Int. J. Appl. Math. Comput. Sci., 2016, Vol. 26, No. 2, 467–478
DOI: 10.1515/amcs-2016-0033

DATA MINING METHODS FOR PREDICTION OF AIR POLLUTION

KRZYSZTOF SIWEK [a], STANISŁAW OSOWSKI [a,b,*]

[a]Faculty of Electrical Engineering
Warsaw University of Technology, pl. Politechniki 1, 00-661 Warsaw, Poland
e-mail: {ksiwek,sto}@iem.pw.edu.pl

[b]Faculty of Electronic Engineering
Military University of Technology, ul. Kaliskiego 2, 00-908 Warsaw, Poland

The paper discusses methods of data mining for prediction of air pollution. Two tasks in such a problem are important: generation and selection of the prognostic features, and the final prognostic system of the pollution for the next day. An advanced set of features, created on the basis of the atmospheric parameters, is proposed. This set is subject to analysis and selection of the most important features from the prediction point of view. Two methods of feature selection are compared. One applies a genetic algorithm (a global approach), and the other—a linear method of stepwise fit (a locally optimized approach). On the basis of such analysis, two sets of the most predictive features are selected. These sets take part in prediction of the atmospheric pollutants PM10, SO2, NO2 and O3. Two approaches to prediction are compared. In the first one, the features selected are directly applied to the random forest (RF), which forms an ensemble of decision trees. In the second case, intermediate predictors built on the basis of neural networks (the multilayer perceptron, the radial basis function and the support vector machine) are used. They create an ensemble integrated into the final prognosis. The paper shows that preselection of the most important features, cooperating with an ensemble of predictors, allows increasing the forecasting accuracy of atmospheric pollution in a significant way.

Keywords: computational intelligence, feature selection, neural networks, random forest, air pollution forecasting.

1. Introduction

An important task in providing the proper quality of our life is protection of the environment from air pollution (Bhanu and Lin, 2003; Brunelli *et al*., 2007; Grivas, 2006; Perez and Trier, 2001). This problem is strictly associated with early prediction of air pollution, concerning the level of SO2, NO2, O3 and particulate matters of diameters up to 10 μm (PM10). Actually, PM is of special importance for a European policy (the new European Air Quality Directive EC/2008/50) defining restrictions for yearly and 24 h average PM10 concentrations. To respect the short term limit values defined by these restrictions and diminish dangerous concentration levels, emission abatement actions have to be planned at least one day in advance. Moreover, according to EU directives, public information on the air quality status and on the predictable trend for the next days should also be provided. Hence, one day ahead forecasting is needed. The paper will discuss the numerical aspects of the air pollution prediction problem, concentrating on the methods of data mining used for building the most accurate model of prediction.

There are two main tasks to be solved. The first one is generation of the best prognostic features influencing the prediction, and the second—building the structure of the predicting system which provides the most accurate forecast. There are a number of papers devoted to this problem (Bhanu and Lin, 2003; Brunelli *et al*., 2007; Grivas, 2006; Agirre-Basurko *et al*., 2006; Mesin *et al*., 2010). However, most of them take into account only primary atmospheric variables (temperature, wind, humidity, etc.), on the basis of which the forecast is made. The derivatives of these variables, like the gradient, the estimated trend of their changes, the forecast made on the basis of such trends, etc., have not been used up to now, although their application might improve the quality of prediction. On the other hand, including all of them in the set of features increases the size of input attributes

*Corresponding author

and may lead to decreasing the generalization ability of the prognostic system. Therefore, special methods of detection of the most important factors influencing the prognosis are necessary. This task is known as the feature selection problem (Guyon and Elisseeff, 2003; Tan *et al*., 2006).

Various sets of potential features might be formed from the parameters measured by meteorological stations (temperature, wind, humidity, insolation) at different hours of the day. The contents of these sets should be analyzed to detect the features which are most important from the prediction point of view (Siwek *et al*., 2011; Osowski *et al*., 2009). In this paper, an analysis and comparison of two approaches to the feature selection will be presented. One applies a genetic algorithm (nonlinear approach) and the other—a linear method of stepwise fit. The former represents a global and the latter a local optimization method. Both the approaches determine the contents of the sets of input variables, treated as the most influential features in the prediction process. Because of different principles of operation the contents of both the sets are usually not the same.

The results of feature selection provide the input information to the system responsible for predicting the average level of air pollution on the next day. Two different systems of prediction will be studied here. In the first one, the features selected are applied to the random forest (RF) of decision trees, which performs two functions at the same time: regression (made by the individual decision trees) and integration (averaging the results of outputs of many decision trees).

In the second approach, the features selected create the inputs to the individual predictors, built on the basis of neural networks: the multilayer perceptron (MLP), the radial basis function (RBF) network and the support vector machine (SVM) of the Gaussian kernel. The universal approximation ability of these networks (Haykin, 2000; Scholkopf and Smola, 2002) will be exploited in this approach. All of them have the reputation of very good universal approximators. Their results are combined together in an ensemble providing the final prognosis of an increased accuracy. The numerical results of prediction of different air pollutants (PM10, SO2, NO2 and O3) will be presented and discussed.

2. Potential set of features

To build a numerical predictive model of any process it is necessary to define the set of input features (also called explanatory variables) on the basis of which the forecasting will be made (Sumi *et al*., 2012). This choice is made based a detailed analysis of the problem. It is known that the factors influencing the pollution level on the next day include atmospheric variables, such as temperature, wind, humidity, pressure, insolation and also the state of pollution following from the previous day. These primary variables form a natural set of parameters on the basis of which the secondary set, well associated with prediction of the pollution level on the next day, will be formed. This selected set of features may be created in a different way, using known mathematical operations, such as the derivative, the gradient, estimation of the trend of their change, extreme values, etc.

Another aspect that should be considered is the dependence of the pollution level on the season of the year and the type of day. Generally, weekdays would have more pollution in the air than weekends. The same is true in the case of seasons, since a higher level of pollution is observed in winter. Table 1 presents some comparative statistical results of PM10 pollution (mean values and standard deviations) corresponding to different seasons and types of days in Warsaw within the years 2001–2014. This fact was taken into account in the model by introducing additional features represented by binary coded types of day and season of the year. As a result of such extension, the set of the potential features considered may be quite large and may contain more than fifty variables.

Table 1. Dependence of PM10 pollution on the season of the year and the type of day (in μg/m^3).

