

Air pollution prediction via multi-label classification

Giorgio Corani and 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)
giorgio{mauro}@idsia.ch

Air pollution prediction via multi-label classification

Giorgio Corani and 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)
giorgio{mauro}@idsia.ch

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 multilabel classifier, which simultaneously predict multiple air pollution variables. To this end we design a multilabel classifier based on Bayesian networks and learn its structure through structural learning. We present experiments in three different case studies regarding the prediction of PM_{2.5} and ozone. The multi-label classifier outperforms the independent approach, allowing to take better decisions.

Keywords: Bayesian networks, air pollution prediction, statistical classification, multilabel classification

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-hour maximum value and 8-hours 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.

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

Our multilabel classifier generalizes the ideas underlying the ETAN classifier to multilabel classification, yielding a model which makes simultaneous inference about *multiple* class variables given the value of the features.

We compare the multilabel 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-hour and 8-hours concentration (4 class variables); (iii) prediction of ozone in Burgas (Bulgaria) for today and tomorrow, considering the threshold for both 1-hour and 8-hours concentration (4 class variables).

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

The application of multilabel 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 multilabel classification could become an important tool for ecological modeling, being able to simultaneously predict the presence/absence of different species accounting for prey-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 $\mathcal{A} := (A_1, \dots, A_k)$ 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, factorizing the joint probability as follows:

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

corresponding to the topology of Fig.(1a). 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

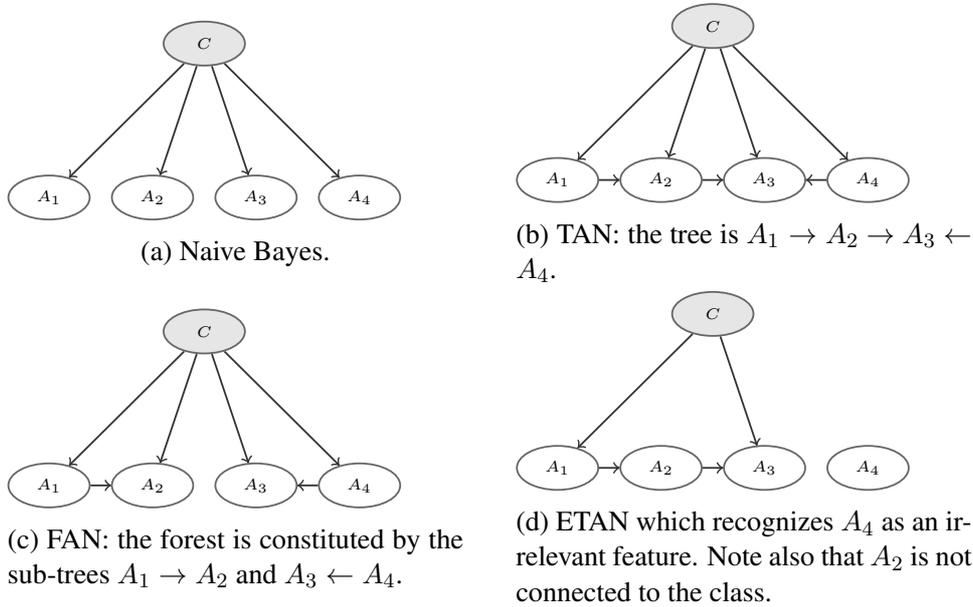


Figure 1: Different Bayesian networks classifiers. The class variables are shown in gray.

in Fig.(1b). 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.(1c). 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. Multilabel classifier

We devise a multilabel 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 multilabel classifier, imposing the following constraints: each class can have as parent at most one other class; each feature can have as parents any number of classes and up to one feature. We call METAN (Multilabel ETAN) the resulting classifier.

Like ETAN, our multilabel classifier a) performs feature selection since the parent set of a feature can be empty; b) connects the features by a forest. However METAN extends ETAN since each feature can have multiple classes as parents and moreover the class variables are linked by a forest.

An example of METAN structure with features is given in Fig.2.

