A feature selection method for air quality forecasting

Luca Mesin[1], Fiammetta Orione[1], Riccardo Taormina[1], Eros Pasero[1],

[1] Department of Electronics, Politecnico di Torino, Torino, Italy
{luca.mesin,fiammetta.orione,riccardo.taormina,eros.pasero}@polito.it

Abstract. Local air quality forecasting can be made on the basis of meteorological and air pollution time series. Such data contain redundant information. Partial mutual information criterion is used to select the regressors which carry the maximal non redundant information to be used to build a prediction model. An application is shown regarding the forecast of PM_{10} concentration with one day of advance, based on the selected features feeding an artificial neural network.

Keywords: Air pollution, artificial neural network, partial mutual information, information theory, input variable selection

1 Introduction

European laws require the analysis and implementation of automatic procedures to prevent the risk for the principal air pollutants to be above alarm thresholds in urban and suburban areas (e.g. the Directive 2002/3/EC for ozone or the Directive 99/30/CE for the particulate matter with an aerodynamic diameter of up to 10 μm called PM_{10}).
Two-three days forecasts of critical air pollution conditions would allow an efficient choice of countermeasures to safeguard citizens' health.

Different procedures have been developed to forecast the time evolution of air pollutant concentration. Meteorological data are usually included in the model as air pollution is highly correlated with them (Cogliani, 2001). Indeed, pollutants are usually entrapped into the planetary boundary layer (PBL), which is the lowest part of the atmosphere. It is directly influenced from soil interaction, as friction and energetic exchange.

Important meteorological parameters involved in air pollution dynamics are air temperature, relative humidity, wind velocity and direction, atmospheric pressure, solar radiation and rain. Principal air pollutants whose concentration should be monitored are Sulphur Dioxide SO_2, Nitrogen Dioxide NO_2, Nitrogen Oxides NO_x, Carbon Monoxide CO, Ozone O_3 and Particulate Matter PM_{10}. Local forecasting can be performed in real time and with low cost technology analyzing time series of weather data and air pollution concentrations. Meteorological and air pollution data contain redundant information. The introduction of irrelevant and redundant information is detrimental for a prediction model, as the training and processing time are increased and noise is introduced, so that accuracy is reduced (May et al., 2008). Hence, a careful selection of useful predictors is needed

In this paper, we are concerned with a specific method to perform local prediction of air pollution. Our analysis carries on the work already developed by the NeMeFo (Neural Meteo Forecasting) research project for meteorological data short-term forecasting (Pasero et al., 2004). Special attention is devoted to the selection of optimal, non redundant features (Guyon and Elisseeff, 2003) from meteorological and air pollution data, decisive to describe the evolution of the system.

Our approach for feature selection is based on the partial mutual information (PMI) criterion (Sharma 2000). Once selected decisive features, prediction is performed using an Artificial Neural Network (ANN). ANNs have been often used as a prognostic tool for air pollution (Perez et al., 2000; Božnar et al., 2004; Cecchetti et al., 2004; Slini et al., 2006; Karatzas et al., 2008).

The method is applied on four hours hourly data, measured by a station located in the urban area of Goteborg, Sweden (Goteborgs Stad Miljo). The aim of the analysis is forecasting the average day concentration of PM_{10}.

2 Methods

2.1 Input variable selection with Partial Mutual Information

2.1.1 Mutual Information

The input variable selection algorithm relies on Partial Mutual Information criterion first introduced by Sharma (2000). The proposed algorithm evolves the concept of Mutual Information (Cover and Thomas, 1991) between two random variables, suitable measure of dependence between signals in nonlinear systems. The mutual information MI(X;Y) of two random variables can be defined as the reduction in uncertainty with respect to Y due to observation of X. For continuous random variables, the MI score is defined as:

$$MI(X;Y) = H(X) + H(Y) - H(X,Y) \tag{1}$$

where *H* is the information entropy defined as:

$$H(Z) = -\int f_Z(z) \ln f_Z(z) dz \tag{2}$$

Mutual Information is always nonnegative and it equals zero only when X and Y are independent variables.

