Modeling of ambient O₃: a comparative study

Biljana Mileva-Boshkoska

Abstract—Air pollution is one of the biggest environmental concerns. Besides passive monitoring, the recent trend is shifting towards prediction of pollutants levels for near and far future. Since the pollution sources are constantly changing the need for adaptive modeling and prediction rises. This results in usage of variety of different machine learning algorithms for modeling. Recently, support vector machines (SVMs) and fuzzy identification are becoming very popular tools for modeling environmental systems. Both methods have short learning and execution time. The aim of this study is to examine the performances of the models built with these two methods.

The modeling techniques are applied to real data for summer and winter 2005.

I. INTRODUCTION

There are no limits to human curiosity and the need for mathematical models. Thus, when devising algebraic, differential, discrete, or any other models from first principles is not feasible, one seeks other avenues to obtain analytical models. Such models are devised by solving two cardinal problems in modern science and engineering [4]:

- Learning from experimental data (examples, samples, measurements, records, patterns or observations) by neural networks (NNS) and support vector machines (SVMs)
- Embedding existing structured human knowledge (experience, expertise, heuristics) into workable mathematics by fuzzy logic models (FLMs).

These problems seem to be very different, and in practice that may well be the case. However, after NN or SVMs modeling from experimental data is complete, and after the knowledge transfer into FLM is finished, these two models are mathematically very similar or even equivalent.

Science deals with modeling of ecological systems in order to explore or improve the environmental conditions in which we live. Air quality is one segment in the environment that is very important for the health of people and it is used by scientists for prediction of its pollution.

One approach for modeling of trends of ambient air concentrations is Support Vector Machines (SVM) [5], for modeling ozone concentrations [2], and for time series forecasting in the environmental applications [3], and is also used in this paper.

Another approach is using Fuzzy Logic (FL) for modeling ozone episodes [7], as well as for prediction of air pollution daily levels [12].

In this paper, we use the SVM and FLM for modeling of ambient air pollution with Ozone (O₃) and we compare the results that we receive from the two modeling approaches.

II. USED TECHNIQUES

A. T-S Fuzzy modeling

Among the different fuzzy models, the Takagi-Sugeno (T-S) fuzzy model [7] has attracted the most attention. The T-S fuzzy model proposed originally by Takagi and Sugeno is suitable for modeling the dynamics of complex nonlinear systems. T-S models assign simple (crisp) equations to the output variable. Most commonly these are either linear or quadratic dependencies on one or more input variables [1].

When modeling with T-S, the input x consists of scalar values:

\[ IF \ x_1 \ IS \ A_1 \ AND \ ... \ AND \ x_n \ IS \ A_n \ THEN \ y = f(x). \]  

(1)

When using the last expression, the output is no fuzzy set, but rather a singleton \( \{x, w\} \). The degree of membership \( w \) of this singleton is equal to the degree of fulfillment of the antecedent:

\[ w = \min \{\mu_{A_1}(x_1), ..., \mu_{A_n}(x_n)\}. \]  

(2)

If several such rules exist, the corresponding output singletons are combined using a fuzzy aggregation operator. Most often the product is used to compute AND operator and a weighted sum is used for aggregation (assuming \( r \) rules):

\[ y_{final} = \frac{\sum_{i=1}^{r} w_i \cdot y_i}{\sum_{i=1}^{r} x_i}. \]  

(3)

In this paper, we use T-S modeling to perform identification of MISO nonlinear dynamical system. Our identification method transforms the input-output process data to a fuzzy T-S model that we afterwards use for prediction of future hourly values of ozone concentrations.

B. Fuzzy c-means clustering algorithm

For identification of the MISO system, we use Fuzzy c-means (FCM) clustering technique. Clustering of numerical data forms the basis of many classification and system modeling algorithms. The purpose of clustering is to identify natural groupings of data from a large data set to produce a concise representation of a system’s behavior. Fuzzy c-means is a data clustering technique in which a dataset is grouped into \( n \) clusters with every data point in the dataset...
vector machine algorithms have been developed for numeric prediction that share many of the properties encountered in the classification case: they produce a model that can usually be expressed in terms of a few support vectors and can be applied to nonlinear problems using kernel functions. Similar with linear regression, the basic idea here is to find a function that approximates the training points well by minimizing the prediction error. The crucial difference is that all deviations up to a user-specified parameter $\xi$ are simply discarded. Also, when minimizing the error, the risk of overfitting is reduced by simultaneously trying to maximize the flatness of the function. Another difference is that what is minimized is normally the predictions’ absolute error instead of the squared error used in linear regression. A user-specified parameter $\xi$ defines a tube around the regression function in which errors are ignored. SVM approximate the learning data set with a function given in a form of:

$$f(x) = \sum_{i=1}^{l} w_i \phi_i(x) + b$$  \hspace{1cm} (4)

meaning that the original data $x \rightarrow \phi(x)$ are mapped into high dimensional space and then construct an optimal hyperplane in this space. $\phi(x)$ represent feature of the inputs, while $w_i$ and $b$ are coefficients. These are estimated by minimizing the risk function [10]:

$$R(f) = \int c(x, y, f(x,y))dp(x,y)$$  \hspace{1cm} (5)

where $c(x, y, f(x))$ is cost function that determines how to penalize estimation errors based on the empirical data $X$ [8]. Given that we do not know the probability measure $dp(x,y)$ we can only use $X$ for estimating a function $f$ that minimizes $R[f]$. A possible approximation consists in replacing the integration by the empirical estimate to get so called empirical risk function

$$R_{emp}[f] = \frac{1}{N} \sum_{i=1}^{N} c(x_i, y_i, f(x_i)).$$  \hspace{1cm} (6)

III. RESULTS FROM MODELING

Data that are used are gathered by the national automatic monitoring network (AMN) by the Ministry of Environment and Physical Planning in Republic of Macedonia. The data that are gathered include many zeros which was the main obstacle in selecting the data for the modeling. We decided to pick a small period of time where we have sufficient data and minimum zero data. The hourly data are taken from the monitoring station Karpoš, in Skopje, for the period 1 - 11 August, 2005 and 1-11 December 2005. The data from the first ten days are taken as training set (1200 data points). We tested the built models with the data taken for the 11th day of the month. Input parameters in the model are hourly concentrations of $O_3, NO_2$, temperature and humidity. The concentration of $O_3$ is predicted as output parameter:

$$O_3(t) = f(NO_2(t-1), NO_2(t), O_3(t-1),$$

$$\text{temperature}(t-1),$$

$$\text{humidity}(t-1))$$  \hspace{1cm} (7)
In (7), \( t - 1 \) means that data are one hour old. Actually we try to predict the concentration of ozon, if we know one hour old data for \( NO_2 \), \( O_3 \), humidity, temperature and the concentration of \( NO_2 \) at the time of measurement. The simulation results have 24 test points relating to the hourly measurements on the eleventh day of the month.

A. Results from Ozon modeling in August 2005 with FL

When modeling with FL, first we have to describe each of the input parameters with membership function. For that purposes we use the subtractive clustering algorithm in which we define the radius as 0.5. After the algorithm is executed, we receive the membership functions of all included parameters. The results show that each of the parameters used for August 2005, is described with seven membership functions which are given on Fig. 1 - 5. If we find the distribution of the input parameters we can see the similarity between the distribution of the parameter and the membership functions that describe it. Actually, the main advantage of FL is that in order to find the membership functions of parameters, we may use the expert knowledge instead of using difficult mathematical calculations. For example, data for \( NO_2 \) are distributed in the interval from 10-85 \( \mu g/m^3 \). However, most of the data are in the interval 10-30 \( \mu g/m^3 \) and 38-45 \( \mu g/m^3 \). If we analyze the membership function given on Fig. 1, we see that there are more membership functions in these two intervals and that the rest of the interval is covered with one membership function or with the intersection of several membership functions. If we compare the data distribution and the membership functions of the rest of the parameters, we will conclude that the intervals where there are more data, there are also more membership functions. In other words, in order to determine the membership functions of the input and output parameters, we may either use SCA or we may use our expert knowledge about the data.

B. Results from Ozone modeling in December 2005 with FL

In this case we also use the subtractive clustering algorithm for determining the number of clusters into the input parameters of the model. The radius in the SCA is also set to 0.5.

Each input parameter in this case is presented with three membership functions which are presented on Fig. 6 - 10. In this case data are more close to each other and they can be grouped into the three clusters. It is also possible to describe the input and output parameters with more than three membership functions, as it is the case previously. In order to do that, we would need to change the radius parameter of the SCA algorithm. As optimization of the FCA algorithm is not the purpose of this study, we continue with first choice for the radius. If we now find the distribution of the data, we may conclude that membership functions are distributed on intervals with most data.