	Weekdays	Weekends and holidays
Spring	33.89±18.05	32.54±19.20
Summer	27.83±8.99	23.42±8.18
Autumn	36.39±17.67	31.03±14.44
Winter	41.97±33.72	37.82±24.81

On the basis of our experience in this field, various descriptors created in a different way are generated. The first subset is composed of environmental parameters forecast for the next day: the 24-hour average value of the temperature, wind speed and direction, humidity, pressure and insolation.

The next one is formed from the known past day parameters: the average, maximum and minimum values of temperature and pressure, the past (already known) average and maximum pollution corresponding to the previous day, the linear trend of hourly pollution, the linear prediction of the pollution for the forecast day made on the basis of this trend, the 2-element binary code of the season of the year (winter, spring, summer and autumn) and the binary code of the type of day (weekdays and weekends).

Additionally, taking into account the influence of the previous day pollution on the level of future pollution on the next day, selected hourly values of pollution of the previous day are also added.

As a result, the set of potential features containing many variables is created. Not all features are equally important in forecasting, hence detection of the most

influential factors is needed. This will be done in our work by using the genetic algorithm (Vafaie and De Jong, 1992; Bhanu and Lin, 2003) and the stepwise linear fit applying the backward and forward selection of variables (Guyon and Elisseeff, 2003; Matlab, 2014).

3. Feature selection using the GA

In feature selection by using the genetic algorithm (GA), the notion of the binary chromosome, representing the selected feature set, is used (Vafaie and De Jong, 1992). In this approach, the chromosome component value of one represents inclusion of the particular feature in the input variable set and the value of zero—deletion of the particular feature from the actual set. The GA consists of selecting parents for reproduction, performing crossover with the parents, and applying the operation of mutation to the bits representing children (Goldberg, 2013; Cloete and Zurada, 2000).

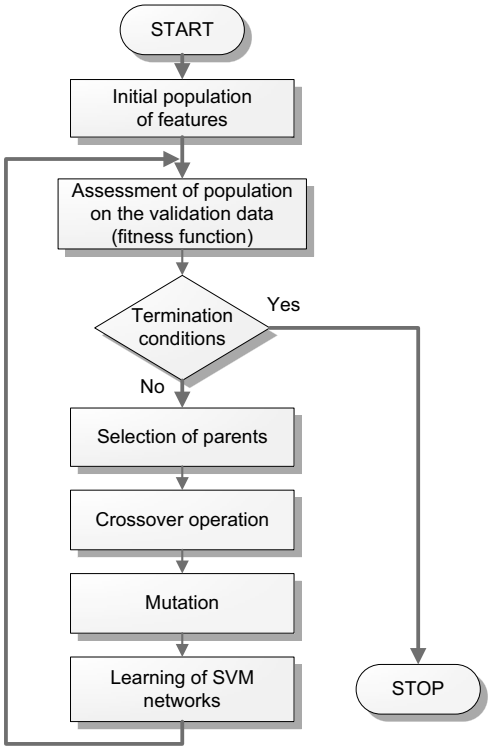

Fig. 1. Illustration of the genetic system of feature selection.

Each chromosome is associated with the input vector $\mathbf{x}$ of the components used as the explanatory variables to the predictor. Vector $\mathbf{x}$ is composed of only the features represented by the value one in the chromosome. The zero value of the chromosome component means the lack of such a feature in the input vector $\mathbf{x}$. The predictor is trained on the learning data set and then tested on the validation data. The testing error function defined for the validation data forms a basis for the definition of the fitness function. The fitness is defined here as an error function taken with a negative sign. The error function is the sum of squares of differences between the real and predicted values of pollution for the days taking part in validation. The genetic algorithm maximizes the value of the fitness function (equivalent to minimization of the error function) by performing the subsequent operations of selection of parents, the crossover among the parents and the mutation. Figure 1 presents the scheme of genetic operations used in our application for feature selection.

4. Stepwise fit of feature selection

Besides the genetic approach, a traditional linear method, generally known as the stepwise linear fit (Sprent and Smeeton, 2007; Matlab, 2014; Zhang, 2009), is also used. It is a method based on successive linear regression, in which the operation of adding and removing the candidate features used as the input attributes to the linear model of the prediction is performed. In general, two types of operations are made within this process:

- forward selection, usually starting with no variables in the model, testing the significance of addition of each variable, adding the variable which most improves the model, and repeating this process until none of the variables improves the designed model according to the assumed criterion;
- backward elimination, starting with some candidate set of variables, testing the deletion of each variable by using a chosen criterion of the model quality, deleting the variable which improves the model to the highest extent by being deleted, and repeating this process until no further improvement is possible.

In practice, both operations interlace each other. At each stage of the process, after a new variable is added, a test is made to check if some variables from the actual set can be deleted without increasing the error of regression. The procedure terminates when the measure of the model quality is (locally) maximized, or when the actual improvement is below some assumed tolerance value.

The impact of the actually investigated feature on the modeled process is measured through the value of its coefficient in linear regression and its statistical change in the process of adding and removing the next features. In each step of adding or removing the particular feature, the F-statistic, is determined, on the basis of which the decision of leaving or removing the given feature from the set is made.

Entering and removing the particular variable from the actual feature set is controlled by two parameters:

penter and *premove* (Sprent and Smeeton, 2007). The parameter *penter* specifies the maximum p-value for a variable to be recommended for adding to the model. The parameter *premove* specifies the minimum p-value for a variable to be removed from the set. The procedure is stopped when adding or removing any feature does not lead to increasing the accuracy of the linear model. Contrary to the genetic algorithm, application of stepwise fit provides only the local optimality of solution. However, in spite of that, this method has a reputation of being highly efficient in practical applications (Zhang, 2009).

5. Predicting systems

The features selected in the previous steps are used as the input attributes to the predicting systems. Predictors of very high efficiency were applied in the work. One of such solutions is the random forest of decision trees as well as neural networks, having the reputation of the best universal approximators: the MLP, RBF and SVM (Haykin, 2000; Siwek *et al*., 2010). They act in an independent way on the same data sets, and their results are fused into the final prognosis.

5.1. Individual predictors. The first system used for prediction is a random forest of decision trees developed by Breiman (2001). The RF is a typical ensemble learning method for classification and regression applying simultaneously many decision trees. The decision trees are trained on part of the available data and output either the class that is the mode of the classes in the classification problem or the mean prediction of the individual trees in regression task. The other part of data is used for out-of-bag testing the trained ensemble of decision trees.