For any given bivariate random sample, the discrete estimation of (1) can be written as:

$$MI(X;Y) = \frac{1}{n} \sum_{i=1}^{n} \ln \frac{f_{X,Y}(x_i, y_i)}{f_X(x_i) f_Y(y_i)} \tag{3}$$

where x_i and y_i are the i^th^ bivariate sample data pair in a sample of size *n*, and $f_X(x_i)$, $f_Y(y_i)$ and $f_{X,Y}(x_i, y_i)$ are respective univariate and joint probability-densities estimated at the sample data points. Robust estimators for the joint and marginal probability density functions in (3) can be computed with Parzen method (Parzen, 1962; Costa et al., 2003):

$$\widehat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}K_h\left(x-x_i\right) \tag{4}$$

where *d* is the number of dimension of the considered random variable and the kernel *K* was assumed as Gaussian with standard deviation *h* (kernel bandwidth). A major issue in kernel density estimation is to find the value for the bandwidth. Small values could lead to data under-smoothing and the resulting MI score could be noise-sensitive thus. A large bandwidth tends to over-smooth the sample probability density-function and underestimate MI value consequently. Although algorithms have been developed to search for an optimal value of the bandwidth *h*, an appropriate first choice is given by the Gaussian reference bandwidth:

$$h=\left(\frac{4}{d+2}\right)^{\frac{1}{d+4}}\sigma n^{-\frac{1}{d+4}} \tag{5}$$

where σ is the standard deviation of the data sample.

2.1.2 Partial Mutual Information

Optimal input set for modeling a certain system can be defined selecting the variables with large Mutual Information with the output. However, this raises a major redundancy issue because the MI criterion does not account for the interdependencies between candidate variables. To overcome this problem, Sharma (2000) has developed an algorithm that exploits the concept of Partial Mutual Information (PMI), which is the nonlinear statistical analog of partial correlation. It represents the information between two observations that is not contained in a third one.

Let Y be an univariate random variable with observations y, X a candidate input with observations x, and Z the multivariate set of the input variables which have already been selected as predictors. The PMI score between candidate X and output Y is computed considering the residuals of both variables once the effects of the existing predictors Z have been taken into account. In other words the arbitrary dependence between variables is removed by computing for each x and y the residuals:

$$x'=x-E[x\,|\,z] \qquad y'=y-E[y\,|\,z] \tag{6}$$

where E[.] denotes the regression of the chosen variable based on the predictors z already selected (i.e., belonging to Z). Using the kernel density estimation approach, the output Y can be estimated as:

$$E[y\,|\,z]=\frac{1}{n}\frac{\sum_{i=1}^{n}y_iK_h(z-z_i)}{\sum_{i=1}^{n}K_h(z-z_i)} \tag{7}$$

The regression estimator $E[x\,|\,z]$ for the candidate predictors *x* to be possibly included in Z is written analogously.

2.1.3 Termination criterion

The above mentioned approach needs a criterion to assess whether each selected variable is indeed a significant predictor for the system output. Different methods to terminate the algorithm have been proposed, e.g. bootstrap estimation technique (Sharma, 2000) or less computationally intensive approaches (May et al., 2008).

This work applied the Hampel test criterion, which is a modification of the Z-test commonly adopted to find outliers within a population of observed values (May et al., 2008).

2.2 Prediction method based on Artificial Neural Network

The prediction algorithm was based on feedforward ANNs with a single hidden layer and a single output (the predicted concentration of pollutant). The hyperbolic tangent function was used as activation function. The Levenberg-Marquardt algorithm (Haykin, 1999) was used to estimate iteratively (using backpropagation) the synaptic weights of the ANN, minimising the sum of the squares of the deviation between the predicted and the target values on a training set.

