Trend Analysis and Prediction of Air and Water Pollutants using Regression algorithm SMOreg

PROF. SWATI VITKAR

[1]Research scholar, JJT University, Jhunjhunu, Rajasthan
Lecturer, S.I.E.S College, Nerul, Navi Mumbai,
Maharashtra

swativitkar@gmail.com

Abstract : Air and water pollution has become very common problem and it can be one of the factors affecting human health. Because of the bad performance of the traditional modelling methods in prediction, in the present paper, the author adopt SMOreg to study the pattern of data, and also its future trend. The author proposes a novel method called SMOreg. The result shows that the new method improves the accuracy of prediction. This paper proposes a scalable, flexible, economic model which will be useful to identify the air and water pollution pattern and also show its future trend and it will also help to visualize environment data and extract meaningful patterns from the environment data so that it will allow the policy makers to take the remedial measures. Java based environment data mining systems enable Non Government Organizations or policy makers to fully exploit their environment analytical capacity when and where it is needed.

***Keywords* — Environment Data Mining, prediction, trend.**

I. INTRODUCTION

Environmental data mining has gained popularity due to the increase in volume of environmental data. Finding patterns and relationships in raw environment data enables the air and water pollution department to characterize and study the current trend of the data as well as predict the future trend of environment data. The primary goal of environment data mining, is to identify environment trends and patterns/series. Mining of environment data provides timely and pertinent information about environment patterns. It will also provide trend analysis to assist Municipal Corporations, which would help them in planning and deployment of resources for the prevention and suppression of pollution activities. Combining historical data with current data sometimes would aid to unearth new clues, thus helping in solving many interesting facts. In environment data analysis, statistical examinations are performed on the air and water data of five zones in order to study its current effluent patterns and to develop the future trend.
Through this pattern of pollution depends on location and time, gradual patterns and trends will emerge which will lead to preventive solutions. The objective of environment data mining is evaluating the probability of a environment and assessing risks. This involves the analysis of data pertaining to observed behaviour and modelling it in order to determine the likelihood of its occurrence again.

Data mining has proven to be a useful methodology in a number of fields. However, its use in environment applications and research is still in its infancy. It can be used in environmental agencies to discover new patterns or trends. It can also be used as an aide in environment prevention. Data mining operations are computationally intensive. The amount of data to mine is usually so large that even the simplest mining operation might require a very long, or sometimes an even prohibitive, amount of time. Efficient storage of data is thus a necessity. In environment analysis, locations play an important role; during the study questions such as how far does the residential area from where the pollution data is collected? which areas have the highest number of water bodies? and many other similar questions are always of interest. For that reason, we decided that a spatial database would best data storage engine for this project.
This project is an attempt to incorporate various data mining techniques like Prediction or forecasting of air and water data, trend analysis.

II. RELATED WORK

[3] This article points out an important source of inefficiency in Platt's sequential minimal optimization (SMO) algorithm that is caused by the use of a single threshold value. Using clues from the KKT conditions for the dual problem, two threshold parameters are employed to derive modifications of SMO. These modified algorithms perform significantly faster than the original SMO on all benchmark data sets tried.
[1] Shevade et al.[1] are successful in extending some improved ideas to Smola and Scholkopf's SMO algorithm[2] for solving regression problems, simply named SMOreg. In this paper, we use SMOreg in exactly the same way as linear regression(LR) is used in locally weighted linear regression[5](LWLR): a local SMOreg is fit to a subset of the training instances that is in the neighborhood of the test instance whose target function

value is to be predicted. The training instances in this neighborhood are weighted, with less weight being assigned to instances that are further from the test instance. A regression prediction is then obtained from SMOreg taking the attribute values of the test instance as input. We called our improved algorithm locally weighted SMOreg, simply LWSMOreg. We conduct extensive empirical comparison for the related algorithms in two groups in terms of relative mean absolute error, using the whole 36 regression data sets obtained from various sources and recommended by Weka[3]. In the first group, we compare SMOreg[1] with NB[4](naive Bayes), KNNDW[5](k-nearest-neighbour with distance weighting), and LR. In the second group, we compare LWSMOreg with SMOreg, LR, and LWLR. Our experimental results show that SMOreg performs well in regression and LWSMOreg significantly outperforms all the other algorithms used to compare.

