DOI: 10.2429/proc.2016.10(1)005 2016;10(1)

Małgorzata KUTYŁOWSKA[1]

SUPPORT VECTOR MACHINES AND NEURAL NETWORKS FOR FORECASTING OF FAILURE RATE OF WATER PIPES

METODA WEKTORÓW NOŚNYCH I SIECI NEURONOWE DO PRZEWIDYWANIA WSKAŹNIKA AWARYJNOŚCI PRZEWODÓW WODOCIĄGOWYCH

Abstract: The failure rate of water pipes was predicted using support vector machines (SVMs) and artificial neural networks (ANNs). Both algorithms are regression methods of so called machine learning. Operational data from the time span 2001-2012 were used for forecasting purposes. The length, diameter and year of construction of the distribution pipes and the house connections were treated as the independent variables. The computations were carried out using the Statistica 12.0 software.

Keywords: pipelines, prediction, radial basis functions

Introduction

Failure frequency is one among other indicators taken into account during the assessment of the reliability level of water supply systems [1, 2]. Nowadays, typical analysis of exploitation data related to the number and kinds of damages should be concerned together with the mathematical modelling. There are a lot of models which could be used for failure rate prediction [3-5]. They help us to assess the deterioration level of water conduits relatively quickly. The regression methods, for example support vector machines (SVMs) and artificial neural networks (ANNs) are nowadays very popular in solving many complex engineering problems [6-8]. It is the reason why such algorithms, based on radial basis functions (RBF), were used in this study to predict the values of failure rate of distribution pipes and house connections. The principal aim of this research was to find out, if an artificial neural network and support vector machine of the RBF type would predict (with an error acceptable for engineering purposes) the failure frequency indicator of water conduits

Materials and methods

The failure rate [λ, fail./(km·a)] of the house connections and the distribution pipes was predicted using the RBF-SVM and the RBF-ANN methods. The two approaches were based on radial basis functions. Exploitation data for the years 2001-2012 received from the water utility were used for forecasting purposes. In relation to SVM modelling, the whole data set was randomly divided into two equal (50%) subsets. The training and the testing sample had 147 data each for the house connections as well as respectively 124 and 125 data for the distribution pipes. The model was built using the training data and then it was tested on

[1] Faculty of Environmental Engineering, Wroclaw University of Science and Technology, Wybrzeże S. Wyspiańskiego 27, 50-370 Wrocław, Poland, phone +48 71 320 40 84, email: malgorzata.kutylowska@pwr.edu.pl
[*] Contribution was presented during ECOpole'16 Conference, Zakopane, 5-8.10.2016

different sample. In relation to ANNs, the methodology was a little bit different because of the peculiarities of such kind of modelling. The artificial neural network learning process consisted of several stages: a training stage (50% of the data), followed by a testing stage (25% of the data) and finally, the validation (25% of the data) of the created models. In the considered case, the whole data set (294 data for the house connections and 249 data for the distribution pipes) was used to learn the ANN. The training, testing and validation samples were chosen randomly from the whole data set. The prognosis was done on the basis of the unknown previously data set (created separately). The models were built separately for the distribution pipes and house connections. The calculations were done in the programme Statistica 12.0.

The relation between the dependent variables (the predicted value) and the independent variable need not to be known because the SVM method is a kind of nonparametric regression algorithm. V-fold cross validation was used to find the optimal model parameters [9]. Tenfold (V = 10) cross validation was used in the considered problem, whereby it was possible to select proper values for such parameters (learning constants) as capacity (C) and epsilon (ε), since they are not known *a priori*. In relation to artificial neural networks, model parameters (*eg* the number of hidden neurons and the type of activation functions) are determined during ANN learning using a suitable training algorithm. Many ANN models, for which the number of hidden neurons ranged from 1 to 30, were tested. The model characterized by the smallest mean-square error and the best fit between the real data and the predicted ones was selected. The results presented later in this paper are for this chosen optimal ANN model.

In both the methods the independent variables were: length (L_r, L_p), diameter (D_r, D_p) the year of construction (Y_r, Y_p) of the distribution pipes and the house connections.

