Environmental Modelling & Software 80 (2016) 259–264

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Air pollution prediction via multi-label classification

Giorgio Corani*, Mauro Scanagatta

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Scuola Universitaria Professionale della Svizzera Italiana (SUPSI), Università della Svizzera Italiana (USI), Galleria 1, Manno, Switzerland

ARTICLE INFO

Article history:
Received 15 July 2015
Received in revised form
19 February 2016
Accepted 24 February 2016

Keywords:
Bayesian networks
Air pollution prediction
Statistical classification
Multi-label classification

ABSTRACT

A Bayesian network classifier can be used to estimate the probability of an air pollutant overcoming a certain threshold. Yet multiple predictions are typically required regarding variables which are stochastically dependent, such as ozone measured in multiple stations or assessed according to by different indicators. The common practice (independent approach) is to devise an independent classifier for each class variable being predicted; yet this approach overlooks the dependencies among the class variables. By appropriately modeling such dependencies one can improve the accuracy of the forecasts. We address this problem by designing a multi-label classifier, which simultaneously predict multiple air pollution variables. To this end we design a multi-label classifier based on Bayesian networks and learn its structure through structural learning. We present experiments in three different case studies regarding the prediction of PM2.5 and ozone. The multi-label classifier outperforms the independent approach, allowing to take better decisions.

1. Introduction

Statistical air pollution prediction is an important task in environmental modeling. The pollutants most commonly studied are ozone (Schlink et al., 2003) and particulate matter (Perez, 2012); see the references therein for a wider bibliography.

Throughout the introduction we assume ozone as the pollutant to be predicted. However our methodology readily applies to any other pollutant.

The decision maker typically needs to know the probability of ozone overcoming a threshold deemed relevant for health. Once we discretize the ozone concentration using this threshold we have a discrete variable. The task is then to estimate the probability of ozone exceeding the threshold on the basis of different *features*, typically constituted by past values of meteorological variables and air pollutants. According to the machine learning terminology this is a *classification* problem. The variable being predicted is referred to as the *class* variable.

Bayesian networks (Koller and Friedman, 2009) are probabilistic models suitable for classification. They represent the joint distribution of a set of random variables via a directed acyclic graph (DAG) and its associated conditional probability tables. The DAG constitutes the *structure* of the network; each node of the DAG represents a random variable. The edges of the DAG encode the assumptions of conditional independence. A Bayesian network performs a probabilistic *inference* when it estimates the posterior probability of the states of some variable(s) given the observation of some other variable(s). In classification we make inference about the class variable given the observation of the features. A state of the art classifier based on Bayesian networks is the extended tree-augmented naive classifier (ETAN) (de Campos et al., 2016), which overcomes the limits of previous algorithms such as naive Bayes and tree-augmented naive classifier (TAN) (Friedman et al., 1997).

Typically the decision maker requires prediction regarding *multiple* variables such as ozone measured in *multiple* stations, assessed according to by *different* indicators (1-h maximum value and 8-h moving average) and over *different* days (typically, today and tomorrow). The common practice is to devise an independent classifier for each class variable being predicted; yet this approach overlooks the dependencies existing among the class variables. By appropriately modeling the dependencies between class variables one can improve the accuracy of the forecasts; this is the focus of this paper.

Multi-label classification (Read et al., 2011) is the machine learning area which studies how to jointly predict multiple dependent class variables (*labels*). We adopt multi-label classification to simultaneously predict multiple air pollution variables. This is the first application of multi-label classification in environmental modeling, as far as we know.

Our multi-label classifier generalizes the ideas underlying the

[*] Corresponding author.
E-mail addresses: giorgio@idsia.ch (G. Corani), mauro@idsia.ch (M. Scanagatta).

http://dx.doi.org/10.1016/j.envsoft.2016.02.030
1364-8152/© 2016 Elsevier Ltd. All rights reserved.

ETAN classifier to multi-label classification, yielding a model which makes simultaneous inference about *multiple* class variables given the value of the features.

We compare the multi-label classifier against the traditional approach of devising an independent classifier (ETAN in our case) for each class variable.

We consider three case studies: (i) prediction of $PM_{2.5}$ in eight stations in Shanghai for today and tomorrow (16 class variables); (ii) prediction of ozone in Berlin for today and tomorrow, considering the threshold for both 1-h and 8-h concentration (4 class variables); (iii) prediction of ozone in Burgas (Bulgaria) for today and tomorrow, considering the threshold for both 1-h and 8-h concentration (4 class variables).

In each case study the multi-label classifier consistently outperforms the independent approach; thus it provides better support for the decision maker.

The application of multi-label classifiers in environmental modeling is not limited to air pollution. Instead, it is suitable to the many applications in which it is required to predicting multiple dependent discrete variables. For instance multi-label classification could become an important tool for ecological modeling, being able to simultaneously predict the presence/absence of different species accounting for pray–predators relations. It could constitute a step forward compared to the development of single-species model. Attempts in this direction are discussed by De'Ath (2002); Chapman and Purse (2011).