For prediction purposes, time is introduced in the structure of the neural network. For just further prediction, the desired output at time step n is a correct prediction of the value attained by the time series at time n+1:

$$y_{n+1} = \varphi\left(\vec{w}\cdot\vec{z}_n + b\right) \qquad \textbf{(8)}$$

where the vector of regressors $\vec{z}_n$ includes information available up to the time step n. Selected features up to time step n were used, obtaining a non linear autoregressive with exogenous inputs (NARX) model (Sjöberg et al., 1994).

3 Results

The input selection method was tested on a dataset for the prediction of the daily average PM10 in Goteborg, urban area. Apart from values of PM_{10} itself, other candidate variables were both meteorological (air temperature, relative humidity, atmospheric pressure, solar radiation, rainfall, wind speed and direction) and chemical ones (SO_2, NO_x, NO_2, O_3 and CO). Daily averages, three previous days maximum and minimum have been included for each variable and 24-hour cumulated variable only for the rain. In this way, the candidate pool was made of more than 120 features, with observations ranging from the beginning of 2002 to the end of 2005. First three years of the database have been used for selecting best input variables, while last year recordings have been arranged for testing the performances of the ANN developed for prediction. The results of the PMI algorithm on the candidate dataset are reported in Fig. 1A. Only three variables were selected using Hampel test, namely the maximum, daily average and minimum concentration of PM_{10} recorded the day before. This entails a drastic reduction from the original pool of candidates.

Once the most significant features have been selected, an ANN has been developed to predict future values of the PM_{10} concentration, using the same dataset of the input selection algorithm for the training. The optimal ANN was found to have 8 hidden neurons, and the results on the test dataset are plotted in Fig. 1B. Root mean square error and correlation coefficient on the test data set were respectively RMSE=6.24 μg/m^3 and CC = 0.91, showing an overall good fitting of the ANN output.

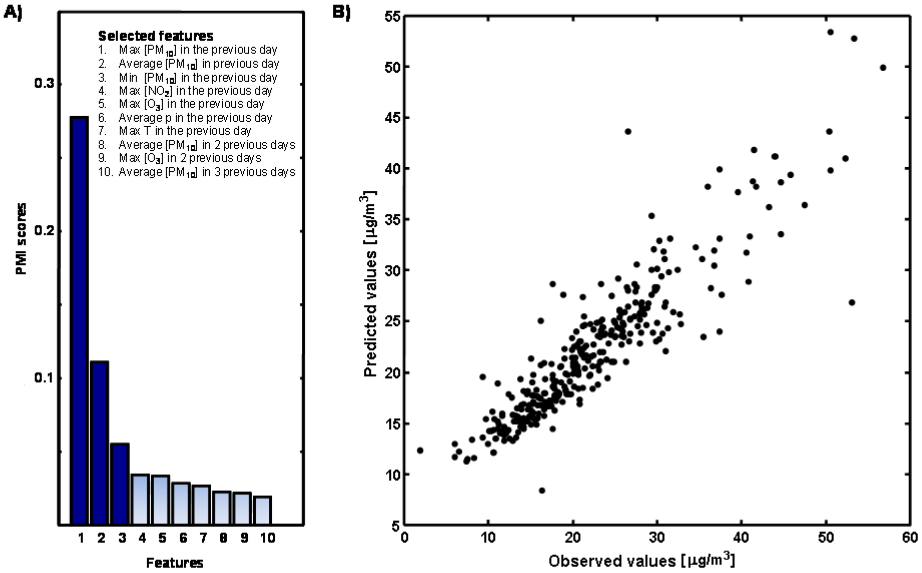

Fig. 1. A) Partial mutual information scores for the 10 most significant features. The first three features were selected by the Hampel test. B) Comparison between the observed and predicted PM_{10} concentration.

