BULETINUL Universităţii Petrol – Gaze din Ploieşti	Vol. LXII No. 1/2010	103 - 108	Seria Matematică - Informatică - Fizică

Air Quality Index Prediction using K-Nearest Neighbor Technique

Elia Georgiana Dragomir

Petroleum-Gas University of Ploiesti, Informatics Department, Ploieşti, Romania
email: elia.dragomir@yahoo.com

Abstract

One of the classical data mining techniques is k-nearest neighbor. This method uses the class of the k nearest neighbor to classify a new instance. The distance is calculated with one of the multiple mathematical distance metrics. In this paper, the technique is used in the air quality forecast domain in order to predict the value of the air quality index. This index is used to categorize the pollution level and to inform the population about some possible episodes of pollution.

Keywords: *air quality index, k nearest neighbor, Euclidean distance*

Introduction

One of the most influent factors on human health is air pollution. There are many different chemical substances that contribute to it. These chemicals come from a variety of sources. On one hand, there are natural sources such as forest fires, volcanic eruptions, wind erosion, pollen dispersal, evaporation of organic compounds and natural radioactivity. And on the other hand, the human industrial activity represents the artificial air pollution sources [14].

Among the many types of air pollutants are nitrogen oxides, sulphur oxides, carbon monoxides, ozone and organic compounds that can evaporate and enter the atmosphere. Large quantities of any air pollutant can affect the population health. In order to prevent them, there have been developed regional, national and international air pollution monitoring networks, which inform people about major pollutants concentrations in real time.

Moreover, an air quality index is calculated on different scales (1-10 in UE, 1-6 in Romania, 0-100 in USA). In Romania, the national network for air quality monitoring contains 117 fixed stations and 17 mobile ones. Beside the concentrations values that are recorded every hour, it can be calculated a quality index for every air pollutant. This index is determined based on the recorded concentrations. The general air quality index is the highest index of quality established for each pollutant [20].

Data mining techniques, as artificial neural networks, genetic algorithms, decision trees, k -nearest neighbor, logistic regression have been successfully used in air quality prediction problems [3][8][9][12]. Artificial neural networks are used to predict chaotic time series of air pollutant concentrations along with kNN in [3], to forecast 8 hours ahead the SO_2, CO, NO_2, NO and O_3 emissions in [8] and to predict the CO maximum ground level concentrations in

[12]. Nearest neighbor technique, self-organizing map, multi-layer perceptron and hybrid methods of the previous are used in [9] in order to handle missing values in air quality data sets.

In this paper, it is presented an application of k-nearest neighbor (kNN) technique, in order to predict the value of the index of air quality, using a database with values recorded for the sulphur dioxide, nitrogen oxides, carbon monoxide and ozone. Because the domain of air quality analysis sometimes includes little or no prior knowledge about the distribution of the data, for a classification study it may be used the k – nearest neighbor technique. In order to perform, this method requires only an integer k, a set of labeled examples (training data) and a metric to measure "closeness" between the instances. The training data is gathered by one of the monitoring station from the national air quality-monitoring network located in Ploiesti. The data used in this application were recorded in June 2009.

This paper contains an introductory part, in which an overview of the domain of the air quality analysis is briefly presented. The second section presents some related work on using kNN in general, and in air quality prediction, in especial. The third section gives a brief description of the k - Nearest Neighbor technique. The experiment is described in detail, along with its statistical results and their interpretation, in the fourth section. The last section concludes the paper and presents some future work ideas.

Related Work

Cover and Hart introduced the idea of nearest neighbor classification, in which the decision rule is to assign an unclassified sample point to the classification of the nearest of a collection of predetermined classified points [4]. An adaptive method of nearest neighbor classification (DANN) was presented by Hastie and Tibshirani [5]. This technique uses local discrimination information to estimate a subspace for global dimension reduction. There is demonstrated that this method can be generalized by applying specialized distance measures for different problems [5].

