Data mining methods for prediction of air pollution

-Extended summary

Krzysztof SIWEK
Warsaw University of Technology, POLAND
ksiwek@iem.pw.edu.pl

Stanislaw OSOWSKI
Warsaw University of Technology
Military University of Technology, Warsaw, POLAND
sto@iem.pw.edu.pl

***Abstract*—The paper discusses the methods of data mining for prediction of air pollution. Two problems in such prediction are important: the generation and selection of the prognostic features, and final prognosis of the pollution level for the next day on the basis of the data of the previous day. In this paper we analyze and compare two methods of feature selection. One applies the genetic algorithm, and the second the linear method of stepwise fit. On the basis of such analysis we are able to select the most important features influencing the prediction. As a mathematical tool for final prediction we apply the neural networks. Three different solutions will be compared: the multilayer perceptron (MLP), radial basis function (RBF) network and support vector machine (SVM).**

Keywords—time series forecasting; feature selection;neural networks, computational intelligence

EXTENDED SUMMARY

The important task in providing the proper quality of our life is protection of environment. This problem is strictly associated with the early prediction of the air pollution, concerning the level of CO_2, NO_x, PM_{10}, O_3. The paper will discuss the numerical aspects of the problem concerning the methods of data mining used in building the model of prediction of the air pollution.

The most important is identification of the environmental factors which have the highest impact on the level of pollution. This problem is known as the feature selection. Among many parameters measured by the meteorological stations (temperature, wind, humidity, insolation) at different hours of the day we have to select those which are most important from the prediction point of view.

In this paper we analyze and compare two methods of selection, which are treated as the most powerful. One applies the genetic algorithm (the nonlinear approach) and the second the linear method of stepwise fit. In genetic algorithm (GA) application each chromosome represents one feature (the value of one means inclusion of the feature in the prediction set and zero – deletion from the actual set of features). GA consists of selecting parents for reproduction, performing crossover with the parents, and applying the operation of mutation to the bits representing children.

Each chromosome is associated with the input vector **x** applied to the neural predictor (the value 1 means real inclusion of the feature and zero – no such feature in a vector). The predictor is trained on the learning data set and then tested on the validation data. The testing error function using the validation data forms the basis for the definition of the fitness function. The fitness is defined as the error function taken with minus sign. The genetic algorithm maximizes the value of the fitness function (equivalent to the minimization of the error function) by performing the subsequent operations of selection of parents, the crossover among the parents and finally the mutation. Figure 1 presents the applied scheme of genetic operations used for feature selection.

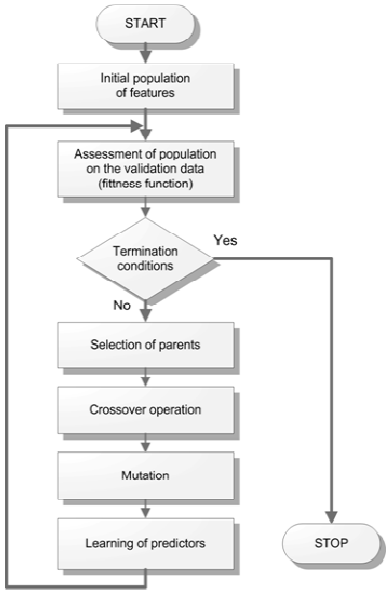

Fig. 1. The illustration of genetic system of feature selection

Genetic algorithms are a very effective way of finding a reasonable solution to a complex problem of feature selection. They do an excellent job of searching through a large and complex search space for which little is known.

Contrary to genetic approach we investigate also the traditional linear method, generally known as stepwise linear fit. It is the method based on the successive linear regression, in which we systematically add and remove the successive candidate features to the set of input attributes of the linear model of the process.

The impact of the actually investigated feature on the modeled process is measured by the value of its coefficient in linear regression and its change at the process of adding and removing the next features. In each step of adding or removing the feature to the set, the F-statistics is determined on the basis of which we decide to leave or remove it from the feature set.

The procedure is stopped when adding or removing any feature does not lead to increase of the accuracy of the linear model. Contrary to the genetic algorithm application the stepwise fit provides only the local optimality of solution. However, in spite of it, this method has a reputation of high quality.

On the basis of this analysis we are able to select the most important features taking part in the prediction. The results of application of both selection methods will be analyzed and compared. The selected features will be used as the input information delivered to the predicting tools.

As a mathematical tools for prediction we apply here the neural networks. Three different solution of predictor will be compared: the multilayer perceptron (MLP), radial basis function (RBF) network and support vector machine (SVM). The results of prediction of 4 main air pollutants (CO_2, NO_2, PM_{10} and O_3) will be presented and discussed.

REFERENCES

[1] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Boston: Addison-Wesley, 1989.

[2] G. Grivas, A. and Chaloulakou, "Artificial neural network models for predictions of PM10 hourly concentrations in greater area of Athens", *Atmospheric Environment*, vol. 40, 2006, pp. 1216-1229.

[3] S. Haykin, *Neural networks, a comprehensive foundation*, New York: Macmillan College Publishing Company, 2000.

[4] M. Misiti, G. Oppenheim, J.M. Poggi, and Y. Misiti, *User manual of Matlab*, Natick : MathWorks, 2010.

[5] L. Nikias and A.P. Petropulu, *Higher-order spectral analysis – a nonlinear signal processing framework*, NJ: Englewood Cliffs, 1993.

[6] B. Schölkopf and A. Smola, *Learning with kernels*, Cambridge: MIT Press, 2002.

[7] K. Siwek, S. Osowski, and B. Swiderski, "Study of dynamics of atmospheric pollution and its association with environmental parameters", *Studies in Computational Intelligence*, vol. 459, pp. 179-190, Heidelberg: Springer, 2013.