4 Discussion

Coarse fraction of PM_{10} derives most from natural source, as wind erosion, sea salt spray, wood waste. Finest fraction of PM_{10} derives most from human activities as transport action or construction ones. Nitrates and sulphur dioxides are found predominantly in fine fraction, less than 2.5 μm in diameter.

Our model selected previous day maximum, minimum and average concentration of PM_{10} as the most important features of which taking account in PM_{10} daily monitoring. This suggests that most of the information needed to forecast future values is indeed contained in the trends of the pollutant itself. In addition, two or more days before observations do not seem to have explaining potential for the prediction. Nitrates are found to have a relative high PMI, although the score is below the threshold for being selected. Meteorological variables have not been selected as well. Due to good prediction performance, this may imply that the meteorological effects as well as chemical interactions between the pollutants might be included in the information provided by the PM_{10} features.

Acknowledgments. This work was sponsored by the national project AWIS (Airport Winter Information System), funded by Piedmont Authority, Italy.

References

1. Božnar, M.Z. ; Mlakar, P.J., Grašič, B.: Neural Networks Based Ozone Forecasting. Proceeding of 9th Int. Conf. on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, June 1-4, 2004, Garmisch-Partenkirchen, Germany (2004)
2. Cecchetti, M., Corani, G., Guariso, G.: Artificial Neural Networks Prediction of PM10 in the Milan Area, Proc. of IEMSs 2004, University of Osnabrück, Germany, June 14-17 (2004)
3. Cogliani, E.: Air pollution forecast in cities by an air pollution index highly correlated with meteorological variables. Atm. Env., 35(16), 2871-2877 (2001)
4. Costa, M., Moniaci, W., Pasero, E.: INFO: an artificial neural system to forecast ice formation on the road, Proceedings of IEEE International Symposium on Computational Intelligence for Measurement Systems and Applications, pp. 216–221, July 29-31 (2003)
5. Cover, T.M., Thomas, J.A. Elements of information theory. John Wiley & Sons, New York, NY (1991)
6. Goteborgs Stad Miljo, http://www.miljo.goteborg.se/luftnet/
7. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection, The Journal of Machine Learning Research, 3, 1157-1182, January (2003).
8. Haykin, S.: Neural Networks: A Comprehensive Foundation. Prentice Hall (1999).
9. Hyvarinen, A.: Survey on Independent Component Analysis. Neural Computing Surveys, 2, 94-128 (1999)
10. Karatzas, K.D., Papadourakis, G., Kyriakidis, I.: Understanding and forecasting atmospheric quality parameters with the aid of ANNs. Proceedings of the IJCNN, Hong Kong, China, pp. 2580-2587, June 1-6 (2008)
11. May, R.J., Maier, H.R., Dandy, G.C., Gayani Fernando, T.M.K.: Non-linear variable selection for artificial neural networks using partial mutual information, Envir. Mod. And Soft., 23, 1312-1326 (2008)
12. Parzen, E.: On Estimation of a Probability Density Function and Mode. Annals of Math. Statistics, 33, 1065-1076, (1962)
13. Pasero, E., Moniaci, W., Meindl, T., Montuori, A.: NeMeFo: Neural Meteorological Forecast. Proceedings of SIRWEC 2004, 12th International Road Weather Conference, Bingen (2004)
14. Perez, P., Trier, A., Reyes, J.: Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. Atmospheric Environment, 34, 1189-1196 (2000)
15. Sharma, A.: Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: 1 - A strategy for system predictor identification. Journal of Hydrology, 239, 232-239 (2000)
16. Sjöberg, J., Hjalmerson, H., Ljung, L.: Neural Networks in System Identification. Preprints 10th IFAC symposium on SYSID, Copenhagen, Denmark. 2, 49-71 (1994)
17. Slini, T., Kaprara, A., Karatzas, K., Moussiopoulos, N.: PM_{10} forecasting for Thessaloniki, Greece, Environmental Modelling & Software, 21(4), 559-565 (2006)