In the air pollution domain there can be mentioned the studies of Martın et al. in which two approaches have been used: artificial neural networks with backpropagation learning rule and k-nearest neighbors classifiers, in order to predict future peaks of carbon monoxide [12]. Another interesting study is made by Athanasiadis, Karatzas and Mitkas [1]. They have applied this method, along with other types of instance-based learners, rule-based classifiers, decision trees, Bayesian classifiers and neural networks in order to determine which method is more reliable for ozone forecasting in Athens area. The comparative analysis of the models performance showed that for the specific test case the k nearest neighbor algorithm have a considerably better performance compared to the statistical methods applied in this study [1].

Nearest Neighbor Technique

Nearest neighbor technique is one of the classification methods used in machine learning. It is based on the idea that a new object is classified based on attributes and training samples, using a majority of K-nearest neighbor category.

In order to apply this technique, it is necessary to have a training set and a test sample, to know the k value (how many neighbors are used in classification) and the mathematical formula of the distance calculated between the instances [18].

In general, this distance is determined with Minkowski distance [18] given by formula 1 and some particularization of it, like Euclidean distance when p=2:

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}, \quad (1)$$

where x_i represents the test sample, y_i is the training data, n is the number of features.

The k nearest neighbor classifier is commonly based on the Euclidean distance (formula 2) between a test sample and the specified training samples.

$$\sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \quad (2)$$

It is based on minimum distance from the test instance to the training samples to determine the k nearest neighbors. After k nearest neighbors are selected, the majority of these k nearest neighbors decides the prediction of the new instance [19].

The general algorithm of computing the k-nearest neighbors is as follows:

- Establish the parameter k that is the number of nearest neighbors;
- Calculate the Euclidian distance between the query-instance and all the training samples;
- Sort the distances for all the training samples and determine the nearest neighbor based on the k-th minimum distance;
- Use the majority of nearest neighbors as the prediction value.

Experiment Description

Our experiment uses the k nearest neighbor method in order to predict the value of the air quality index, based on a training data recorded in June 2009. The parameters selected to create the data sets are: sulphur dioxide [µg/m^3], nitrogen monoxide [µg/m^3], nitrogen dioxide [µg/m^3], carbon monoxide [mg/m^3] and ozone [µg/m^3]. The daily mean values of these parameters are used in order to establish the quality index for each pollutant; the different ranges for these values are represented in Table 1. The general air quality index takes the higher value of the pollutants indexes.

Table 1 Mean concentration domains for each quality index

SO_2 (ug/m3)	**NO_2, NO (ug/m3)**	**O_3 (ug/m3)**	**CO (mg/m3)**	**Quality Index**
0 – 49, (9)	0 – 49,(9)	0 – 39,(9)	0 - 2,(9)	1(excellent)
50 – 74,(9)	50 – 99,(9)	40 – 79,(9)	3 - 4,(9)	2(very good)
75 – 124,(9)	100 – 139,(9)	80 – 119,(9)	5 - 6,(9)	3(good)
125 – 349,(9)	140 – 199,(9)	120 – 179,(9)	7 – 9,(9)	4(medium)
350 – 499,(9)	200 – 399,(9)	180 – 239,(9)	10 – 14,(9)	5(bad)
>500	>400	>240	>15	6(very bad)

Only the data recorded in 29 days of June 2009 are available for this experiment. This is a drawback because the training set should have had more instances in order to create a model more accurate and precise.

k-NN-based Prediction and Statistical Results

The experiment has been made using Weka (Waikato Environment for Knowledge Analysis), a data mining specialized software. Weka is a popular suite of machine learning written in Java, developed at the University of Waikato. WEKA is a free software available under the GNU General Public License. It contains a collection of visualization tools and algorithms for data

analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality [21].

The Weka algorithm's output includes some general information about the used scheme, the number of instances (29) and the attributes (6), as well as the attributes names and the *test mode.* For this technique it was used the *10-fold cross-validation*. During this type of training, a set of input samples is split up into 10 equal size partitions. Nine of the partitions are used for training and one for testing, until each partition is used for testing.

The predictions on the test data are given in Figure 1. For each of the instances in the test set, the following is displayed: the instance number, which is followed by the actual classification of the air quality index, then the predicted classification [22]. The last column represents the error classification: if the actual and the predicted values are equal, then the error is zero. Otherwise, it is displayed the error value: as a negative number if the predicted value of the air quality index is smaller than the actual one, or as a positive number if the prediction gives a value greater than the actual value.

