ALFA BK UNIVERSITY
FACULTY OF INFORMATION TECHNOLOGIES
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
International Scientific Conference

Alfa BK Univerzitet

UDK:351.777.6:004.8
DOI: 10.46793/ALFATECHproc25.078K
Original scientific paper

Optimization of AI Methods for Air Pollution Prediction

Goran KEKOVIĆ[1], Rade BOŽOVIĆ[1]

Abstract: One of the biggest problems of large urban areas is air pollution, and in this regard, artificial intelligence (AI) methods can predict the level of pollution using a wide range of parameters. The use of artificial neural networks (ANN) based on Levenberg-Marquardt algorithm with Bayesian regularization (LMBR) is considered in this paper. It is shown that this algorithm achieves very high prediction accuracy, competitive with radial basis neural networks, which are commonly used for regression tasks. It is also shown that by choosing the optimal sample size, in addition to tuned ANN parameters, a balance can be achieved between the desired accuracy of the method and the deviation between simulated and real data**.** Relative error was used as a measure of that deviation. At the same time, it has been shown that sample size is not always a decisive factor affecting the efficiency of the AI method itself, but that a complete picture can be obtained by taking into account the entire structure of the input data.

Keywords: Air pollution; Artificial neural networks; Bayesian regularization; Levenberg - Marquardt algorithm.

1 INTRODUCTION

In order to solve the problem of air pollution, modern technologies include the application of AI methods to predict pollution levels. There is a wide range of applied AI methods, which can generally be classified into five categories: Fuzzy Logic [1], Hidden Markov Models [2], Ensemble Models [3], Artificial Neural Networks [4] and Deep Learning [5]. From the point of view of our work, the category of ANNs is the most interesting and these methods are quite robust against all types of relations (monotonic, non-monotic, etc.) in input data. However, despite this, these and other AI methods are dependent on the structure of the input data set and in this sense, there exist just recommendations based on the results of individual studies. The situation was alleviated by pre-processing the data that includes input variable selection techniques (IVS). These techniques can be classified into three categories [6].

The first category consists of input variable filtering methods, which are mainly based on the application of statistical analysis methods [7]. Methods in this category are characterized by being independent of the type of ML method applied and being relatively simple to implement. An example of a method in this category is the mrMR algorithm (mr-minimum redundancy, MR-maximum relevance) [8]. More recently, two statistical tests have been applied to select input variables based on their relevance to the output variable.

The second category includes the so-called wrapper methods [9]. These methods include:exhaustive feature selection, sequential forward and backward selection and recursive feature elimination. The methods of exhaustive feature selection and recursive feature elimination are computationally demanding and they cover more of the parameter space than other methods. Sequential forward selection consists of sequentially adding variables to an initial empty set (SFS method), while in the core of sequential backward selection is eliminating variables from the complete initial set (SBS method). The selection of variables is carried out by following the values of the loss function or some other metrics of the effectiveness of the AI method. The SFS method is less computationally demanding than the SBS method, which is otherwise very impractical when there is a large number of input variables. These methods are characterized by the fact that for each specific set of input variables, it is necessary to determine the accuracy of a given AI method. This means that these methods are more demanding in terms of computing time and complexity compared to filtration methods.

Finally, methods within the third category, known as embedded methods, are the most complex for practical implementation [10]. The denominator of these methods is that the selection of input variables is done during the training of the AI method. This is the reason for their complexity and they are especially complex when applied to artificial neural networks (ANN), since their structure also depends on the number of input variables. Typical examples of these methods are the applications of the CART and ID3 algorithms in decision trees, where at each stage the set of attributes that has the best discrimination property between the categories of the target variable [11].

By applying IVS techniques, in principle, an optimal set of predictors can be obtained, which achieves the same or higher degree of efficiency as with full sets of predictors. This simultaneously saves computing time and memory resources. However, the optimal set of predictors does not have to be (which is most often the case) optimal for all types of AI methods. This means that these methods are still dependent on the structure of the input data.

Therefore, the term optimization has a much broader meaning in AI methods. This does not mean that every data set can be solved by tuning the parameters of an ANNs in its hyperspace. A comprehensive approach also requires taking into account the relationships between the input variables, in order to assess which method is most effective. Such a detailed approach cannot be the subject of a single paper, but it is still possible to examine the impact of individual parameters of the data set. Probably the most important and influential parameter is the sample size, and this is what this paper is dedicated to.

