Y: TP)

In Figure 6, when TP rate is equal to 0.5, FP rate is 0.3, so Series Damage of pipe cannot be well selected, that means：When TP rate gets to 0.5, the FP rate at same time reaches 0.3. Category of SD cannot be classified. Besides, the ROC area reaches to 0.651, the ROC area is lower than NPC.

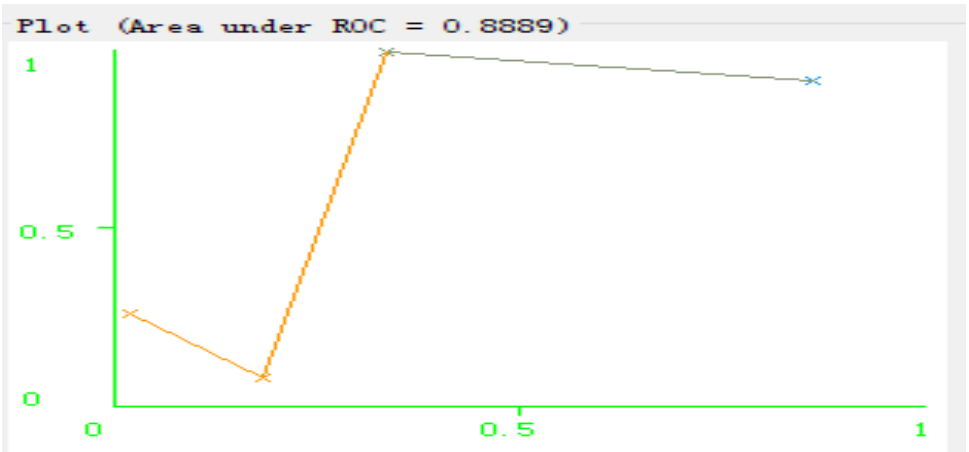

Figure 7. ROC curve for AU(X: FP, Y: TP)

In Figure 7, when TP rate is nearly equal to1, FP rate is 0.3, so Asset unserviceable of pipe

can be well selected, that means：Category of AU can be classified. Besides, the ROC area reaches to 0.8889, this show that classification is effective in this test.

CONCLUSION

In this paper, adopting the decision tree method is to classify successfully the fault data in water supply network. The levels of decision tree is three, it can smoothly get results: a series of rules about classification in pipe data, for example: Load(Low)、(D-P、D—T and Diameter)、(AU、MD、NPC、MD、NPC)、（NPC and MD）. In further work, using these rules can classify input data(new detection data)， it is equal to establish prediction method for conducting the further work. Most important, in this test, a series of assessment methods （ROC） are to assess the classification.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support of the Shenzhen Peacock Plan (KQCX20130628155525052), Shenzhen Fundamental Research Program (JCYJ20120616213618826), and Seventh Framework Programme (FP7) Marie Curie Actions (PIRSES-GA-2012-318985).

REFERENCE

[1] Dalius Misiunas, “Failure monitoring and asset condition assessment in water supply system”[D],Sweden: Lund University, (2005).

[2] Jilong Sun, Ronghe Wang, Xiaoxue Wang, Haibo Yang, Junhui Ping, “Spatial cluster analysis of bursting pipes in water supply networks”,12th International Conference on Computing and control for the water Industry,(2013).

[3] Gary Stein ， Bing Chen ， Annie S. Wu ， Kien A. Hua , “Decision tree classifier for network intrusion detection with GA-based feature selection”, ACM-SE 43 Proceedings of the 43rd annual Southeast regional conference - Volume 2 Pages 136-141,(2013).

[4] Ricardo Cerri, Gisele L. Pappa, Andre Carlos P.L.F. Carvalho, Alex A.Freitas, “An extensive evaluation of decision tree-based hierarchical multilabel classification methods and performance measures”, Computational Intelligence, Volume 27,(2013).

[5] Sanjay Kumar Malik, Sarika Chaudhary, “Comparative study of decision tree algorithms for data analysis”, International Journal of research in Computer Engineering and Electronic. Page1 Vol: 2, (2013).

[6] Quinlan, J. R. “C4.5: Programs for Machine Learning”. Morgan Kaufmann Publishers, (1993).

[7] Quinlan, J.R., “Unknown attribute values in induction. Proceed-ings of the sixth International Machine Learning Workshop” (PP.164-168). San Mateo, CA: Morgan Kaufmann, (1989).

[8] Ian H. Witten, Eibe Frank , Mark A.Hall, “Date Mining practical machine learning tools and techniques”, Third Edition Elservier, (2011).