Decision trees forming the RF are of multivariate form. They use a modified tree learning algorithm that selects a random subset of the available features (feature bagging). Thanks to this the correlation of the trees is reduced. If some features are very strong predictors for the target variable, these features will be selected in many of the trees, causing them to become correlated. A random choice of variables reduces the scale of this problem. Typically, for a dataset with N features, $\sqrt{N}$ features are used in each split. Moreover, each decision tree is trained on a different set of randomly chosen observations.

The MLP is a typical multilayer network structure applying sigmoidal neurons (Haykin, 2000; Cloete and Zurada, 2000). The information put to the input of the network is processed locally in each unit by computing the dot product between the corresponding input vector and the weighting vector of the neuron. Before training, the weights are initialized randomly. Training the network to produce a desired output vector when presented with an input vector involves systematically changing the weights of all neurons until the network produces the desired output within a given tolerance. The procedure is repeated over the entire training set. Learning is just reduced to the minimization of the Euclidean error measure over this set. The most effective learning approach applies gradient information and uses second order optimization algorithms, like Levenberg–Marquard or the conjugate gradient. The gradient vector in the multilayer network is computed using the backpropagation algorithm (Haykin, 2000).

Radial basis function networks are systems performing the role of local approximation (Haykin, 2000). The structure of the network is similar to the MLP, except the activation function, which is Gaussian. Its main advantage is great simplification of the learning algorithm following from association of the network parameters with the distribution of the learning data points in the multidimensional space. RBF networks implement a nonlinear transformation of the data from the input space to a high dimensional space. The superposition of the hidden neuron signals with proper weights helps to obtain an approximation of multidimensional data with a desired accuracy. Each output neuron of the RBF network performs a simple weighted summing operation,

$$y(\mathbf{x}) = w_0 + \sum_{j=1}^{K} w_j \phi_j(\mathbf{x}), \tag{1}$$

where the nonlinear activation functions $\phi_j(\mathbf{x})$ are Gaussian. The learning problem of the RBF network is split into two stages. The first one is the choice of the number of hidden units representing the Gaussian nonlinear functions and adaptation of parameters of these functions (centers and widths). These problems are solved through clusterization of the input data and association of cluster centers with the centers of the Gaussian function. In the second step, the synaptic weights of the linear output neurons are adapted by using singular value decomposition. There are also algorithms (for example, the orthogonal least square) which join both the stages in one common procedure (Haykin, 2000).

The support vector machine is another powerful neural alike structure developed by Vapnik (Scholkopf and Smola, 2002). Solution of the prediction problem needs its application in regression mode (SVR). The number of hidden units (kernel functions) is automatically determined in the learning procedure on the basis of training data. The learning of SVR is to minimize the weights of the network, while keeping the output signals as close as possible to their destination values within the predefined tolerance limit ε (Haykin, 2000; Scholkopf and Smola, 2002). The regularization constant C is applied for balancing between the values of weights and the prediction error on the learning data.

In practice, the learning procedure is composed of two stages, the so-called primary and dual tasks. The optimization problem is finally transformed to the dual problem of maximization of the quadratic function defined for a set of Lagrange multipliers. Its solution is relatively easy and there are many efficient algorithms leading to a global minimum. Details of the learning procedures of SVR can be found in many textbooks (e.g., Haykin, 2000; Scholkopf and Smola, 2002). To get reliable results of learning, the proper choice of hyperparameters: ε (assumed tolerance), the width σ of Gaussian functions and C (user specified regularization parameter) should be made. Their optimal values are usually determined in an introductory step of experiments by using a small percentage of learning data.

5.2. Final predicting systems. Individual solutions corresponding to the three neural networks (the MLP, RBF and SVM) are combined with feature sets, selected either by the genetic algorithm or stepwise fit. They are trained on the learning data and then tested on a separate testing set. The results of these predictors are merged together to produce the final forecasts of pollution on the next day. The weighted average and random forest applied as integrators are used in this step. In the weighted average approach the results of individual predictors are summed up with the weights proportional to the accuracy of the corresponding predictor on the learning data. The RF integrator treats the results of individual predictors as the input attributes and performs the prediction process on these data. This approach to prediction of environmental pollution is summarized in Fig. 2 (left).

Another approach investigated in the paper is direct application of selected features to the RF, performing the role of a predicting and integrating system at the same time. This solution is presented in Fig. 2 (right). The direct RF system will be supplied by the features selected by the GA, stepwise fit and a combined set of features chosen by both the methods.

5.3. Experimental setup. In this research both the predicting systems were investigated and compared. Their performance was checked on the observation data measured in Warsaw within the years 2001–2014. The feature selection stage was performed in the first stage of experiments on a separate set of data, not used further in the prediction process. Two selection methods applied results in two separate sets of features. On the basis of their contents the appropriate features were used in further experiments as the input attributes to the three neural predictors (the MLP, RBF and SVM) integrated into the final system. In training MLP, 12 hidden units were applied. The RBF network generated the best results at 300 Gaussian basis functions of the width equal to

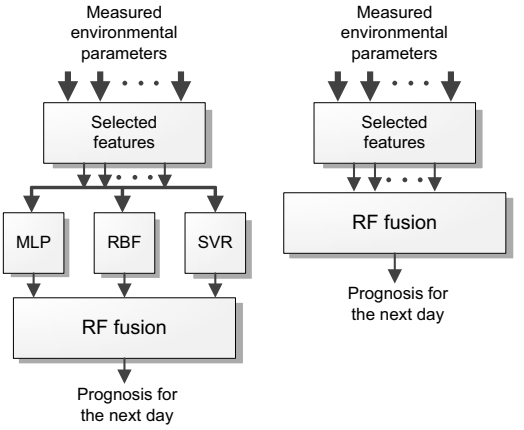

Fig. 2. General diagrams of the investigated predictors in PM10 prognosis: the ensemble system applying the intermediate neural predictors (left) and the direct application of the RF in the role of a predictor and an integrator (right).

1. The hyperparameters of the SVM were as follows: $C = 100$, $\varepsilon = 0.01$ and the unity width of the Gaussian function. All of them were selected in the introductory steps of experiments.

Additionally, the selected features were also applied directly to the RF predicting system. The RF was composed of 100 tress. Four variables in each node were used in splitting the data. Three variants were checked: (a) the features selected by the GA, (b) those by the stepwise fit and (c) their combined set.

The prediction experiments were performed 10 times using randomly selected learning and testing subsets. Then the average error for the testing data across all 10 trials was computed. The same testing data sets were used for all individual predictors.

