Knowledge-Based Systems 22 (2009) 336–343

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

# Bayesian networks based rare event prediction with sensor data

Seong-Pyo Cheon [a], Sungshin Kim [a,*], So-Young Lee [b], Chong-Bum Lee [c]

[a] School of Electrical and Computer Engineering, Pusan National University, 609735 Busan, Republic of Korea
[b] Department of Atmospheric Sciences, Pusan National University, 609735 Busan, Republic of Korea
[c] Department of Environment Engineering, Kangwon National University, Kangwon-do, Republic of Korea

ARTICLE INFO

*Article history:*
Received 16 May 2007
Received in revised form 29 August 2008
Accepted 17 February 2009
Available online 27 February 2009

*Keywords:*
Bayesian network
Rare event prediction
Ozone forecasting

ABSTRACT

A Bayesian network is a powerful graphical model. It is advantageous for real-world data analysis and finding relations among variables. Knowledge presentation and rule generation, based on a Bayesian approach, have been studied and reported in many research papers across various fields. Since a Bayesian network has both causal and probabilistic semantics, it is regarded as an ideal representation to combine background knowledge and real data. Rare event predictions have been performed using several methods, but remain a challenge. We design and implement a Bayesian network model to forecast daily ozone states. We evaluate the proposed Bayesian network model, comparing it to traditional decision tree models, to examine its utility.



## 1. Introduction

Scientists have reported the adverse effects of high ozone ($O_3$) concentration [4,20]. Kelsall et al. [17] researched the cause and effect of air pollution and mortality using 15 years data from Philadelphia. Schwartz et al. [26] found associations between acute respiratory symptoms in children and summer air pollution. In the light of these health effects, many models have been suggested and developed during the past two decades. Previous models are mainly based on empirical models and/or expert opinions, statistical models, causal models, or combinations of these [5,15,16,23]. Despite the various models developed, many scientists still seek more accurate and reliable models, because $O_3$ changes its reaction mechanism depending on altitude, location, and other factors. In this paper, Bayesian approaches are adopted to combine real measured data and expert knowledge to overcome complexity and nonlinearity of $O_3$ reactions. The proposed approach has the following advantages compared to traditional models:

- It is an almost ideal combination of sensor data and knowledge.
- It is simple to handle missing or abnormal sensor data.
- It is easy to understand a forecast result and its cause and effect.
- It is easy to analyze high concentration $O_3$ episodes.
- It is easily applied and customized to other areas.

In this paper, we especially consider the rare event state prediction field with a Bayesian network in daily maximum $O_3$ forecasts in Seoul, Korea. The construction of the proposed Bayesian network is shown in Fig. 1. This Bayesian approach starts from historic data and expert knowledge. The historic data contains weather and pollution data when high concentration $O_3$ events have occurred. A variety of data analysis methods have been applied to extract features and create episodes from historic data. Conversely, expert knowledge included $O_3$ related chemical equations and hypotheses make a skeletal Bayesian network. Data analysis and the skeletal Bayesian network are completed as a Bayesian network model by the learning algorithm and belief propagation.

The Bayesian network related to forecasting $O_3$ is discussed in Section 2. Section 3 includes the design procedures for the Bayesian network model, methods of representing the expert knowledge, and atmospheric chemistry hypotheses. The results of an evaluation of the proposed model and typical decision tree models are compared and discussed in Section 4. Conclusions will be addressed in Section 5.

## 2. Bayesian networks for forecasting ozone states

Since Good [8] established the probability representation and Bayesian inference methods in 1961, many researchers have illustrated the usefulness and performance of Bayesian networks in various fields. Pearl et al. [14] developed CONVINCE, the first expert system using Bayesian networks. Munin [1], Pathfinder [9], MS-Windows diagnosis-and-repair modules [10], and Office Assistant [12] have also been developed and verified the abilities of a Bayesian network. This study considers only the essential ideas

* Corresponding author. Tel.: +82 51 510 2374; fax: +82 51 513 0212.
*E-mail addresses:* buzz74@pusan.ac.kr (S.-P. Cheon), sskim@pusan.ac.kr (S. Kim), shanny00@pusan.ac.kr (S.-Y. Lee), cbl@kangwon.ac.kr (C.-B. Lee).