6. Case Study

The present study was confined to NaviMumbai(19^0 2' 2.20" N, 73^0 0' 43.71"E). This city lies across the Thane creek, north east of Mumbai and flanked by the Thane creek waters on its west, south-west and north-west contours.

Navi Mumbai was developed as a planed city by City and Development Corporation of Maharashtra (CIDCO). There are 6 nodes and approximately 30-50 sectors in each node.

In several nodes the city has fresh water lakes (ponds). The Navi Mumbai Municipal Corporation (NMMC) and CIDCO have also impounded creek water creating a series of water holding ponds.

The present case study was confined to 7 different zones of Navi Mumbai which are: Nerul, Vashi, Belapur, Airoli, Ghansoli, Turbhe, Koparkhairne. Each zone with more than two water bodies and air stations. After allocating the air and water pollutants data from these different zones at uniform interval of time, we applied data mining techniques for fining the missing data values and studied the trend of the data also predicted its future trend. By storing the rules in another database we could able to find out the diseases caused by these air and water pollutants from the same area. Also we are able to find out the upcoming number of health victims due to these reasons.

B. Database Design

The Environment data and visualization system was built using the following software tools. All packages are free or Open Source software. Java 6 is a powerful object oriented language. The purpose of this course project is to develop a java application which is capable of searching and visualizing environment report data. Datasets that are used for this project are the environment reports of NMMC spanning last 06 years.

The first step of this project involved creating and storing the database using PostgreSQL. The next major step involved applying the mining algorithms on the air and water parameters data to extract meaningful patterns from the data. The final step is creating a GIS based front end (visualization) to interact with data stored at the back end to represent the data. The Google Maps API offers a 2D mapping interface with a robust overlay capability. PostgreSQl database with support for geometry and geospatial query capability used in conjunction with PostGIS. WEKA is a data mining tool with a collection of machine learning algorithms. We chose PostgreSQL 9 with the PostGIS spatial extension. PostgreSQL is a high performance open source DBMS with a rich set of features, good scalability, and support for very large tables (32 TB). Being open source PostgreSQL receives many contributions from academic institutions, making it one of the most up-to-date DBMS's. PostGIS spatially enables PostgreSQL, adding support for geographic data types. It extends the SQL commands with a set of functions and operators that trivialize certain operations which would have otherwise required multiple queries or were simply undoable in SQL, such as calculating the distance between two objects, testing for intersection, calculating the bounding box of a set of points, and many others.

Java Interface : By using Java interface as front end various graph on the studied parameters of air and water are generated using JCharts.

C. Data Mining

Data mining is the part of the project where the environment reports stored in the database are processed using different data mining algorithms such as Classification, Association, Clustering and Outlier Detection. The purpose of data mining is to find associations and other relations between various parameters of air and water, location, time, etc. To accomplish these tasks; I have used open source machine learning and data mining software tool such as, WEKA. In the project, WEKA was employed to perform the data mining tasks as it implements a wide range of data mining algorithms and also has visualization support even though other software tools do perform the same tasks.

D. About WEKA?

WEKA is an Machine learning and data mining software tool written in Java, few of the main features that are included in WEKA are as follows:

- Data pre-processing tools, learning & evaluation methods
- Graphical user interface included for data visualization

The dataset format used by WEKA[7] is referred to as "Attribute-Relation File Format (ARFF)," this format is intuitive and easy to build.

To implement the data mining tasks algorithms involved in WEKA, the data set should be represented in such a format. This can be done by getting data from the database directly through functions provided by WEKA, which automatically converts the data into the "ARFF" format. Another way to import the dataset into WEKA is through "Comma Separated File Format (.csv)." This is facilitated in the project by implementing a converter (JAVA Program) that converts CSV files through ARFF format.

E. Implementations in WEKA

weka::classifiers::functions::SMOreg Class Reference

Inheritance diagram for weka::classifiers::functions::SMOreg:

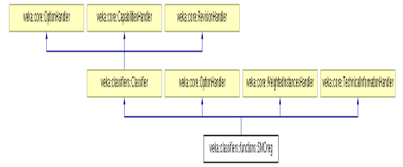

Fig 1 : Weka Functions

A **support vector machine** (**SVM**) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the input, making the SVM a non-probabilistic binary linear classifier.

