Expert Systems with Applications 27 (2004) 331–340

Expert Systems with Applications

www.elsevier.com/locate/eswa

Data mining to aid policy making in air pollution management

Sheng-Tun Li[a,*], Li-Yen Shue[b]

[a]*Institute of Information Management, National Cheng-Kung University, Tainan, Taiwan, ROC*
[b]*Department of Information Management, National Kaohsiung First University of Science and Technology, Kaohsiung, Taiwan, ROC*

Abstract

In the past two decades, the heavy environmental loading has led to the deterioration of air quality in Taiwan. The task of controlling and improving air quality has attracted a great deal of national attention. The Taiwanese government has since set up the National Air Quality Monitoring Network (TAQMN) to monitor nationwide air quality and adopted an array of measures to combat this problem. This study applies data mining to uncover the hidden knowledge of air pollution distribution in the voluminous data retrieved from monitoring stations in TAQMN. The mining process consists of data acquisition from Web sites of 71 data gathering stations nationwide, data pre-processing using multi-scale wavelet transforms, data pattern identification using cluster analysis, and final analysis in mapping the identified clusters to geographical locations. The application of multi-scale wavelet transforms contributes greatly in removing noises and identifying the trend of data. In addition, the proposed two-level self-organization map neural network demonstrates its ability in identifying clusters on the high-dimensional wavelet-transformed space. The identified distribution of suspended particulate PM10 represents a complete, national picture of the present air quality situation, which contrasts the present pollution districts, and could serve as an important reference for government agencies in evaluating present and devising future air pollution policies.

Keywords: Data mining; Air pollution management; Multi-scale analysis; Self-organization neural network; Decision support

1. Introduction

Data mining, also known as knowledge discovery in databases (KDD) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), is the process of discovering useful knowledge from large amount of data stored in databases, data warehouses, or other information repositories. It is a hybrid disciplinary (Zhou, 2003) that integrates technologies of databases, statistics, machine learning, signal processing, and high-performance computing. This rapidly emerging technology is motivated by the need for new techniques to help analyze, understand or even visualize the huge amounts of stored data gathered from business and scientific applications. The major data mining functions that are developed in commercial and research communities include summarization, association, classification, prediction and clustering (Zhou, 2003). Data mining has been shown capable of providing a significant competitive advantage to an organization by exploiting the potential knowledge of large databases (Bose & Mahapatra, 2001). Recently, a number of data mining applications and prototypes have been developed for a variety of domains (Liao, 2003; Mitra, Pal, & Mitra, 2002), including marketing, banking, finance, manufacturing, and health care. In addition, data mining has also been applied to other types of scientific data (Abidi, 2001; Read, 2000) such as bioinformatical, astronomical, and medical data.

In general, techniques and functions that are to be applied in a data mining process depend very much on the application domain and the nature of the data available. This creative process generally involves phases of data understanding, data preparation, modeling, and evaluation (Fayyad et al., 1996). Data understanding starts with an initial data collection and proceeds with activities to get familiar with the data, to identify data quality problems, and to discover first insights into the data. Data preparation covers all activities that construct the final dataset to be modeled from the initial raw data. The tasks of this phase may include data cleaning for removing noise and inconsistent data, and data transformation for extracting the embedded features. The modeling phase applies various modeling techniques, determines the optimal values for parameters in models, and finds the one most suitable to meet the objectives. The evaluation phase evaluates the model found in the last stage to

* Corresponding author. Address: No. 1, Ta-Hsueh Road, Tainan 701, Taiwan, ROC. Tel.: + 886-6-2757575x53126, fax: +886-6-2362162.

E-mail address: stli@mail.ncku.edu.tw (S.-T. Li).

0957-4174/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2004.05.015

confirm its validity to fit the problem requirements. No matter which areas data mining is applied to, most of the efforts are directed toward the data preparation phase (Pyle, 1999). In this study of mining air pollution data, our data preparation phase particularly emphasizes the data scale issue.