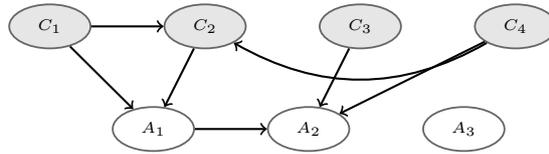


Figure 2: Example of METAN classifier; the class variables are shown in gray. The class variables are linked by a forest; the feature variables are linked by a forest; feature A_3 is recognized as irrelevant for classification.

In multilabel classification we make inference about the class variables given the observed value of all features. Given the METAN structure, this is equivalent to make inference on a forest, which is extremely quick.

3.1. Structural learning

To identify the highest-scoring METAN structure we need to solve a problem of structural learning. We score a candidate DAG \mathcal{G} through the BIC score (Koller and Friedman, 2009, Chap.18.3) which is the log-likelihood minus a term which

accounts for the model complexity. The BIC score of DAG \mathcal{G} is as follows:

$$\text{BIC}(\mathcal{G}) = \sum_{X \in \mathcal{X}} \sum_{\pi_x \in |\Pi_x|} \sum_{x \in |X|} N_{x, \pi_x} \log \hat{\theta}_{x|\pi_x} - \frac{\log N}{2} (|X| - 1) (|\Pi_x|) \quad (2)$$

$$= \sum_{X \in \mathcal{X}} \text{BIC}(X, \Pi_x) \quad (3)$$

where \mathcal{X} denotes the set of random variables that we are considering; X is a generic random variable belonging to \mathcal{X} ; Π_x is the parent set of X ; x is a generic value of X and π_x is a generic value of Π_x ; $|X|$ denotes the *arity* (number of states) of variable X and $|\Pi_x|$ denotes the product of the arities of the parents of X . Moreover, $\hat{\theta}_{x|\pi_x}$ is the maximum likelihood estimate of the conditional probability $P(X = x | \Pi_x = \pi_x)$; N_{x, π_x} denotes the number of times in which in the data set it appears the configuration $(X = x, \Pi = \pi_x)$. Assuming all models to be equally probable a priori, the BIC is approximately proportional to the posterior probability of the model.

The BIC score is decomposable: as shown in Eq.3, it corresponds to the sum of the local BIC scores of each variable X given its parent set Π . The structural learning problem is typically addressed in two steps. First we identify independently for each variable a list of candidate parent sets (*parent set identification*). Then we select for each node the parent set which yields the highest-scoring DAG without introducing cycles (*structure optimization*).

During parent set identification we compute the local score for each variable X and of each candidate parent set Π . In order to make this approach feasible, one typically has to set a limit on the maximum number of parents (typically two or three) per node. The number of parents of a node is its *in-degree*. Yet a problem is that there might exist high-scoring parent sets with in-degree higher than the limit and which are overlooked during structural learning. To overcome this issue we adopt an innovative approach (Scanagatta et al., 2015) for parent set identification. It first approximately estimates the score of a large number of parent sets without limit on the in-degree and then exactly computes the score of only the most promising parent sets. In the METAN case the number of possible parent sets is especially high for the feature variables.

As for the structure optimization problem, we encode it as an integer program following Cussens (2011). For each variable X and each candidate parent set Π we define a binary variable $I(X, \Pi)$ which is one if Π is parent of X in the optimal DAG and zero otherwise.

Let us denote the set of features as \mathcal{F} and set of classes as \mathcal{C} , so that: $\mathcal{X} =$

$\mathcal{F} \cup \mathcal{C}$. The structural learning problem of METAN can be then casted as follows:

Instantiate each $I(X, \Pi)$ to maximize:

$$BIC(\mathcal{G}) = \sum_{X, \Pi} BIC(X, \Pi) I(X, \Pi)$$

subject to:

- the $I(X_i, \Pi)$ representing a DAG;
- the parent set of each class variable being either empty or containing another class variable: $\forall C_i \in \mathcal{C} : \Pi_{C_i} \in \{C_j, \emptyset\}, j \neq i$
- the parent set of each feature variable containing any number of classes and up to one another feature: $\forall F_i \in \mathcal{F} : \Pi_{F_i} = \mathcal{C}_{F_i} \cup \mathcal{F}_{F_i} : \mathcal{C}_{F_i} \subset \mathcal{C}, \mathcal{F}_{F_i} \in \{F_j, \emptyset\}, j \neq i$,

where \mathcal{C}_{F_i} and \mathcal{F}_{F_i} denotes the class variables and the feature variables which are parents of F_i . The first constraint ensure that only valid DAG which do not contain cycles constitute a feasible solution of the problem. The last two constraints guarantee that the obtained structure is a METAN. We solve the above optimization problem using the Gobnilp open-source package for structural learning (Cussens, 2011).

Gobnilp identifies extremely quick the optimal METAN; in our experiments the process was always completed in less than a second.

Before running the structural learning solver, we discretize all the feature variables in four bins.

4. Assessing the predictions

We can assess the posterior probabilities computed by a probabilistic model through different indicators. Consider predicting the class variable C .

The *accuracy* is the proportion of correct classifications. A prediction is accurate if the predicted class matches the actual class. Accuracy is maximized by returning as a prediction the most probable class c^* :

$$c^* := \arg \max_{c \in \mathcal{C}} P(c|\mathbf{a}),$$

where \mathcal{C} denotes the set of possible values of the class variable. Accuracy is commonly used but it has a severe drawback: a model can achieve high accuracy without being informative. Consider a classifier which always predicts the most

common class (*majority predictor*). If the most common class is very frequent, the majority predictor is highly accurate without providing any useful information.

A more robust indicator which does not depend on the proportion of the different classes is the area under the receiver operating curve (AUC). Consider a two-classes problem in which the samples are either positive or negative. Assume to rank the samples according to the posterior probability estimated by the model of being positive; the AUC is then the probability that a randomly chosen positive sample is ranked before a randomly chosen negative sample. This is the most straightforward explanation of AUC; see (Ling et al., 2003; Bradley, 1997) for a more detailed discussion and alternative interpretations.

The *success index* is another important indicator in the literature of air pollution prediction. The idea is as follows. A model can make two types of error: false positive (false alarms) and false negatives (missed alarms). Denote by a the correctly predicted exceedances, by f all the predicted exceedances, by m all observed exceedances and by n the total number of observations. The true positive rate is the fraction of correctly predicted exceedances: $tpr=a/m$. The false positive rate is $fpr=(f-a)/(n-m)$. The success index is $si=tpr-fpr$ (Schlink et al., 2003). The success rate is maximized by returning as a prediction the most probable class c^* (Dorling et al., 2003).

Finally we discuss the cost-sensitive (Elkan, 2001) indicators; they are based on the consideration that we have to decide whether to issue an alarm and that a missed alarm is more costly than a false alarm. Let us denote by c_{fa} the cost a false alarm (false positive) and by c_{ma} the cost of a missed alarm (false negative). Let us also assume that a correct prediction implies no cost. To minimize the expected cost of our decision, we should issue an alarm (Dorling et al., 2003) if the posterior probability of ozone being *high* is greater than the threshold $t(c_{fa}, c_{ma})$:

$$t(c_{fa}, c_{ma}) = \frac{c_{fa}}{c_{fa} + c_{ma}}. \quad (4)$$

We issue the alarm according to the previously discussed policy and we count how many false positives and false negatives the model generates. We then compute the mean cost-per-decision of the model.

The relative cost of false alarms and missed alarms is not easy to quantify. We consider two scenarios. In the first we set $c_{fa}=1$ and $c_{ma}=5$; we denote the mean cost-per-decision attained by the model in this scenario c_5 . In the second scenario we set $c_{fa}=1$ and $c_{ma}=10$; we denote as c_{10} the mean cost-per-decision in this case.

We compute the indicators via ten runs of ten-folds cross-validation, as recommended by Witten and Frank (2005). Thus we perform 100 training/test exper-

iments and we report the mean value of the indicator over the 100 experiments. The cross-validation folds are identically generated for the independent approach and for the multilabel classifier.