=== Predictions on test data ===

inst#,	actual,	predicted,	error	inst#,	actual,	predicted,	error
1	4	2	-2	1	2	3	1
2	3	3	0	2	3	3	0
3	4	3	-1	3	3	3	0
1	4	4	0	1	4	3	-1
2	3	2	-1	2	4	4	0
3	2	2	0	3	2	2	0
1	4	3	-1	1	3	3	0
2	2	2	0	2	2	2	0
3	3	3	0	3	3	2	-1
1	2	3	1	1	3	3	0
2	2	2	0	2	2	3	1
3	3	4	1	3	3	3	0
1	3	3	0	1	3	3	0
2	4	4	0	2	4	4	0
3	3	3	0				

Fig. 1. The prediction on test data output

There are 19 of 29 instances with prediction error equal with zero, which represents a percent of 65.51% right predictions for this model. If a new data measurement is available, there it is a probability of 65.51% for this model to predict accurately the general air quality index. This result is relatively good taking into account that only 29 instances were used for the training data, as well as the test set. Better results may be obtained if there will be used as input sets data that have been recorded for a longer period of time.

Weka provides also complementary information about some important statistical parameters. One of these parameters is the *correlation coefficient*, which is a measure of the agreement between two individuals, with a 0.5614 value in our case. It is considered that a weak correlation has a coefficient closer to zero and a strong correlation between the attributes is reflected in a coefficient value over 0.50. The correlation coefficient from this experiment shows a strong predictive relationship among these attributes.

=== Summary ===	
Correlation coefficient	0.5614
Mean absolute error	0.3793
Root mean squared error	0.6695
Relative absolute error	64.5706 %
Root relative squared error	87.1035 %
Total Number of Instances	29

Fig. 2. The statistical output

Other parameters are *mean absolute error*, which is a quantity that is used to measure how close forecasts or predictions are to the eventual outcomes, *root mean squared error*, which constitutes a good measure of the model's accuracy, *root relative squared error* (the average of the actual values), and *relative absolute error* that is similar to the relative squared error. This summary for our experiment is presented in Figure 2.

Conclusions

Air pollution is an important issue nowadays, being a factor which influences both human health and activities. In order to avoid the consequences of pollution episodes, the authorities have developed air quality monitoring networks that inform people in real time about the air quality index.

Using different techniques, such as statistical methods, artificial intelligence or data mining techniques, collections of recorded data stored in data warehouses can be used to forecast future values of the most important air pollution parameters, as well as of the air quality index. One of these techniques is k nearest neighbor, a method that relies on the distance between the closest k neighbors of one new instance to classify it. It can be used also to predict the value of a certain parameter.

In this paper, we have presented an experiment done in order to determine the particularities of applying this technique for air quality analysis. Aiming at generating a prediction for the air quality index, training data that were collected in June 2009 were used as input data for the algorithm. The experimental results show that among the parameters that have been selected for this experiment, there is a strong correlation, and, therefore, these can be used in the forecasting process.

The results were relatively good, if we consider that for 19 of the 29 instances the prediction error was zero. The accuracy of the model can be improved by taking into consideration a longer period of time for the model's training set.

As future work, seeking to obtain more precise predictions, versions of this algorithm can include different values for parameter k and other measurement data recorded for a longer period of time. Other research direction refers to the implementation of other prediction methods, in the same field, followed by a comparison of the results.

References

1. Athanasiadis, I.N., Karatzas, K., Mitkas, P. - Classification techniques for air quality forecasting*, BESAI 2006 Workshop on Binding Environmental Sciences and Artificial Intelligence*, part of the 17th European Conference on Artificial Intelligence, 2006
2. Brunelli, U. et al. - Three hours ahead prevision of SO2 pollutant concentration using an Elman neural based forecaster*, Building and Environment*, Volume 43, Issue 3, March 2008, pp. 304-314
3. Gautam, A.K., Chelani, A.B., Jain, V.K., Devotta, S. - A new scheme to predict chaotic time series of air pollutant concentrations using artificial neural network and nearest neighbor searching, Atmospheric Environment, volume 42, Issue 18, June 2008, pg 4409 - 4417
4. Hart, P.E., Cover, T.M. - Nearest neighbor pattern classification*, IEEE Transactions on Information Theory*, IT-13, 1967
5. Hastie, T., Tibshirani, R. - Discriminant adaptive nearest neighbor classification, *IEEE Trans. Pattern Anal. Mach. Intell*. 18(6), 607–616 (1996)
6. Haupt, S.E., Pasini A., Marzban C. - *Artificial Intelligence Methods in the Environmental Sciences,* Springer, 2009
7. Hopgood, A. - *Intelligent Systems for Engineers and Scientists 2d edition*, CRC Press, 2001