Algorithm

SMO is an iterative algorithm for solving the optimization problem described above. SMO breaks this problem into a series of smallest possible sub-problems, which are then solved analytically. Because of the linear equality constraint involving the Lagrange multipliers α_i, the smallest possible problem involves two such multipliers. Then, for any two multipliers α_1 and α_2, the constraints are reduced to:

$$0 \le \alpha_1, \alpha_2 \le C,$$

$$y_1\alpha_1 + y_2\alpha_2 = k$$

and this reduced problem can be solved analytically.

The algorithm proceeds as follows:

1. Find a Lagrange multiplier α_1 that violates the Karush–Kuhn–Tucker (KKT) conditions for the optimization problem.
2. Pick a second multiplier $\alpha_2$and optimize the pair (α_1, α_2).
3. Repeat steps 1 and 2 until convergence.

When all the Lagrange multipliers satisfy the KKT conditions (within a user-defined tolerance), the problem has been solved. Although this algorithm is guaranteed to converge, heuristics are used to choose the pair of multipliers so as to accelerate the rate of convergence.

In this study SMOreg algorithm is used to predict half life of peptides. Sequential Minimization Optimization for regression (SMOreg) is a new algorithm for training SVM. This implementation globally replaced all missing values, transformed nominal attributes into binary ones and also normalized all attributes by default. SMOreg is a support vector machine for regression problems. It differs significantly from standard multiple linear regression. It's beyond the scope of these forums to discuss the theory behind kernel methods like SVMs.

The SMOreg regression method is applied on the given data as follows.

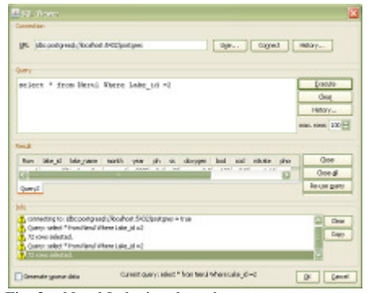

Fig. 2 : Nerul Lake is selected.

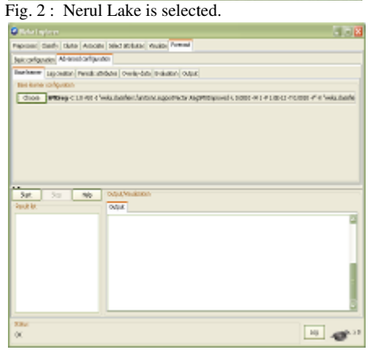

Fig. 3 : SMOreg algorithm is selected.

Fig 4 : Prediction for the parameter pH for the year 2014.

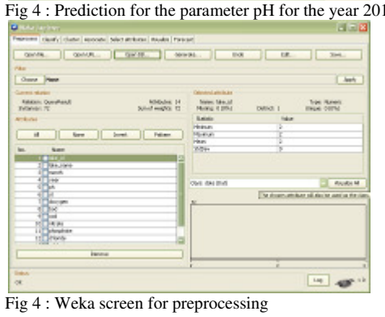

Fig 4 : Weka screen for preprocessing

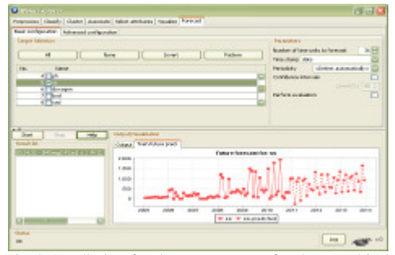

Fig 5 : Prediction for the parameter ss for the year 2014.

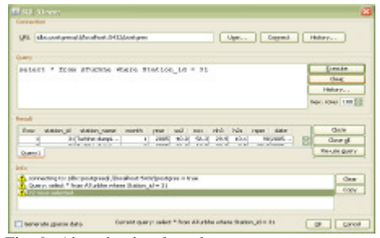

Fig. 6 : Airstation is selected

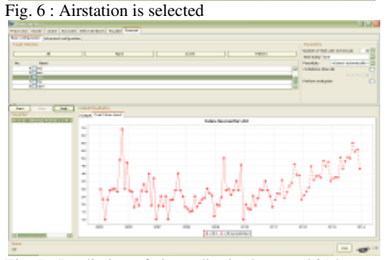

Fig. 7 : Prediction of air quality in the year 2014.