5. Predicting $PM_{2.5}$ in Shanghai

We consider the Shanghai data set published by Zheng et al. (2013). The data covers the period between February 2013 to February 2014. It contains data regarding 10 stations; however we removed two stations from the analysis as they contain about 75% of missing data. The following variables are measured at each station: $PM_{2.5}$ air quality index (AQI), NO_2 air quality index, temperature, pressure, humidity, wind, weather. Weather is a categorical variable, with categories {snowy, cloudy, sunny, overcast, rainy, foggy, dusty}. All variables are recorded with hourly frequency. We average them on the daily scale. For weather, which is categorical, we take the most frequent value during the day. Moreover we add two further variables: the season of the year and the day of the week.

The air quality index (AQI) is an index for reporting daily air quality, described for instance in Zheng et al. (2013) and also at the url <http://www.airnow.gov>. We binarize the AQI at 150. An AQI smaller than 150 (*low* in our terminology) correspond to a $PM_{2.5}$ AQI which is either good (0–50), moderate (50–100) or unhealthy for sensitive groups (100–150). An AQI larger than 150 (*high* in our terminology) implies that the $PM_{2.5}$ AQI can be unhealthy (151–200), very unhealthy (200–300), hazardous (>300).

Averaging over all the eight stations, the proportion of data in which the $PM_{2.5}$ air quality index is low and high are respectively 49% and 51%. The prediction task is to predict whether the $PM_{2.5}$ air quality index will be high or low in each of the eight stations, today and tomorrow. This yields 16 class variables to be predicted. We learn 1600 ETAN structures as we perform 100 training experiments (10 runs of 10 folds cross-validation) for each class variable.

We can assess the importance of each feature by measuring how frequently it is connected to the class variable within the learned ETAN structure. The feature which is most frequently connected to the class variable is yesterday’s $PM_{2.5}$, confirming the well-known persistency of $PM_{2.5}$. It is connected to the class variable in about 87% of the experiments. Pressure and humidity are instead the least connected features. They are connected to the class in less of the 10% of the experiments.

In Fig.3 we show the structure (a forest) which the multilabel classifier learns among the 16 class variables. There are two nodes for each station, representing

the $PM_{2.5}$ of today and tomorrow. Most links connect the $PM_{2.5}$ recorded in two different stations in the same day, suggesting that spatial dependence dominates temporal dependence in this data set.

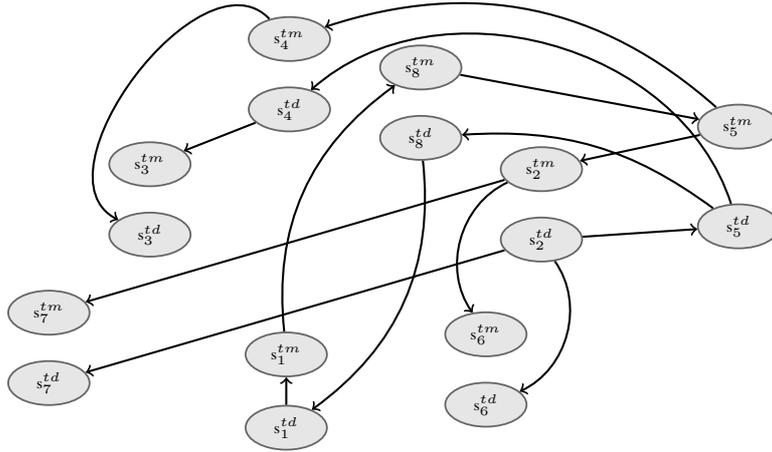


Figure 3: The graph of the METAN model linking the class variables in the Shanghai case study. By s_i^{td} we denote the air quality index on station i for today; by s_i^{tm} we denote the air quality index on station i for tomorrow. Most links connect class variables referring to ozone in different stations in the same day.