The purpose of this study is to apply data mining technology to identify the national air quality distribution of Taiwan, whose hourly air quality data are continuously collected and archived through a network of 71 EPA stations. In dealing with voluminous data, we combine both wavelet transform (WT) and self-organization map (SOM) neural networks as our data mining technology. The former is accredited with capability of investigating temporal variation with different scales, and the latter is known to be effective in isolating clusters in high-dimensional space. With both technologies, one can benefit from better understanding and interpretation of the pollution data. The rest of this paper is organized as follows. Section 2 provides a brief review of air pollution management in Taiwan. Section 3 presents the issues of mining air quality data from the EPA Web site and the underlying technologies for dealing with the issues. Section 4 elaborates on the mining procedure that consists of data acquisition, missing-value handling, data transform, modeling, and performance evaluation. Section 5 discusses the mining results and its comparisons with official distribution districts. Section 6 concludes this paper.

2. Air pollution management in Taiwan

The island of Taiwan runs from north to south like a sweet potato and is divided into an eastern seaboard and western seaboard by the mountain range that runs also from north to south. With a total area of 35,873 square kilometers and only 26% of it being plain, her population of more than 21 millions represents one of the densest countries in the world. The western seaboard contains a much wider area of plain than the eastern one and is hence much densely populated and also heavily industrialized. Presently, for every square kilometer, Taiwan has 611 people, 453 vehicles, 2.78 factories. Her present population and vehicle density are about 2 times those of Japan, 3 times those of Germany and British, 22 times those of America, and the factory density, ranging from 2.4 to 69.5 times, is even worse. With such a heavy environmental loading, the industrial and related pollution have, in the past 30 years, caused air quality to deteriorate alarmingly, and the air quality control and improvement task has become an urgent task for successive governments. The EPA (Environmental Protection Administration) of Taiwan was formally established in 1987 and was given the mission to control and improve national air quality. It later set up the National Air Quality Monitoring Network (TAQMN) in 1990 to monitor nationwide air quality.

TAQMN, presently, consists of 71 air quality monitoring stations on the Taiwan Island, and can automatically collect and monitor air quality on an hourly basis (EPA, 2000). In addition to others, each monitor station may collect one or more of the five major types of priority pollutants: PM10 (suspended particulate), SO2 (sulfur dioxides), NO2 (nitrogen dioxide), CO (carbon monoxide), and O3 (ozone), with PM10 and O3 being the main air pollutants. EPA also maintains a Web site for each station for publishing archived and real-time pollutant information and forecasting as well. The locations of these stations are mainly based on population distribution, as shown in Fig. 1, and are distributed among eight areas on the island: T-K (Taipei–Keelung), Ilan, T-H-M (Taoyuan–Hsinchu–Miaoli), T-C (Taichung–Changhua), Nantou, Y-C (Yunlin–Chiai), T-K-P (Tainan–Kaohsiung–Pingtung), and H-T (Hualien–Taitung).

In order to control air quality and reverse the trend, successive governments have adopted the polluter-pay principle and have devised a strategy that consists of both control component and incentive component. One main task of the control component is to map out air quality districts with different ratings and impose different pollution levy for each rating. The incentive component provides financial incentives to encourage industry owners to replace existing equipments with newer ones or install better pollution reduction devices. At the same time, a Pollution Prevention Fund was legislated for pooling all levies together for the common goal. The central government is currently retaining 40% of this fund for national infrastructure use and refunding 60% back to the district governments for local use. This 60% refund from the Pollution Prevention Fund has arisen much interest among local governments, and it was not hard to see the involvement of politics in drawing up the air quality districts for getting the maximum refund. Thus, the ideal process of determining air quality districts, which should be based on air quality, population density, and different types of land utilization, has been greatly compromised during the process. The present air quality control district, as is shown in Fig. 1, is based on the official government administrative districts.

In this study, we intend to investigate the current air quality distribution by applying data mining to the pollution

Fig. 1. Air quality administration districts on Taiwan.