The results of prediction are compared on the basis of their statistics. The following definitions of errors were applied;

- the mean absolute error (MAE),

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \mid y(i) - d(i) \mid; \tag{2}$$

- the mean absolute percentage error (MAPE),

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mid y(i) - d(i) \mid}{d(i)} \cdot 100\%; \tag{3}$$

- the maximum percentage error (MAX),

$$\text{MAX} = \max\left\{\frac{\mid y(i) - \hat{d}(i) \mid}{d(i)} \cdot 100\%\right\}; \tag{4}$$

- the root mean squared error (RMS),

$$\text{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \mid y(i) - d(i) \mid^2}. \qquad (5)$$

The variables y and d used in these definitions represent the results of prediction and the real values of daily mean pollution, respectively. Additionally, the Pearson correlation coefficient R between the real pollutions and their predictions made by our systems will also be given. Only the errors related to the testing data not taking part in learning in 10 repetitions of the calculations will be presented.

6. Results of numerical experiments

The numerical experiments were performed for different air pollutants, including PM10, NO2, SO2 and O3, all measured in two meteorological stations in Warsaw. The period considered extended from 2001 to 2014. The measurements were done every hour and contain the pollution level of each pollutant as well as the values of the basic meteorological parameters (temperature, speed and direction of wind, humidity and insolation). Around 20% of randomly chosen data from the database of the years 2004–2013 were used in the first stage of feature generation and selection. The remaining 80% were used only in the predicting experiments (learning and testing phases). The data of the year 2014 were left for only on-line prediction in the last step of the experiments.

6.1. Feature generation and selection. The first results of experiments are related to the PM10 pollution. A set of pairs ($\mathbf{x}$, d), $\mathbf{x}$ representing the potential input vector to the neural predictor and d the daily average of the PM10 pollution, is created. The vector $\mathbf{x}$ is composed of selected normalized features created from the meteorological data.

The following potential features were defined on the basis of the 24-hour average and also the minimum and maximum values of the meteorological parameters prognosed for the next day: $f1$, average temperature; $f2$, minimum temperature; $f3$, maximum temperature; $f4$, average humidity; $f5$, minimum humidity; $f6$, maximum humidity; $f7$, mean insolation; $f8$, average wind speed in x coordinate; $f9$, minimum wind speed in x coordinate; $f10$, maximum wind speed in x coordinate; $f11$, average wind speed in y coordinate; $f12$, minimum wind speed in y coordinate; $f13$, maximum wind speed in y coordinate. The forecast values were taken from the database of the Interdisciplinary Center for Mathematical and Computational Modeling (ICM) in Warsaw.

The next features were based on the historical data from the previous day. They include $f14$, average temperature; $f15$, maximum temperature; $f16$, minimum temperature; $f17$, the 24-hour average pollution; $f18$, maximum pollution; $f19$, minimum pollution; $f20$, the 24-hour average humidity; $f21$, maximum humidity; $f22$, minimum humidity.

The other features exploit the hourly linear trend of pollution observed on the previous day: $f23$, the value of the linear trend of change in the pollution estimated on the basis of succeeding hour measurements; $f24$, the predicted value of the average pollution of the next day following from this linear trend; $f25$, the linear trend of the change of temperature; $f26$, the linear trend of the change of humidity; $f27$, the predicted value of the average humidity for the next day following from this trend; $f28$, the linear trend of the change of the wind speed. The other 24 features (from $f29$ to $f52$) corresponded to 24 hourly values of pollution of the previous day.

The other two features ($f53$ and $f54$) represent the binary code of the season of the year (winter, 11; spring, 10; summer, 01; autumn, 00) and the last $f55$—the code of the type of the day (weekday: 1, weekend: 0). The normalization of data was made by dividing the real values of the particular features by their mean.

The randomly selected 20% of the data set was used in the feature selection process to discover the input variables having the highest impact on the prognosed average values of pollution on the next day. The stepwise fit and genetic algorithm were used in this stage. Additionally (for comparison purposes), the correlation of the single feature with the forecast value of the PM10 level was also checked. The stepwise fit method was applied at the values of *penter* $= 0.06$ and *premove* $= 0.08$. They were chosen as the result of introductory experiments by setting their different values and choosing the ones providing the best results of prediction on the learning data. The following parameters of genetic operations were used in the experiments: the mutation probability 2%, probability of crossover 0.8, the roulette rule in selection, population of chromosomes equal to 70, the initial random choice of zero or one for chromosome elements in all populations. This choice of parameters was preceded by some introductory experiments, aimed at getting the best fit of the predicted results to the real data.

The results of application of these two dedicated selection procedures were compared with simple checking of the correlation of the particular feature with the average value of pollution for the forecast day.

After selection of the most important features, the main experiments of prediction using the remaining 80% of data were performed. In these and all further experiments, one third of these data were left for testing only, while the remaining data were used for training the predictors.

In application of the neural predictors, the training data set was split into three separate parts, each one used to train only one of the predictors applied: the MLP, RBF or Gaussian kernel SVR. However, the testing set was common to all trained predictors. The experiments were repeated 10 times at a random choice of training and testing data.

6.2. Results of experiments for PM10 prediction. The first numerical experiments are presented for prediction of PM10, the pollutant which is extremely important because of its direct impact on human health via inhalation. Application of all features resulted in very high errors. The best results corresponded to RF application and the mean absolute percentage error (MAPE) was 29.82% in this case. The only way to reduce this error is to reduce the number of features by applying the proper selection procedure from full set of 55 features.

Table 2 presents the sets of features selected either by the GA or by the stepwise fit. Column 1 shows the notation of selected features. Only 24 (out of 55) features were selected by any of these two methods. The selected variables (presented in columns 2 and 4) are denoted by one. The zero value in any of these two columns means no selection of the feature by the particular selection method. In the case of stepwise fit the p-values of the t-test (Sprent and Smeeton, 2007) are also depicted. If the p-value of the particular feature was below the assumed threshold of 0.06, the feature was treated as significant (selected for prediction and denoted by 1 in the second column). As can be seen, the stepwise fit selected 19 features and the same population of features was created by the GA. However, only 14 features were commonly selected by both the methods. In the last column of the table the values of the correlation coefficient of the pollution level with the feature selected either by the genetic algorithm or the stepwise fit are shown.

In most cases the values of the correlation coefficients are not compatible with the selected set of features. This is a confirmation of the observation that a simple correlation principle is not a good choice for the feature selection in prediction problems.