In Table 1 we compare the performance of the independent approach and the multilabel classifier. The multilabel classifier outperforms the independent approach yielding higher accuracy, higher auc, higher success index and lower costs. We also check the statistical significance of such differences through a Wilcoxon signed-rank test ($\alpha=0.05$). Consider for instance accuracy. We have two samples containing 16 values each: the accuracies achieved in cross-validation by the two classifiers on the 16 class variables. The test checks whether the mean differences between the two samples is significantly different from 0. The difference in favor of the multilabel classifier is statistically significant on most indicators.

6. Predicting ozone in Berlin

We analyze the ozone data of Berlin, already analyzed in Schlink et al. (2003). The data regard a single station in the years 1997-1999. We remove the period October-March of every year which is not relevant for ozone prediction. Eventually the data set contains 178 daily recordings.

Mean values over 16 class variables			
	multilabel	independent	significance
accuracy	0.64	0.61	✓
auc	0.63	0.58	✓
c ₅	1.74	2.29	–
c ₁₀	3.87	4.84	–
success index	0.29	0.21	✓

Table 1: Indicators of performance of the multilabel classifier and the independent model.

Two types of exceedances are considered in Schlink et al. (2003). The first is the threshold of $180 \mu\text{g}/\text{m}^3$ on the maximum hourly ozone concentration (max_{1h}). There are only 4 exceedances of this threshold in the data set, which makes it hard to collect reliable statistics on how the model can predict its exceedances. We slightly decrease the max_{1h} threshold to $150 \mu\text{g}/\text{m}^3$, obtaining nine exceedances (about 5% of the data set). This allows a more robust comparison among competing predictive models. The second class variable is the 8-hours moving average ozone concentration (max_{8h}), whose threshold is $120 \mu\text{g}/\text{m}^3$. This threshold is exceeded about 5% of times in the data set.

The features are yesterday’s values of: air temperature, wind speed, wind direction, NO_x (nitrogen oxides), NO_2 (nitric dioxide), NO (nitrogen oxide), O_3 (ozone). Moreover we introduce the day of the week and the season of the year as further features. We have thus 9 features. We consider the prediction of four class variables, namely the exceedance of max_{1h} and max_{8h} , both for today and tomorrow. The stochastic dependence between max_{1h} and max_{8h} are due to the fact that these are two similar indicators, and yet they characterize in a different way the ozone pollution. It is sensible to predict them jointly.

When running the independent approach we learn 400 ETANs: 10 runs of 10-folds cross-validation for four class variables. The variables most connected to the class is the air temperature (connected in 73% out of the 400 ETANs), followed by ozone concentration (61%), season (45%), NO_2 (24 %). The remaining features are connected less frequently.

The structure of the multilabel classifier is shown in Fig.4. Again there are direct links between the class variable referring to the same day.

We report in Table 2 the performance indicators for the independent approach (an independent ETAN for each class variable) and the METAN classifier. As

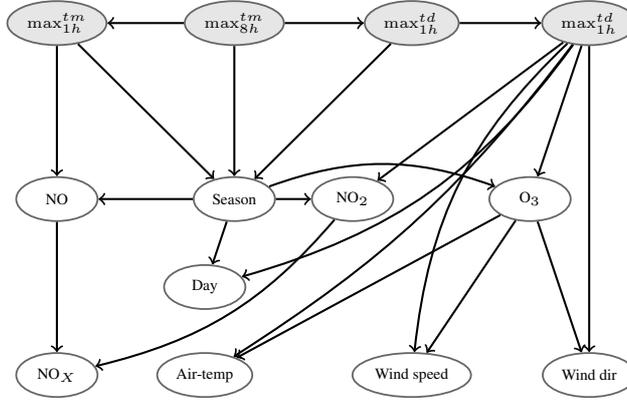


Figure 4: The METAN structure learned in the Berlin case study. By \max_{1h}^{td} and \max_{1h}^{tm} we denote the \max_{1h} indicator for today and tomorrow; the same notation applies to the \max_{8h} indicator.

we have here only four class variables, performing the significance test in the same way of the previous section would be not meaningful (each sample analyzed by the test would contain only four measures). We thus address the statistical comparison of the two classifiers in a different way. On each indicator and on each class variable we compare the performance of the two models analyzing the 100 results yielded by the 10 runs of 10-folds cross-validation through the correlated t -test, which is designed for analyzing cross-validation results (Witten and Frank, 2005). Statistically significant differences are boldfaced in Table 2.