2. Bayesian networks classifiers

We denote by C the *class* variable and by A:=(A_1, …,A_k) the set of features, typically constituted by the past observations of meteorological and air pollution variables. For a generic variable A, we denote as $P(a)$ the probability that $A = a$.

There are different approaches for classification based on Bayesian networks.

The Naive Bayes classifier assumes the stochastic independence of the features given the class, factorization the joint probability as follows:

$$P(c,\mathrm{a}) := P(c)\prod_{j=1}^{k} P(a_j|c), \qquad (1)$$

corresponding to the topology of Fig. 1(a). However the posterior probabilities computed by naive Bayes are biased by such unrealistic assumption (Hand and Yu, 2001).

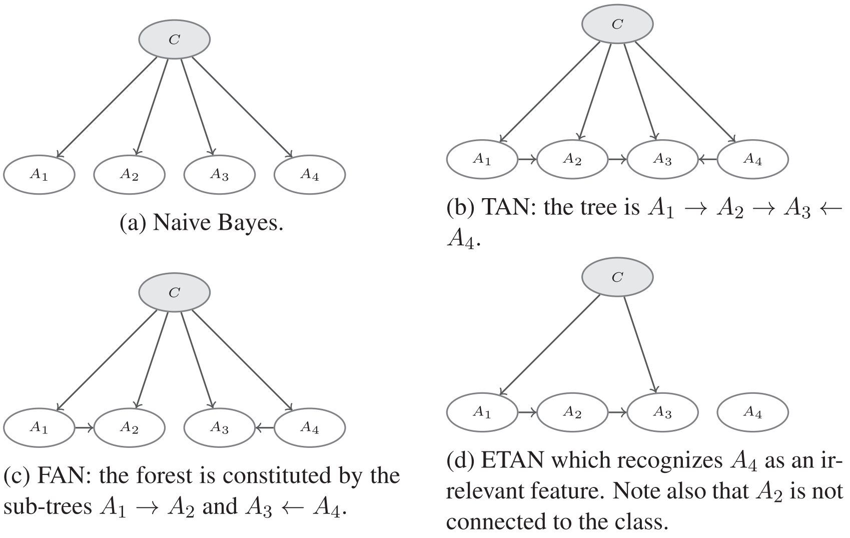
The tree-augmented naive classifier (TAN) (Friedman et al., 1997) relaxes this assumption, augmenting the naive Bayes structure with a tree which connects the features. A tree is a graph in which any two vertices are connected by a unique path. As a result one feature has only the class as parent, while the remaining k-1 features have two parents: the class and another feature. An example is shown in Fig. 1(b). The optimal tree is identified by maximizing a score which evaluates how well the graph fits the joint probability distribution of the variables. A discussion of the scores for Bayesian networks is given in (Koller and Friedman, 2009; Chap.18.3). The structural learning algorithm which exactly identifies the maximum-scoring tree has been devised by Friedman et al. (1997).

TAN is further improved by the forest-augmented naive classifier (FAN). A FAN augments the naive Bayes with a forest. A forest is a set of disjoint trees; it is more general than a tree, as it includes the tree as a special case. Thus the BIC score of FAN is higher or equal than the BIC score of TAN. An example of FAN is given in Fig. 1(c). The structural learning algorithm of FAN (Koller and Friedman, 2009; Chap.18.4.1) is obtained as a slight modification of the TAN algorithm.

A limit of both TAN and FAN is that they do not perform feature selection; each feature is forcedly connected to the class without checking if it is relevant. The extended tree-augmented naive (ETAN) (de Campos et al., 2016) overcomes this problem. ETAN allows each feature to have as parent either (i) the class; (ii) the class and a feature; (iii) a feature without the class; (iv) no parent, in which case the feature is recognized as irrelevant. The structural learning algorithm of ETAN (de Campos et al., 2016) exactly identifies the highest-scoring graph which satisfies the previous constraints. This algorithm is more complex than that of TAN and FAN. The ETAN includes naive Bayes, TAN and FAN as special cases; thus it achieves a higher BIC score (equal score in the worst case) than all of them.

3. Multi-label classifier

We devise a multi-label classifier which represents the joint distribution of all the class variables and the features used to predict them. We learn from data the structure of the multi-label classifier, imposing the following constraints: each class can have



(a) Naive Bayes.

(b) TAN: the tree is $A_1 \rightarrow A_2 \rightarrow A_3 \leftarrow A_4$.

(c) FAN: the forest is constituted by the sub-trees $A_1 \rightarrow A_2$ and $A_3 \leftarrow A_4$.

(d) ETAN which recognizes A_4 as an irrelevant feature. Note also that A_2 is not connected to the class.

Fig. 1. Different Bayesian networks classifiers. The class variables are shown in gray.