The next task of experiments was predicting the mean value of PM10 for the next day with application of the features selected by the stepwise fit and by the genetic approach. The statistical results of prediction for each selection method and the predictors applied are given in Tables 3 and 4. All of them refer to the same data set not taking part in learning (approximately one third of the population of data taking part in the prediction experiments). The row denoted by RF refers to the direct application of the RF working simultaneously as the predictor and the integrator.

The next rows correspond to the application of individual neural predictors and their integration using

Table 2. Sets of features selected as the best by the stepwise fit and genetic approaches in the PM10 prediction problem.

Feature	Stepwise fit selection	p-value	Genetic selection	Correlation coefficient
$f17$	1	0	1	0.5998
$f18$	1	0.0166	0	0.5234
$f19$	1	0.0594	1	0.5524
$f1$	1	0.0232	1	-0.2197
$f2$	1	0	1	-0.2826
$f3$	1	0	1	-0.1628
$f4$	0	0.4828	1	-0.0227
$f5$	0	0.5181	1	0.0105
$f6$	1	0	0	-0.0789
$f27$	1	0	1	0.0032
$f24$	0	0.1460	1	-0.0216
$f7$	1	0.0145	1	-0.0812
$f9$	1	0	1	0.2938
$f10$	1	0.0030	1	-0.2928
$f11$	1	0.0193	0	0.0144
$f12$	1	0.0003	1	0.3073
$f13$	1	0.0001	1	-0.2919
$f29$	1	0	1	0.6696
$f32$	1	0.0002	0	0.5938
$f51$	0	0.9733	1	0.3615
$f52$	1	0.0001	0	0.3521
$f53$	1	0	1	0.0560
$f54$	0	0.1447	1	0.1521
$f55$	1	0.0060	1	0.1598

weighted averaging (w_avg) and application of the RF for fusion (RF fusion). In the weighted average case three individual results of neural networks create the input signal to the integrator. The weighted averaging denoted by w_avg is defined by

$$y(i) = \sum_{j=1}^{3} w_j z_j(i), \tag{6}$$

where $z_j(i)$ represents the pollution value for the i-th day predicted by the j-th predictor and w_j is the weight adjusted according to its relative accuracy with application of the bilinear formula (Osowski *et al.*, 2009). According to this formula the predictors of higher accuracy have higher impact on the final forecast.

It can be seen that in both feature selection methods the prediction accuracy is increased by an ensemble. However, the direct application of selected features to the RF integrator resulted in a better average accuracy (17.83% of the MAPE in comparison with the best result of 22.73% in the weighted average fusion). In this best case, the standard deviation of the MAPE results in 10 runs was 1.67%. Another observation is that the weighted average method of integrating three results of individual

Table 3. Quality measures of the RF, MLP, RBF, SVM and ensemble predictions of PM10 with application of the genetic algorithm.

	MAPE %	MAE μg/m^3	RMS μg/m^3	MAX %	R
RF	*17.92*	*5.405*	*8.36*	*134.17*	*0.924*
MLP	25.43	7.624	10.42	136.26	0.882
RBF	26.37	8.067	13.83	181.90	0.822
SVM	23.57	7.216	10.83	168.21	0.872
w_avg	22.68	6.979	10.35	150.06	0.885
RF fusion	23.36	7.269	11.39	116.80	0.858

Table 4. Quality measures of the RF, MLP, RBF, SVM and ensemble predictions of PM10 with application of the stepwise fit.

	MAPE %	MAE μg/m^3	RMS μg/m^3	MAX %	R
RF	*17.83*	*5.381*	*8.39*	*142.11*	*0.924*
MLP	27.11	8.240	12.78	143.12	0.822
RBF	26.64	8.526	16.81	204.14	0.783
SVM	23.66	7.109	12.82	179.31	0.816
w_avg	22.73	7.114	10.95	159.88	0.869
RF fusion	23.04	7.277	11.77	133.97	0.847

neural predictors was better than integration made by the RF.

Table 5. Quality measures of an ensemble of predictors of PM10 at fusing the genetic algorithm and stepwise fit selection results.

	MAPE %	MAE μg/m^3	RMS μg/m^3	MAX %	R
RF	*18.87*	*5.66*	*8.67*	*149.07*	*0.918*
w_avg	22.39	6.944	10.39	222.91	0.883
RF fusion	23.42	7.369	12.38	156.83	0.835

The features resulting from the stepwise fit and the genetic algorithm were merged in the last step. In the first approach they were input directly to the random forest, performing the role of predictors and integrators (the results denoted in the tables by RF). In the second approach the results of three neural predictors supplied by the features selected either by the genetic or by stepwise fit were combined together. In the latter case six series of prediction results taking part in an integration were used.

The results of such data processing are depicted in Table 5. The row denoted as RF presents the results of the direct application of the merged features to the RF network and the next rows—the results of weighted average and RF integration of six predictions made by neural networks.

As can be seen, the direct application of all selected features as an input to the RF also generated the most accurate results (smaller prediction errors and a higher value of correlation between the predicted and real values of pollution levels). However, the results are slightly worse than in the previous cases. Increasing the number of input signals to the predictor by combining the unique features chosen by the genetic algorithm and stepwise fit together did not improve the performance of the systems.

The obtained results are also compared to the naive method of prediction (Tan *et al.*, 2006). The numerical experiments showed absolute superiority of our approach. The improvement of results related to the MAPE was more than three times.

Figure 3 presents the estimated and real distribution of pollution for the tested days in a graphical way. They correspond to the best method checked in investigations. In most cases the results of an automatic predicting system are close to the real values, preserving well the trend for most days. The instantaneous prediction errors defined as the difference between the real and estimated values, shown in the lower subimage of Fig. 3, confirm limited values of mispredictions.

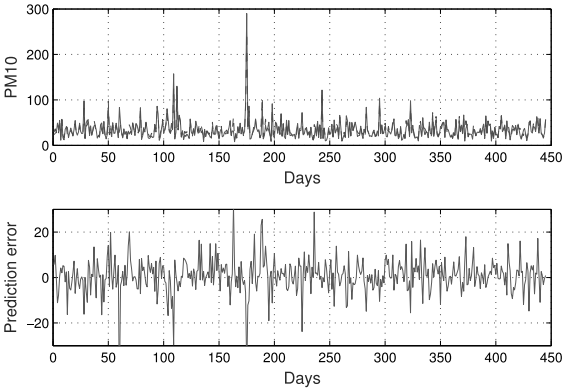

Fig. 3. Estimated and real distribution of PM10 pollution in μg/m^3 (top) and the prediction error (bottom) for the days taking part in testing.