Results and discussion

The parameters of the built ANN and SVM models for the different kinds of water pipelines are presented in Table 1. The validation error was considered for selecting an SVM model most accurately forecasting the failure rate. The validation error for the house connections and distribution pipes equalled to respectively 0.11 and 0.08. Nevertheless, the failure rate prediction on the basis of the testing sample was not satisfactory from the predicted/real data fit point of view. Moreover, the number of SVMs for the distribution pipes was high and as much as 82% of them were localized SVMs, *ie* with weights equal to ± the capacity value (Table 1), indicating a more complicated model structure. In the case of any kind of modelling, it is necessary to answer the question whether the aim is to obtain a perfect data fit at any cost, *ie* at the expense of model architecture complication, or rather to reveal the correlations between the dependent and independent variables.

In the case of the ANN models, Pearson's correlation coefficient (R), a determination coefficient (R^2) and a relative mean-square prediction error (amounting to about 20% for the distribution pipes and the house connections) would be compared. The value of this error is rather high in comparison to other results obtained using multilayer perceptron instead of radial basis functions. According to the literature [10] RBF ANNs were also less useful for hourly water demand prediction than the multilayer perceptron. Despite the fact that there were three times more hidden neurons in the house connections model than in the distribution pipes model (Table 1), the prediction results are worse and characterized by larger

discrepancies between the experimental and forecasted data (Figs. 1 and 2). Because of the nature of RBF ANNs, the activation functions and the training method were pre-imposed, which also can have a bearing on modelling quality in comparison with, *eg* artificial neural networks using the multilayer perceptron, where it is possible to use several different functions, such as the sigmoidal function, the exponential function and so on [11].

Table 1

Parameters of SVM and ANN models

Type of conduit/parameter	Distribution pipes	House connections
SVM model		
Gamma	0.333	0.333
Capacity (C)	3	1
Epsilon (ε)	0.2	0.5
Number of support vectors (localized)	56 (46)	14 (7)
Cross-validation error	0.081	0.110
ANN model		
Number of hidden neurons	8	27
Activation functions: hidden/output layer	Gaussian/linear	Gaussian/linear
Training algorithm	RBFT	RBFT
Correlation coefficient (learning/prognosis step)	0.956/0.859	0.997/0.897
Determination coefficient (learning/prognosis step)	0.914/0.737	0.994/0.805

The results of failure rate prediction for the learning sample are presented in Table 2 while the ones for the testing sample (the SVM model) and the prognosis stage (the ANN model) are shown in Figures 1 and 2. Such distinction, between the testing sample for the SVM model and the prognosis stage for the ANN model, is necessary to draw, since the testing sample data and the prognosis stage data were unknown to the model previously. Using such approach it is possible to establish the quality of the model and its applicability to failure rate prediction.

Table 2

Results of failure rate prediction - learning step

Year	House connections			Distribution pipes		
	Experimental	ANN-RBF	SVM-RBF	Experimental	ANN-RBF	SVM-RBF
2001	0.94	0.95	1.17	0.34	0.36	0.38
2002	0.84	0.84	1.16	0.34	0.36	0.38
2003	1.59	1.58	1.26	0.50	0.47	0.48
2004	1.07	1.07	1.32	0.37	0.39	0.41
2005	1.00	1.00	1.32	0.57	0.48	0.52
2006	1.15	1.15	1.33	0.42	0.42	0.42
2007	0.83	0.80	1.12	0.31	0.30	0.33
2008	0.65	0.63	0.79	0.22	0.21	0.22
2009	0.61	0.62	0.70	0.25	0.25	0.24
2010	0.50	0.50	0.63	0.27	0.26	0.24
2011	0.23	0.33	0.57	0.10	0.21	0.20
2012	0.38	0.38	0.51	0.24	0.25	0.24

An analysis of Table 2 clearly shows that prediction of failure rate λ of house connections using the RBF-ANN model is better than for the RBF-SVM model. For the

distribution pipes the differences in failure rate predictions between the two modelling methods are not so significant and one can say that these two methods are equally effective, as indicated by the fact that coefficients R = 0.96 and R^2 = 0.92 are identical for both methods. Considerable errors (over 100%) occur in the estimates of the failure rate for the distribution pipes only in 2011, which is undoubtedly due to the fact that this indicator is very much different from the values for the other analyzed years (could be treated as outlier). A similar situation is observed during forecasting the failure rate for the house connections in 2011.