0950-7051/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2009.02.004

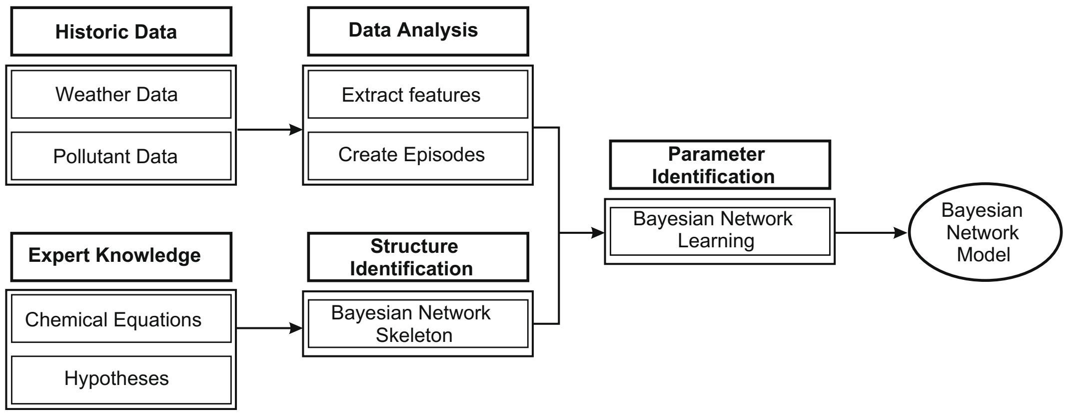


**Fig. 1.** Construction of the proposed Bayesian network model.

about the structure and use of Bayesian networks as provided by Heckerman. He briefly addressed what Bayesian networks and Bayesian methods have to offer. There are summarized as follows: Bayesian networks can readily handle incomplete data sets. Bayesian networks allow one to learn about causal relationships. Bayesian networks, in conjunction with Bayesian statistical techniques, facilitate the combination of domain knowledge and data. Bayesian methods, in conjunction with Bayesian networks and other kinds of models, offer an efficient and principled approach to avoid over-fitting data [11].

### 2.1. Bayesian network representation

Bayesian networks are directed acyclic graphs (DAGs), in which nodes represent random variables and the lack of an arc between two nodes represents the conditional probabilistic independence of the two unlinked variables. The network's structure is a factorization of a joint probability distribution [19]. A node without parents is called a *root node* and a node without children is a *leaf node*. Consider an example concerning $O_3$ *State* related variables in the atmosphere. As shown in Fig. 2, the $O_3$ state of the previous day (*Pre* $O_3$), non-methane hydrocarbons (NMHC), sulfur dioxide ($SO_2$), $O_3$, and nitric oxide (NO) are root nodes and the $O_3$ *State* is a leaf node. Root nodes, which have no causal node, may only be associated with unconditional probabilities, other nodes. That is, *Pollutants* and $O_3$ *State* must have own conditional probability table. Table 1 shows a conditional probability table that is related to the *Pollutants* node and its parent nodes. We assume NMHC, $SO_2$, $O_3$, and NO have their own *States* whether *true* (T) or *false* (F), respectively. If *States* of NMHC, $SO_2$, $O_3$, and NO are all 'T', the *True State* probability of *Pollutants* is about 0.95.

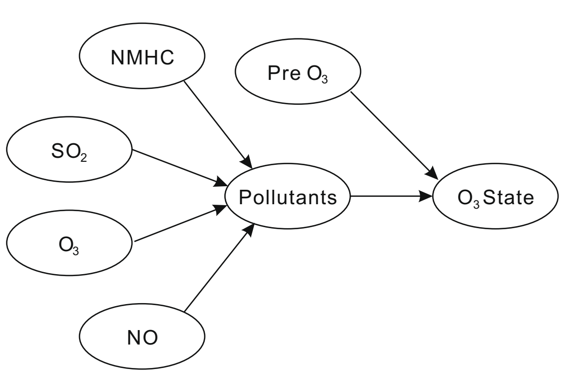


**Fig. 2.** Simple representative of a Bayesian network.

## 3. Applied bayesian network model for ozone forecasting

Creating a Bayesian model, poses challenging issues that must be solved. There are summarized as follows:

- The data are deficient, since the occurrence of high concentrations of $O_3$ are rare. Existing data sets generally include noise and abnormal data.
- Knowledge acquisition is difficult; it is not easy to make logical rules from complex knowledge.
- The learning procedure cannot be specified systematically, since high $O_3$ episodes are created purely from expert knowledge.

Fig. 3 shows the schematic diagram of $O_3$ related systems such as on-line monitoring stations/systems, the prediction system, and the warning system. Seoul has 27 monitoring stations. Usually, one or two forecasters have to monitor and analyze all the points simultaneously. Therefore, the proposed Bayesian network model may assist the forecaster in the following cases:

- While a forecaster is analyzing a specific area, the BN model can automatically alert a forecaster when abnormal signals or symptoms are detected.
- When a forecaster wants to determine the causes of a high concentration $O_3$ event, the BN model can display candidate causes listed from previous similar cases.

**Table 1**
Example conditional probability table in a Bayesian network.

| NMHC | $SO_2$ | $O_3$ | NO | Pollutants |
|---|---|---|---|---|
| T | T | T | T | 0.95 |
| T | T | T | F | 0.80 |
| T | T | F | T | 0.85 |
| T | T | F | F | 0.50 |
| T | F | T | T | 0.85 |
| T | F | T | F | 0.50 |
| T | F | F | T | 0.50 |
| T | F | F | F | 0.20 |
| F | T | T | T | 0.85 |
| F | T | T | F | 0.50 |
| F | T | F | T | 0.50 |
| F | T | F | F | 0.20 |
| F | F | T | T | 0.50 |
| F | F | T | F | 0.20 |
| F | F | F | T | 0.20 |
| F | F | F | F | 0.05 |