C. Results from Ozone modeling with SVM

In order to perform function approximation from input-output data pairs with support vectors, first we have to determine the so called kernel function. The aim of the kernel function is to map the input output data into higher dimension space where it is possible to perform regression on the data. For the purposes of this study, we choose the
Gaussian function as kernel function:

$$K(x_i, x_j) = e^{-\frac{(x_i - x_j)^2}{2\sigma^2}}$$  \hspace{1cm} (8)

When we use the Gaussian function, first we have to determine the three free parameters which is the main difficulty when using the SVM. The three parameters that have to be determined are the penalty factor, the standard deviation, also known as speed parameter, and the width of the tube. The penalty factor tells how much to penalize those data that are predicted wrongly. The wide of the tube tells which errors from the prediction should be neglected and considered as zero. Theoretically, the value of the speed parameter \(\sigma\) influences a lot on the prediction performances of the model. Very small \((\sigma \to 0)\) or very large \((\sigma \to \infty)\) values of \(\sigma\) may lead to bad prediction results. If \((\sigma \to 0)\), all training data become support vectors. In that case, when using the SVM, the FL model will not be able to provide good prediction results. From the other side, if \((\sigma \to \infty)\), all training data will be considered as one point and the SVM model may produce same results for any new input data to the model. Therefore, these two extreme cases should be avoided.

In this study, we use the results from Gaussian modeling of ambient ozone with SVM with different kernels published in [6]. The value of the penalty factor is set to \(C=100\), the standard deviation is set to \(\sigma = 2\) and the width tube is \(\epsilon = 0.1\).

D. Comparison of the results

Results of comparison of the two models built with FL and with SVM with Gaussian kernel are given in Fig. 11 and 12 and the comparison of mean average errors is given in Table I. The results show that both algorithms may be used for prediction of ozone concentrations and they both give similar performances. In the first case, when modeling the ozone concentrations for August, the FL models perform better than SVM in terms of MAE. In the second case, when modeling the ozone concentrations for December, the SVM performs better than FL. If we look at the data in December, we may conclude that they are clustered within three different regions while the data in August are distributed more uniformly. That result in smaller number of membership functions for the input parameters in December compared to August.

At the end we measured the time for building the models. Time needed to build the models is approximately 0.2 seconds in both cases. That shows that both models may
be used for the same purpose of prediction of ambient air modeling with expectation of very close performance results.

IV. CONCLUSIONS AND FUTURE WORKS

A. Conclusions

The paper describes two methods that are used to predict the hourly concentrations of Ozone in the ambient air using SVM and Fuzzy logic at the municipality Karpoš III, in Skopje, Macedonia. Usually, Fuzzy logic is better to use than SVM when there is structured expert knowledge about the problem that is considered. FL is not a magic but a way of representation of a mathematical relationship with rules that we can actually understand and that we would normally describe it with complicated formulas. With FL, we can describe with membership function only those data intervals that actually matter considerably in the modeling and avoid the extreme and error data. On the other hand, Support vector machines give you an opportunity to approximate processes and complex relations among parameters that are usually unknown to the expert. The study shows that both methods are applicable for prediction of concentrations for ozone, and even more, the obtained results from the two models are very close to each other. In August, FL gives better results than SVM. In December, SVM perform better than FL. In this case, it is obvious that when we describe the input and output parameters with more membership functions, we receive better results than when we describe the input and output parameters with less membership functions. However, it should be noted that this is not the general case for modeling with fuzzy logic. Although it is not possible to use the exact same models to predict the concentrations on the other measurement places in the country, still the presented methodology is general and it may be used for building new models for the other measurement places. The new models will be trained with data measured at the local measurement sites.

<table>
<thead>
<tr>
<th>Month</th>
<th>SVM</th>
<th>FL</th>
</tr>
</thead>
<tbody>
<tr>
<td>August</td>
<td>8.6105</td>
<td>7.6167</td>
</tr>
<tr>
<td>December</td>
<td>1.0929</td>
<td>1.3177</td>
</tr>
</tbody>
</table>
Fig. 12. Comparison for August 2005

REFERENCES