Parallel correlations (as the ones described above) between real and forecasted data were found at the testing step (the SVM model) and the prognosis step (the ANN model), as shown in Figures 1 and 2. In relation to the house connections, good agreement between experimental and predicted values is generated by the ANN model, but for some years (*eg* 2003, 2006 and 2009) the discrepancies are much larger than the ones observed at the learning step. Despite many divergences, the trend in the changes of the forecasted values is similar to the trend in the variation of real values. In the years 2006-2008 a similar configuration is observed for the SVM model, but most of the λ values are much higher than the real ones.

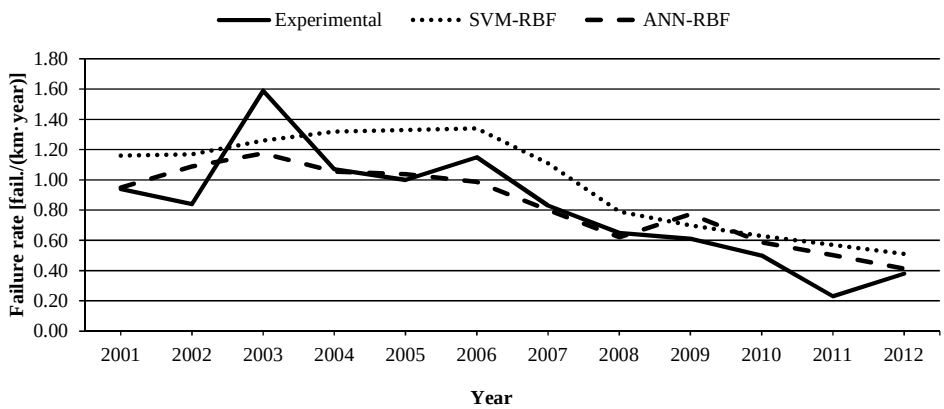

Fig. 1. Prediction results of house connections' failure rate - testing-SVM/prognosis-ANN

The estimation of the failure rate of the distribution pipes (Fig. 2) by the SVM method and the ANN method was characterized by acceptable agreement between the forecasted and real values in both cases. The Pearson correlation coefficient for the SVM model and the ANN model equalled to respectively 0.96 and 0.86, indicating that the SVM method is slightly better for predicting the failure rate of distribution pipes than the ANN method. The different situation is observed for house connections. It should also be noted that the results of predicting the failure rate of the distribution pipes and the house connections (Table 2, Figs. 1 and 2) by means of the SVM-RBF model are very similar for both the learning sample and the testing sample. Whereas the results of learning and prognosis by the ANN-RBF model show larger discrepancies for both types of pipelines. Even though at the learning stage the agreement between the experimental and forecasted values is satisfactory, the prognosis stage (using new data) gives a larger (but still acceptable from the engineering point of view) error. This is evidence of greater effectiveness of RBF-based

training by means of SVMs than ANNs. However, at the present step of the studies it cannot be explicitly pointed out which of the algorithm is better and should be widely adopted in the modelling of the failure rate of water pipelines. Further research in this area is needed, also on operational data from other water utilities, permitting more in-depth analyses and broader generalizations.

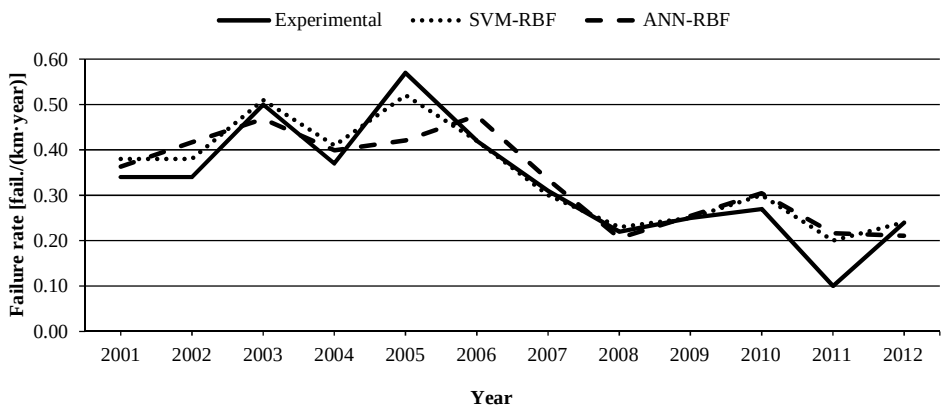

Fig. 2. Prediction results of distribution pipes' failure rate - testing-SVM/prognosis-ANN

Conclusions

Support vector machines and artificial neural networks were used for the prediction of the failure rate of the house connections and distribution pipes in one of the Polish city. Exploitation data for the time span 2001-2012 were chosen for prediction purposes. The problem is quite important for the correct and quick approximation of the reliability level. Both methods can be used to establish the failure rate of municipal systems. A relatively large database must be available in order to identify the relevant correlations between dependent and independent variables.