2 Methodology

For air pollution prediction, our choice was the LMBR algorithm, which is well described in the literature and there is no need to repeat it in detail here [12]. We will only list the basic settings of this algorithm. The essence of this algorithm is to minimize the loss function *F*:

$$F = \alpha E_W + \beta E_D \quad (1)$$

where E_w represents the sum of the squares of the neural weights $w_{ij}^{(k)}$(index *k* refers to a specific neural layer) across all neural layers, while E_d is the classical loss function (or random mean square error). After arbitrary initialization of the parameters *α, β,* neural weights $w_{ij}^{(k)}$ and biases parameters$b^{(k)}$, the input vectors in the form of matrix $X(n\ x\ m)$(*n*-number of samples, *m*-number of attributes) are propagated forward through the neural network. In the next step, the Hessian matrix of the second derivatives $H \cong 2\beta J^T J + 2\alpha I_k$ and the loss function *F* are determined and finally the corrections of the neural weights $W^{(k)}$ (in matrix form) are made:

$$W_{new}^{(k)} = W_{old}^{(k)} - \mu H^{-1} J e \quad (2)$$

In the above formula, μ is the Levenberg's damping factor, *J* is the Jacobian, and *e* represents the loss function summed over all samples. With the calculated neural weights according to Eq. (2) and E_w,E_D, the parameters *α, β* are calculated again which corresponds to minimizing the loss function *F*. The entire cycle of calculating the parameters $W^{(k)}$, *α, β,* E_w, $E_D$continues until the condition of convergence is met.

Since air pollution prediction with ANN is a regression task, it is necessary to set the maximum deviation between simulated and real data, or simply put, the tolerance. In this sense, we introduced the relative error r_i as a measure and it is given by the formula:

$$r_i = \frac{|y_i - y_{it}|}{y_{it}} \quad (3)$$

where y_i, y_{it} represent the simulated and real values of the target variable (*i=1,2,,,,n*). In addition, the value of $\Delta_{gr} = 0{,}04$ (or 4%) is the selected limit value of the relative error: $r_i \leq \Delta_{gr}$. To establish a relationship between the relative error and the random mean square error of the ANN, it can be very easily shown that the following formula holds:

$$E_D = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_{it})^2 \leq \left(M\Delta_{gr}\right)^2 \quad (4)$$

where is $M = \max(y_{it}), it = 1{,}2 \ldots N$, so that at $M = 1$ it turns out that$E_D \leq 10^{-4}$(in dimensionless form).

The structure of the ANN was as follows: the input layer had 12 neurons, the hidden layer had 10, and the output layer had one neuron. The optimization procedure is shown below.

Pseudo-algorithm of optimization procedure

INPUT:$\alpha, \beta, W, b, \Delta_i(\%)(i = 1{,}2{,}3{,}4)$; *X(n x m); structure of ANN; tuning ANN*; *accuracy=cell(1,4);*

for i=1:4
out=[];
for j=1:500:n
out=[out;ANN(1:j,Δ_i*)];*
end
accuracy{i}=out;
end

OUTPUT: *accuracy*

As can be seen from the above pseudo-algorithm, the air pollution prediction simulation is performed for several different values of the threshold $\Delta_i \in \{1\%, 2\%, 3\%4\%\}$. For each of these values, the sample size was continuously increased in steps of 500 and a simulation with ANN was performed on each of these samples and the accuracy was determined. In this way, four different curves of the dependence of the ANN accuracy on the sample size were determined, at a fixed threshold. As a result of the optimization, a curve was selected where the accuracy was >*95%* and where there no major oscillations in the accuracy with the change in the sample size.

Finally, a database of 9358 samples obtained by averaging the measured air pollution components per hour, was used [13]. The data were measured in a significantly polluted area in an Italian city between March 2004 and February 2005. The output or targeted variable was Benzene (C_6H_6) concentration (*mg/m*^{3}) and the input features or predictors were: hourly averaged concentration CO (*mg/m*^{3}), PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted), true hourly averaged overall NMHC (Nonmethane hydrocarbons) concentration (*mg/m*^{3}), PT08.S2 (titanium dioxide) hourly averaged sensor response (nominally NMHC targeted), true hourly averaged NOx concentration (*ppb*), PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted), true hourly averaged NO_2 concentration (*μg/m*^{3}), PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO_2 targeted), PT08.S5 (indium oxide) hourly averaged sensor response (nominally O_3 targeted), *T(*^{o}*C)* and relative humidity RH(%)and absolute air humidity AH.

In preparation for numerical simulations using the LMBR algorithm, the data were traditionally first manually filtered to remove incomplete data. Extreme points were removed by graphical display of the data and visual inspection. Then, the data were converted to the interval [0,1], according to the formula, which is a necessary and standard step in preparation for working with artificial neural networks.