Figure 4 presents the histogram of the testing errors for PM10. As can be seen, only a small number of samples were predicted with the errors above 50 μg/m^3. The curve resembles the normal distribution of the center located at zero.

The presented results of experiments were obtained in 10 runs using the data arranged in a random way. This was done to assess the prediction properties of the system in the most objective way. The next tests were made on a separate data set corresponding to the year 2014, arranged chronologically by using the previously trained predicting systems. The prediction of the daily mean pollution for the next day was done on-line, applying the data of the previous days and using the forecast atmospheric parameters. No repetition of experiments was applied in this case. The best results of the quality measures are

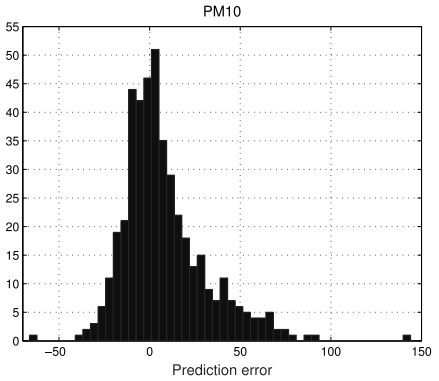

Fig. 4. Histogram of percentage prediction errors for the testing data of PM10 not taking part in learning.

presented in Table 6. The errors are on a similar level as in the best results in the previous experiments and fall inside the range of one standard deviation. This means that the developed system is able to work on-line in forecasting the pollution level from day to day with satisfactory accuracy.

Table 6. Quality measures of an ensemble of predictors of PM10 at fusing the genetic algorithm and stepwise fit selection results in on-line testing for the data of 2014.

	MAPE %	MAE μg/m^3	RMS μg/m^3	MAX %	R
RF	*19.96*	*6.04*	*9.17*	*138.42*	*0.933*
w_avg	23.93	7.44	11.13	172.81	0.885
RF fusion	23.64	7.59	11.81	147.20	0.872

It is interesting to compare the proposed method with the solutions reported by other authors. Unfortunately, there is no standard database available on Internet. Therefore, the comparison is made for different data sets. An objective interpretation of this comparison needs to take this into account. For example, Grivas (2006) reported the correlation coefficient between the predicted and real values of pollution, changing from 0.70 to 0.82 (depending on the year). Our best result is 0.92.

Most papers do not declare the values of the MAPE, the basic universal measure of prediction quality, which is independent of the level of pollution in the investigated region. Instead, they concentrate on the MAE. The relation of the MAE to the mean of PM10 reported by Grivas (2006) was changing from 0.213 to 0.255. In our case this ratio was 0.138.

6.3. Results of experiments for other air pollutants. Similar experiments were conducted for other pollutants: SO2, NO2 and O3. In the first phase of experiments the significant features were selected from the whole set of 55 elements. The genetic algorithm and the stepwise fit of feature selection were applied. Each algorithm selected a limited number of them.

In the case of SO2 the genetic algorithm chose 13 features ($f1$, $f2$, $f3$, $f6$, $f7$, $f8$, $f11$, $f12$, $f17$, $f18$, $f23$, $f27$, $f30$) while stepwise fit only 12 ($f1$, $f2$, $f3$, $f7$, $f11$, $f18$, $f19$, $f23$, $f27$, $f29$, $f41$, $f47$).

According to the genetic algorithm, the most important features for NO2 prediction included $f1$, $f2$, $f3$, $f7$, $f8$, $f9$, $f12$, $f16$, $f18$, $f19$, $f22$, $f23$, $f24$, $f27$ and $f50$ (15 features) and according to stepwise fit the set of $f1$, $f2$, $f6$, $f7$, $f8$, $f9$, $f12$, $f13$, $f16$, $f18$, $f19$, $f23$, $f24$, $f27$, $f28$ and $f50$ was selected (16 features).

In the case of O3 the optimal set of features according to the genetic algorithm contained $f2$, $f3$, $f4$, $f8$, $f9$, $f10$, $f11$, $f12$, $f14$, $f15$, $f16$, $f19$, $f23$, $f24$, $f27$, $f30$, $f35$ and $f42$ (18 features). Stepwise fit also selected 18 features ($f1$, $f2$, $f3$, $f7$, $f8$, $f9$, $f10$, $f11$, $f14$, $f15$, $f16$, $f17$, $f19$, $f23$, $f24$, $f27$, $f30$, $f34$). The number and contents of the selected features were slightly different for each pollutant and the selection method applied.

The selected feature sets were used as the input attributes to the predicting systems and then took part in the numerical experiments for predicting the daily average pollution level of each pollutant. Tables 7 and 8 show the quality measures (MAPE, MAE, RMS, MAX and R) characterizing the prediction quality for each pollutant with application of the genetic and stepwise fit algorithms. The results refer to the testing data not taking part in learning (approximately one third of the extracted population of data) and present the average of 10 runs of the prediction processes.

In all cases better results were obtained in the direct application of the selected features to the RF. Similarly to the PM10 case, the results of predictions related to all three neural predictors supplied by the feature sets selected by both the genetic and the stepwise fit algorithms (six individual solutions of predictors) were combined together. The results of such integration for these three pollutants with application of the weighted average and RF are depicted in Table 9.

The results show that increasing the number of input signals to predictors leads to a decrease in the prediction accuracy. This was observed for all pollutants. The histograms presenting the statistical distribution of prediction errors for each pollutant (Fig. 5) confirmed the good quality of prediction. Most errors are located close to zero. Only single cases were forecast with higher error values.

Our best results for all pollutants were compared with the standard ARX linear model implemented in Matlab (Matlab, 2014). The comparison was made on the same data sets and in the same organization of calculations as in our basic models. The optimal parameters of ARX were chosen after a series of introductory experiments:

Table 7. Quality measures of the RF, MLP, RBF, SVM and ensemble predictions of the SO2, NO2 and O3 levels with application of the genetic algorithm.