The optimal SVM model had gamma coefficient equalled to 0.33 for both the distribution pipes and the house connections. The capacity C and the number of SVMs were respectively 3 and 4 times greater in relation to the model describing the failure frequency of the distribution pipes. The error of the V-fold cross validation amounted to 0.110 and 0.081 for the model describing the failure rate of respectively the house connections and distribution pipes. The length, diameter and year of laying the water pipes in the ground were treated as independent variables. The optimal ANN-RBF model contained 27 and 8 hidden neurons for respectively house connections and distribution pipes. The coefficients R and R^2 are slightly higher at the step of learning than during prognosis of ANN model.

Acknowledgments

The work was realized within the allocation No. B50519 awarded for Faculty of Environmental Engineering Wroclaw University of Science and Technology by Ministry of

Science and Higher Education. The grant was allocated for scientific researches of young scientists in years 2015-2016.

References

[1] Tscheikner-Gratl F, Sitzenfrei R, Rauch W, Kleidorfer M. Struct Infrastruct Eng. 2016;12(3):366-380. DOI: 10.1080/15732479.2015.1017730.
[2] Kowalski D, Miszta-Kruk K. Eng Failure Analysis. 2013;35:736-742. DOI: 10.1016/j.engfailanal.2013.07.017.
[3] Scheidegger A, Leitao JP, Scholten L. Water Res. 2015;83:237-247. DOI: 10.1016/j.watres.2015.06.027.
[4] Nishiyama M, Filion Y. Can J Civ Eng. 2014;41(10):918-923. DOI: 10.1139/cjce-2014-0114.
[5] Tscheikner-Gratl F, Sitzenfrei R, Rauch W, Kleidorfer M. Struct Infrastruct Eng. 2016;12(3):366-380. DOI: 10.1080/15732479.2015.1017730.
[6] Kolasa-Więcek A. Ecol Chem Eng S. 2013;20(2):419-428. DOI: 10.2478/eces-2013-0030.
[7] Aydogdu M, Firat M. Water Resour Manage. 2015;29:1575-1590. DOI: 10.1007/s11269-014-0895-5.
[8] Rodolfi E, Servili F, Magini R, Napolitano F, Russo F, Alfonso L. Proced Eng. 2014;89:648-655. DOI: 10.1016/j.proeng.2014.11.490.
[9] Statistica 12.0, Electronic Manual. http://www.statsoft.pl/textbook/stathome_stat.html?http%3A%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstmachlearn.html
[10] Siwoń Z, Cieżak W, Cieżak J. Modele neuronowe szeregów czasowych godzinowego poboru wody w osiedlach mieszkaniowych [Neural network models of hourly water demand time series in housing areas]. Ochr Środ. 2011;33(2):23-26. http://www.os.not.pl/docs/czasopismo/2011/2-2011/Siwon_2-2011.pdf.
[11] Kutyłowska M. Eng Failure Analysis. 2015;47:41-48. DOI: 10.1016/j.engfailanal.2014.10.007.

METODA WEKTORÓW NOŚNYCH I SIECI NEURONOWE DO PRZEWIDYWANIA WSKAŹNIKA AWARYJNOŚCI PRZEWODÓW WODOCIĄGOWYCH

Wydział Inżynierii Środowiska, Politechnika Wrocławska

Abstrakt: Wskaźnik awaryjności przewodów wodociągowych przewidywano za pomocą metody wektorów nośnych (SVM) i sztucznych sieci neuronowych (SSN). Oba algorytmy należą do metod regresyjnych, nazywanych metodami uczenia maszyn. Dane eksploatacyjne z lat 2001-2012 zostały wykorzystane w celach predykcji. Długość, średnica i rok budowy przewodów rozdzielczych i przyłączy były zmiennymi niezależnymi. Obliczenia przeprowadzono w programie Statistica 12.0.

Słowa kluczowe: rurociągi, przewidywanie, radialne funkcje bazowe