2.1 Data preprocessing

In preparation for numerical simulations using the LMBR algorithm, the input data are filtered in a traditional way by removing incomplete data. Since the data are in the form of a matrix $X(n\ x\ m)$, if any row of data is incomplete, then the entire row is removed: $X(i,:) = [\ ]$. By visual inspection of the graphical representation of the input data, outliers are also removed. Then, the data were converted to the interval $[a, b]$according to the formula:

$$x'_{ij} = a + \frac{x_{ij} - x_{min}}{x_{max} - x_{min}}(b - a) \quad (5)$$

where is $a = 0, b = 1$ andx_{max} , x_{min} are the maximum and minimum values of the input data for $i = 1{,}2, \ldots n; j = 1{,}2 \ldots m$. This is necessary and standard step in preparation for

working with artificial neural networks. Beside that, the interval $[-1,1]$ is also often used. After numerical simulations with artificial neural networks, the data were then restored back to the original interval.

3 Results and Discussion

The results of the optimization procedure are shown in Figure 1 below and the *x,y*-axes indicate the sample size *N* and accuracy *ACC(%)* of the LMBR algorithm.

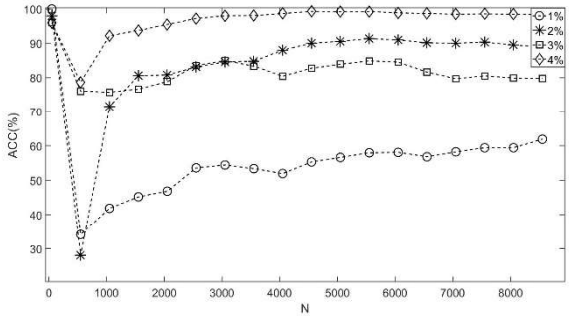

Figure 1 Accuracy *ACC(%)* vs. sample size *N*

In Fig.1 can be seen that the accuracy of the LMBR algorithm is highest at a threshold of $\Delta = 4\%$, for almost all sample size values. Two characteristic areas are clearly distinguished. The limiting value of the sample size between these areas is *N=3000.* In the first area, where *N<3000*, the value of the *ACC(%)* parameter does not exceed *95%* except in the case, $\Delta = 4\%$. On the other hand, at *N>3000*, the accuracy of the LMBR algorithm, at a relative error threshold of *4%* exceed *95%,* while at other values of relative error the accuracy does not exceed *90%*.

Also, in the above figure, an interesting property can be observed for sample size *N<1000*. Namely, contrary to our expectation, the accuracy first decreases with increasing sample size and a global minimum can be observed for all values of the relative error threshold. Then the values of the *ACC(%)* parameter start to increase, so that after *N>1000* we have a relatively stable increase in accuracy, for all values of the sample size, at all values of *Δ*. This unusual property of AI methods in that their accuracy decreases with sample size has also been observed by other authors. For example, researchers have investigated the application of machine learning methods to classify autistic and non-autistic individuals [14]. They have shown that even with a small sample size, high accuracy of machine learning (ML) methods can be achieved by combining various methods of their validation. By dividing the input data set into training and validation sets and using nested cross-validation, they achieved the best results in the sense that the efficiency of the ML methods was independent of the sample size.

Another interesting result can be seen in Fig. 1. The curve of the values of *ACC(%)* at a higher value of the relative error threshold is not always above the value at a lower threshold, for *N>3000*. Thus, the values of *ACC(%)* at $\Delta = 2\%$, are higher than the values at $\Delta = 3\%$, contrary to our expectation. It can also be seen from the figure that all accuracy curves above the value of sample size *N>3000* enter saturation. As previously pointed out, this sample size limit is also the limit above which the desired level of accuracy is achieved at a given relative error threshold. Note that above this value, there are also smaller accuracy oscillations, which are practically negligible for $\Delta \geq 4\%$.

Based on these results, it can be concluded that the optimal region for simulating real air pollution data with the LMBR algorithm is determined by the parameters: $N \geq 3000, \Delta \geq 4\%$. We confirmed these results by radial neural networks, where we achieved accuracy nearly of approximately 100% on the entire dataset. These ANNs has also been used in other research for air pollution prediction [15]. The key question that arises at this point is: can this result be generalized to other algorithms and other databases? There is no exact answer to this question, since there is no parameter space of the input data in which their properties could be exactly related to the properties of AI methods. There are only recommendations based on the results of other research.

All these results point to the need to define a parameter space of input data in which their properties could be more closely related to the properties of AI methods. There is not much research on this topic, and we would like to highlight one comprehensive study that examined the accuracy of AI methods depending on the sample size effect according to Cohen's scale and quality of input data [16].

4 CONCLUSION

In this paper, we have examined the possibility of predicting air pollution using the LMBR algorithm and shown that it can be a shortlist of candidates, in addition to radial neural networks. It should be emphasized that other AI methods can also be considered. This is because AI methods are still dependent on the structure of the input data, which, as we have seen, greatly affects the generalization property of ANNs. In addition, we have shown that the size of the selected relative error greatly affects the efficiency of the LMBR algorithm. The question that naturally arises here is whether the application of other ANN algorithms would lead to the same result. These and other questions such as: whether some other metric instead of the relative error would lead to better results and whether the parameter space of the input data could be defined, are left for future work.