Pollutant	Predictor	MAPE %	MAE μg/m^3	RMS μg/m^3	MAX %	R
SO2	*RF*	*22.75*	*1.914*	*2.795*	*160.13*	*0.904*
	MLP	29.60	2.419	3.406	191.31	0.829
	RBF	32.31	2.748	4.194	218.82	0.741
	SVM	25.64	2.323	3.456	200.11	0.828
	w_avg	26.79	2.308	3.403	238.27	0.830
	RF fusion	27.64	2.364	3.377	243.14	0.832
NO2	*RF*	*18.40*	*3.98*	*6.117*	*148.84*	*0.928*
	MLP	29.95	6.439	9.115	247.12	0.839
	RBF	31.52	6.674	9.696	252.92	0.823
	SVM	25.72	5.916	8.911	250.91	0.851
	w_avg	26.78	5.852	8.583	312.11	0.861
	RF fusion	27.50	5.986	8.603	231.33	0.860
O3	*RF*	*20.19*	*8.648*	*11.91*	*197.33*	*0.846*
	MLP	27.53	10.311	13.41	255.71	0.770
	RBF	26.45	10.624	13.78	257.52	0.743
	SVM	22.64	9.278	12.34	257.33	0.797
	w_avg	23.81	9.446	12.18	211.71	0.801
	RF fusion	24.44	9.689	12.73	214.51	0.784

Table 8. Quality measures of the RF, MLP, RBF, SVM and ensemble predictions of the SO2, NO2 and O3 levels with application of the stepwise fit.

Pollutant	Predictor	MAPE %	MAE μg/m^3	RMS μg/m^3	MAX %	R
SO2	*RF*	*18.35*	*1.572*	*2.379*	*165.31*	*0.932*
	MLP	29.09	2.488	3.566	170.91	0.811
	RBF	31.19	2.744	4.447	183.52	0.731
	SVM	26.93	2.368	3.435	207.33	0.830
	w_avg	27.43	2.397	3.561	162.15	0.815
	RF fusion	28.01	2.385	3.445	179.25	0.825
NO2	*RF*	*18.27*	*3.944*	*6.059*	*161.51*	*0.931*
	MLP	30.65	6.542	9.352	250.13	0.833
	RBF	36.50	8.602	10.99	237.92	0.715
	SVM	27.10	6.157	9.291	223.17	0.836
	w_avg	29.10	6.555	10.98	213.52	0.777
	RF fusion	29.73	6.417	9.153	202.71	0.839
O3	*RF*	*17.31*	*7.501*	*10.59*	*152.25*	*0.882*
	MLP	24.36	9.536	12.60	265.85	0.778
	RBF	27.80	11.45	13.91	255.91	0.745
	SVM	25.20	10.23	13.94	222.22	0.757
	w_avg	23.91	9.579	13.56	204.64	0.749
	RF fusion	24.97	9.617	12.59	216.81	0.785

input-output delay equal to 0, orders of polynomials $A(z)$ and $B(z)$ equal to 8 and 3, respectively. Table 10 presents this comparison for all investigated pollutants. The advantage of our approach is evident. All quality measures of our best solution are superior to ARX.

Our research is also well compared with other results presented in different publications. For example, Agirre-Basurko *et al*. (2006) reported the average correlation coefficient R for the predicted NO2 on the level of 0.6, while the appropriate value in our experiments was 0.93. In the case of O3, their best average result was $R = 0.62$ and our result is 0.88. The reported

Table 9. Quality measures of an ensemble of predictors of SO2, NO2 and O3 after fusing the genetic algorithm and the stepwise fit selection results.

Pollutant	Predictor	MAPE %	MAE μg/m^3	RMS μg/m^3	MAX %	R
SO2	*RF*	*19.09*	*1.626*	*2.442*	*119.21*	*0.928*
	w_avg	26.38	2.273	3.351	186.11	0.835
	RF fusion	*27.64*	*2.364*	*3.377*	*198.62*	*0.832*
NO2	*RF*	*19.55*	*4.203*	*6.351*	*147.50*	*0.923*
	w_avg	27.62	6.121	9.264	199.10	0.837
	RF fusion	*27.50*	*5.986*	*8.603*	*173.23*	*0.860*
O3	*RF*	*17.70*	*7.657*	*10.73*	*162.53*	*0.879*
	w_avg	23.34	9.261	12.34	188.72	0.792
	RF fusion	*24.44*	*9.689*	*12.70*	*201.71*	*0.784*

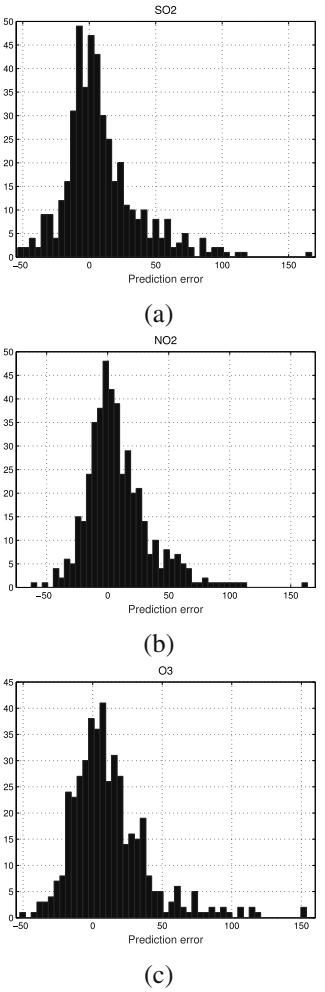

Fig. 5. Histograms of percentage prediction errors of the best predictors for testing data representing SO2 (a), NO2 (b), O3 (c).

Table 10. Comparative results of our best method (RF) with the linear ARX model for prediction of PM10, SO2, NO2 and O3.

Pollutant	Predictor	MAPE %	MAE μg/m^3	RMS μg/m^3	MAX %	R
PM10	*RF*	*17.83*	*5.381*	*8.396*	*142.11*	*0.924*
	ARX	29.42	9.221	14.07	267.11	0.757
SO2	*RF*	*18.35*	*1.572*	*2.379*	*165.31*	*0.932*
	ARX	31.64	2.743	16.05	285.25	0.800
NO2	*RF*	*18.27*	*3.944*	*6.059*	*161.51*	*0.931*
	ARX	31.69	7.061	10.27	241.23	0.780
O3	*RF*	*17.31*	*7.501*	*10.59*	*152.25*	*0.882*
	ARX	27.53	12.09	16.05	279.96	0.684

relative average error of prediction of NO2 in the paper of Perez and Trier (2001) (limited to only some days) was 35%, and our best result of the MAPE it was equal 18.27%. However, these comparisons might not be fully objective, since they refer to different regions of the world, where the mechanisms of pollution creation might present different degrees of difficulties.