BENEFICIARIES

- Health Practitioners
- Policy makers (CIDCO,NMMC)
- Socialists
- NGO's working for environment
- Any environment conscious individual

V. CONCLUSION

This paper proposes that data mining techniques are valuable tools that could be used to good effect in the environmental and natural resource science field, and are thus of interest to NGO's, Municipal corporations etc.
The project is a good starting point for implementation of data mining for real world examples. This project has brought the insight into various techniques not only in the field of data mining but also in database utilization, visualization, etc.
The advantage of this project is the use of open-source data mining tools, even though WEKA is a very useful alternative many other tools exist that are more robust and feature rich. Utilization of such tools would proved for more open and feature rich application.

References

[1] Shevade, S.K., Keerthi, S.S., Bhattacharyya, C., Murthy, K.R.K.: Improvements to SMO Algorithm for SVM Regression, Technical Report CD-99-16, Control Division, Dept. of Mechanical and Production Engineering, National University of Singapore, Singapore (1999)

[2] Smola, A.J., Scholkopf, B.: A tutorial on support vector regression, NeuroCOLT2 Technical Report Series NC2-TR-1998-030, ESPRIT working group on Neural and Computational Learning Theory Neuro-COLT 2 (1998)

[3] Witten, I.H., Frank, E.: Data mining-Practical Machine Learning Tools and Techniques with Java Implementation. Morgan Kaufmann, San Francisco (2000), http://prdownloads.sourceforge.net/weka/datasets-numeric.jar

[4] Frank, E., Trigg, L., Holmes, G., Witten, I.H.: Naive bayes for regression. Machine Learning 41, 5–15 (2000)
[5] Mitchell, T.M.: Instance-Based Learning. In: Machine Learning, ch. 8, McGraw-Hill, New York (1997)
[6] Platt, J.: Fast training of support vector machines using sequential minimal optimization. In: Scholkopf, B., Burges, C.J.C., Smola, A.J. (eds.) Advances in Kernel Methods. Support Vector Learning, pp. 185–208. MIT Press, Cambridge (1999)
[7] Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to Platt's SMO algorithm for SVM classifier design. Technical Report CD-99-14, Dept. of Mechanical and Production Engineering, Natl. Univ. Singapore, Singapore (1999)
[8] Atkeson, C.G., Moore, A.W., Schaal, S.: Locally Weighted Learning. Artificial Intelligence Review 11(1-5), 11–73 (1997)
[9] Nemallapudi Chaitanya, ElGammal Mahmoud, Sunkara Anish CS 5984 Project Proposal- Crime Data Mining and Visualization.
[10] Frank, E., Hall, M., Pfahringer, B.: Locally Weighted Naive Bayes. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence (2003), pp. 249–256. Morgan Kaufmann, San Francisco (2003).
[11] Nadeau, C., Bengio, Y.: Inference for the generalization error. Advances in Neural Information Processing Systems 12, 307–313 (1999)
[12] Chaoqun Li, Liangxiao Jiang Using locally weighted learning to improve SMOreg for regression, PRICAI'06 Proceedings of the 9th Pacific Rim international conference on Artificial intelligence Pages 375-384 Springer-Verlag Berlin, Heidelberg ©2006.
[13] S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy, "Improvements to the SMO Algorithm for SVM Regression" IEEE Transactions on Neural Networks, 1999. A.J. Smola, B. Schoelkopf (1998). A tutorial on support vector regression.
[14] S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy, *Improvements to SMO Algorithm for SVM Regression*. Technical Report CD-99-16, Control Division Dept of Mechanical and Production Engineering, National University of Singapore.
[15] Alex J. Smola, Bernhard Scholkopf, "A Tutorial on Support Vector Regression", NeuroCOLT2 Technical Report Series - NC2-TR-1998-030 (1998).
[16] Chen., H., et al.,mining: an overview and case studies, in Proceedings of
[17] the 2003 annual national conference on Digital government research.
[18] 2003, Digital Government Society of North America: Boston, MA.
[19] ESR 2010-11
[20] PostgreSQL: http://www.postgresql.org/about/
[21] PostGIS: http://postgis.refractions.net/
[22] JDMP (http://www.jdmp.org)
[23] WEKA (www.cs.waikato.ac.nz/ml/weka/)
[24] API: (http://weka.sourceforge.net/doc/)
[25] JSON: http://www.json.org/
[26] JSON for Java: http://www.json.org/java/
[27] Google Maps. 2010 [cited 2012 September 18th]; Available from:
[28] http://maps.google.com/