REFERENCES

[1] Chao, B. & Guang, Q.H. (2024). Air pollution concentration fuzzy evaluation based on evidence theory and the K-nearest neighbor algorithm. *Frontiers of Environmental Science andEngineering 12*, **1-**47. https://doi.org/fenvs.2024.1243962.

[2] Liu, Y., Wen, L., Lin, Z. et al. (2024). Air quality historical correlation model based on time series. *Scientific Reports 14,* 22791. https://doi.org/10.1038/s41598-024-74246.

[3] Donnelly, A., Misstear, B., Broderick, B., (2015). Real time air quality forecasting using integrated parametric and non-parametric regression techniques. *Atmos. Environ. 103*, 53. https://doi.org/10.1016/j.atmosenv.2014.12.011.

[4] Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., Di Tommaso, S., Colangeli, C., Rosatelli, G., Di Carlo, P., (2017). Recursive neural network model for analysis and forecast of PM10 and PM2. 5. *Atmospheric Pollution Research8*(4), 652-659. https://doi.org/10.1016/j.apr.2016.12.014.

[5] Huang, C.J., Kuo, P.H., (2018). A deep CNN-LSTM model for particulate matter (PM2.5) forecasting in smart cities. *Sensors18* (7), 2220. https://doi.org/10.3390/s18072220.

[6] Cateni, S., Colla, V., Vannucci, M., (2023). Improving the Stability of the Variable Selection with Small Datasets in Classification and Regression Tasks. *Neural Processing Letters55*, 5331–5356. https://doi.org/10.1007/s11063-022-10916-4.

[7] Guyon, I., Elisseeff, A., & Kaelbling, L. P. (Ed.). (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*(7-8), 1157–1182. https://doi.org/10.1162/153244303322753616.

[8] Radovic, Ghalwash, M., Filipovic, N., Obradovic. Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics18*(9), 2-14. https://doi.10.1186/s12859-016-1423-9.

[9] Kohavi, R. & John, G.H. (1997). Wrappers for feature selection. *Artif Intell 97*(1–2), 273–324. https://doi.org/ 10.1016/s0004-3702(97)00043-x.

[10] Rodriguez-Galiano, V.F., Luque-Espinar, M., Chica-Olmo, J.A.M., Mendes. M.P. (2018). Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. *Science of The Total Environment 624*, 661-672. https://doi.org/10.1016/j.scitotenv.2017.12.152.

[11] Singh, S., Giri, M., (2014). Comparative Study Id3, Cart And C4.5 Decision Tree Algorithm: A Survey. International *Journal of Advanced Information Science and Technology (IJAIST) 3(*7), https://doi:10.15693/ijaist/2014.v3i7.47-52.

[12] Uwimana, E., Zhou, Y. & Sall, N.M. A short-term load demand forecasting: Levenberg–Marquardt (LM), Bayesian regularization (BR), and scaled conjugate gradient (SCG) optimization algorithm analysis. *J Supercomput 81*, 55 (2025). https://doi.org/10.1007/s11227-024-06513-y.

[13] Vito, S. (2008). Air Quality [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C59K5F.

[14] Vabalas, A, et al. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE14*(11). https://doi.org/10.1371/journal.pone.0224365.

[15] Yiran S., Ding L.Y., Yantao T., Yaowu S. (2015). Air:fuel ratio prediction and NMPC for SIengines with modified Volterra model and RBF network, Eng. *Appl. Artif. Int. 45*, 313-324. https://doi.org/10.1016/j.engappai.2015.07.008.

[16] Rajput, D., Wang, W.J., Chen, C.C. (2023). Evaluation of a decided sample size in machine learning applications. *Bioinforma*tics *24*, 1-17. https://doi.org/10.1186/s12859-023-05156-9.

Contact information:

Goran KEKOVIĆ, PhD, Assistant Professor
(Corresponding author)
Year of birth:1967
Institution: Faculty of Information Technology, Alfa BK University
Postal address:Marshal Tolbukhin Boulevard 8, 11070 Belgrade, Serbia
I-Mail: goran.kekovic@alfa.edu.rs
https://orcid.org/0000-0003-1429-0582

Rade BOŽOVIĆ, PhD, Assistant Professor
Year of birth:1983
Institution: Faculty of Information Technology, Alfa BK University
Postal address: Marshal Tolbukhin Boulevard 8, 11070 Belgrade, Serbia
E-Mail: rade.bozovic@alfa.edu.rs
https://orcid.org/0000-0002-8580-714X