7. Conclusions

The paper presented and compared different solutions of a system predicting the daily average air pollution of PM10, SO2, NO2 and O3 for the next day. An important point in this research is the generation and selection of the explanatory variables (prognostic features), which play the most significant role in the prediction process.

The genetic algorithm (the global optimization approach) and the stepwise fit (the local approach) were used for such selection. Their results were compared with the ordinary correlation of the single feature with the predicted variable. The features selected were used as the input attributes to the RF and an ensemble of neural networks working in regression mode.

Three different solutions of such networks, the MLP, RBF and SVM, were tried. Their choice was dictated by the need for independence of the forecast results, which is an important condition in their fusion for getting the improved final results of prediction.

The experiments showed that a simple application of the correlation of the feature with the predicted variable is not a good choice, since it resulted in an unacceptable level of the prediction errors. The proposed approaches (the genetic algorithm and the stepwise fit) led to a much better accuracy of prediction.

The application of several predictors and feature selection methods allowed integrating their results into one final forecast. The best results of integration were obtained in the direct application of selected features to the RF, performing at the same time the role of regression and integration. The numerical results presented in the paper have confirmed the superiority of such an approach for all pollutants considered.

The developed system is already under tests at the National Center for Nuclear Research (NCBJ) in Świerk, Poland, and is used to predict the next day PM10 pollution in Warsaw. The observed average accuracy of prediction made in this institution in the last year is on a similar level as the results presented in the paper.

References

Agirre-Basurko, E., Ibarra-Berastegi, G. and Madriaga, I. (2006). Regression and multilayer perceptron-based models for forecast hourly O3 and nO2 levels in the Bilbao area, *Environmental Modelling and Software* **21**(4): 430–446.

Bhanu, B. and Lin, Y. (2003). Genetic algorithm based feature selection for target detection in SAR images, *Image and Vision Computing* **21**(4): 591–608.

Breiman, L. (2001). Random forests, *Machine Learning* **45**(11): 5–32.

Brunelli, U., Piazza, V., Pignato, L. and Sorbello, F.and Vitabile, S. (2007). Two-day ahead prediction of daily maximum concentrations of SO2, O3, PM10, NO2, CO in urban area of Palermo, Italy, *Atmospheric Environment* **41**(14): 2967–2995.

Cloete, I. and Zurada, J. (2000). *Knowledge-based Neurocomputing*, MIT Press, Cambridge, MA.

Goldberg, D. (2013). *Genetic Algorithms in Search, Optimization, and Machine Learning*, Pearson Education, Upper Saddle River, NJ.

Grivas, G. and Chaloulakou, A. (2006). Artificial neural network models for predictions of PM10 hourly concentrations in greater area of Athens, *Atmospheric Environment* **40**(7): 1216–1229.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection, *Journal of Machine Learning Research* **3**(1): 1158–1182.

Haykin, S. (2000). *Neural Networks. A Comprehensive Foundation, 2nd Edition*, Prentice-Hall, Englewood Cliffs, NJ.

Matlab (2014). *Matlab User Manual—Statistics Toolbox*, MathWorks, Natic, MA.

Mesin, L., Taormina, R. and Pasero, E. (2010). A feature selection method for air quality forecasting, *Proceedings of the International Conference on Artificial Neural Networks, Thessaloniki, Greece*, pp. 489–494.

Osowski, S., Siwek, K. and Szupiluk, R. (2009). Ensemble neural network approach for accurate load forecasting in a power system, *International Journal of Applied Mathematics and Computer Science* **19**(2): 303–315, DOI: 10.2478/v10006-009-0026-2.

Perez, P. and Trier, A. (2001). Prediction of NO and NO2 concentrations near a street with heavy traffic in Santiago, Chile, *Atmospheric Environment* **35**(21): 1783–1789.

Scholkopf, B. and Smola, A. (2002). *Learning with Kernels*, MIT Press, Cambridge, MA.

Siwek, K., Osowski, S. and Sowiński, M. (2010). Neural predictor ensemble for accurate forecasting of PM10 pollution, *Proceedings of the International Joint Conference on Neural Networks, Barcelona, Spain*, pp. 1–7.

Siwek, K., Osowski, S. and Sowiński, M. (2011). Evolving the ensemble of predictors model for forecasting the daily average PM10, *International Journal of Environment and Pollution* **46**(3/4): 199–215.

Sprent, P. and Smeeton, N. (2007). *Applied Nonparametric Statistical Methods*, Chapman and Hall/CRC, Boca Raton, FL.

Sumi, S.M., Zaman, M.F. and Hirose, H. (2012). A rainfall forecasting method using machine learning models and its application to the Fukuoka city case, *International Journal of Applied Mathematics and Computer Science* **22**(4): 841–854, DOI: 10.2478/v10006-012-0062-1.

Tan, P.N., Steinbach, M. and Kumar, V. (2006). *Introduction to Data Mining*, Pearson Education, Boston, MA.

Vafaie, H. and De Jong, K. (1992). Genetic algorithms as a tool for feature selection in machine learning, *Proceedings of the 4th International Conference on Tools with Artificial Intelligence, Arlington, VA, USA*, pp. 1–6.

Zhang, T. (2009). Adaptive forward-backward greedy algorithm for sparse learning with linear models, *in* D. Koller *et al.* (Eds.), *NIPS: Proceedings of Neural Information Processing Systems*, MIT Press, Cambridge, MA, pp. 1921–1928.

Krzysztof Siwek was born in Poland in 1971. He received the M.Sc., Ph.D. and D.Sc. degrees from the Warsaw University of Technology, Poland, in 1995, 2001, 2013, respectively, all in electrical engineering. Currently he is a professor of electrical engineering at the Institute of the Theory of Electrical Engineering, Measurement and Information Systems, Warsaw University of Technology. His research and teaching interests are in computational intelligence, neural networks and their applications in prediction problems.

Stanisław Osowski was born in Poland in 1948. He received the M.Sc., Ph.D., and D.Sc. degrees from the Warsaw University of Technology, Poland, in 1972, 1975, and 1981, respectively, all in electrical engineering. Currently, he is a professor of electrical engineering at the Institute of the Theory of Electrical Engineering, Measurement and Information Systems, Warsaw University of Technology. His research and teaching interests are in the areas of neural networks, data mining and their application in different areas of biomedical engineering and techniques. He is an author or co-author of more than 200 scientific papers and ten books.

Received: 29 January 2015
Revised: 3 June 2015
Re-revised: 25 August 2015
Accepted: 20 September 2015