8. Ibarra-Berastegi, G., Elias, A., Barona, A., Saenz, J., Ezcurra, A., Diaz de Argona J. - From diagnosis to prognosis for forecasting air pollution using neural networks: Air pollution monitoring in Bilbao, *Environmental Modelling & Software*, vol. 23, number 5, May 2008, pp. 622-637
9. Junninen, H. et al. - Methods for imputation of missing values in air quality data sets, *Atmospheric Environment*, Volume 38, Issue 18, June 2004, pp. 2895-2907
10. Karatzas, K., Kaltsatos, S. - *Air pollution modelling with the aid of computational intelligence methods in Thessaloniki, Greece, Simulation Modelling Practice and Theory*, Volume 15, Issue 10, November 2007, pp. 1310-1319
11. Kurt, A., Gulbagci, B., Karaca, F., Alagha, O. - An *online air pollution forecasting system using neural networks, Environment International*, vol 34, Issue 5, July 2008, pp. 592-598
12. Martin, M.L. et al. - Prediction of CO maximum ground level concentrations in the Bay of Algeciras, Spain using artificial neural networks, *Chemosphere*, volume 70, Issue 7, January 2008, pp. 1190 - 1195
13. Oprea, M., Nichita, C., Dunea, D. - *Aplicatii ale inteligentei artificiale in protectia mediului*, Editura Universitatii Petrol – Gaze din Ploiesti, 2008
14. Petre, M.- *Tehnologii necatalitice pentru depoluarea atmosferei*, Editura Universităţii Petrol-Gaze din Ploieşti, 2007
15. Petre, E.G.- A Decision Tree for Weather Prediction, *Bulletin of PG University of Ploieşti, Series Mathematics, Informatics, Physics,* vol. LXI, nr. 1/2009, pp. 53-58
16. Song, Y., Huang, J., Zhou, D., Zha, H., Giles, C.L. - IKNN: Informative K-Nearest Neighbor Pattern Classification, *Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases*, Warsaw, Poland, 2007, pp. 248 - 264
17. Weinberger, K.Q., Blitzer, J., Saul, L.K. - *Distance metric learning for large margin nearest neighbor classification,* NIPS, 2005
18. Leon, F. - *Inteligenta Artificiala*, http://eureka.cs.tuiasi.ro/~fleon/Lab_AIA/knn.pdf, accessed on 15 March 2010
19. *** - *K Nearest Neighbor Technique*, http://people.revoledu.com/kardi/tutorial/KNN/HowTo_KNN.html , accessed on 15 February 2010
20. *** - *Reţeaua Naţională de Monitorizare a Calităţii Aerului*, http://www.calitateaer.ro/index.php, accesssed on 15 January 2009
21. *** - *Weka Software Documentation*, http://www.cs.waikato.ac.nz/ml/weka/, accessed on 20 March 2010
22. *** - *Weka Documentation*, http://wekadocs.com/node/13, accessed on 21 March 2010

Predicţia indexului de calitate a aerului folosind tehnica celor mai apropiaţi *k* vecini

Rezumat

Una dintre tehnicile clasice de data mining este tehnica celor mai apropiaţi k vecini. Această metodă foloseşte tipul clasei celor mai apropiaţi k vecini pentru a clasifica o nouă instanţă. Distanţa dintre elemente este calculată folosind una dintre multiplele metrici matematice pentru distanţe. În această lucrare, această tehnică este aplicată în domeniul predicţiei calităţii aerului cu scopul de a prezice valoarea indicelui de calitate a aerului. Acest index este util în clasificarea nivelului de poluare şi în informarea populaţiei despre eventualele episoade de poluare.