The multilabel classifier model significantly improves some indicators such as auc and success index. It is significantly worse than the independent approach on no indicator.

	\max_{1h}^{td}	\max_{8h}^{td}	\max_{1h}^{tm}	\max_{1h}^{tm}
Accuracy	0.94 / 0.94	0.87 / 0.87	0.94 / 0.95	0.86 / 0.86
auc	0.85 / 0.71	0.85 / 0.85	0.73 / 0.68	0.75 / 0.50
c_5	0.24 / 0.25	0.49 / 0.50	0.25 / 0.25	0.67 / 0.67
c_{10}	0.49 / 0.49	1.06 / 1.12	0.51 / 0.51	1.35 / 1.35
success index	0.13 / 0.00	0.50 / 0.50	0.34 / 0.06	0.26 / 0.22

Table 2: Performance of the multilabel classifier and of the independent approach on the Berlin case study. Boldfaced entries indicate significant differences according to the correlated t -test.

6.1. Case study: predicting ozone in Burgas

We also re-analyze the Burgas dataset already analyzed by (Petelin et al., 2013). This is another case study regarding ozone and thus we remove the period October-March. The data set eventually contains 208 daily recordings. Again we consider the \max_{1h} and the \max_{8h} , and we set the threshold for with the goal of having 5% exceedances for both of them. This results in the threshold of $120 \mu\text{g}/\text{m}^3$ for \max_{1h} , and $100 \mu\text{g}/\text{m}^3$ for \max_{8h} .

The features are yesterday’s values of: air temperature, air humidity, atmospheric pressure, solar radiation, wind speed, NO_2 , SO_2 , phenol (carbolic acid), benzene, ozone. We introduce the day of the week and the season as further features, for a total of 12 features. We consider four class variables as in the previous case study: \max_{1h} and \max_{8h} for today and tomorrow. We thus learn 400 ETANs.

The variables most connected to the class are air temperature (69% out of the 400 learned ETANs), air radiation (65 %), ozone (53 %), air pressure (39 %), wind speed (22 %). All the others features are connected less frequently to the class.



Figure 5: The forest between the four variables of the Burgas case study. By \max_{1h} we denote the \max_{1h} indicator and by \max_{8h} we denote the \max_{8h} indicator; td and tm indicate today and tomorrow.

The forest linking the class variables is shown in Fig. 5. It shows a strong relationship between the \max_{1h} and the \max_{8h} on the same day and a weaker relationship between the \max_{1h} of two consecutive days.

	\max_{1h}^{td}	\max_{8h}^{td}	\max_{1h}^{tm}	\max_{8h}^{tm}
Accuracy	0.85 / 0.83	0.82 / 0.77	0.79 / 0.73	0.80 / 0.77
auc	0.71 / 0.68	0.59 / 0.50	0.64 / 0.53	0.55 / 0.50
c_5	1.11 / 1.52	1.32 / 2.17	1.99 / 2.37	1.72 / 3.10
c_{10}	3.06 / 3.45	3.14 / 5.41	4.61 / 4.99	5.20 / 6.62
success index	0.60 / 0.57	0.46 / 0.18	0.50 / 0.35	0.35 / 0.11

Table 3: Indicators of performance of the multilabel classifier and of the independent approach.

Following the same approach of the previous section, we report in Table 3 the performance indicators for the independent and the correlated model. The

multilabel classifier improves on most indicators compared to the independent approach. Overall the analysis supports the same claims of the previous case studies: the correlated models estimates posterior probabilities which are consistently more accurate than those of the independent model.

7. Conclusions

We have applied multilabel classification to the problem of predicting multiple air pollution variables. It delivers more accurate predictions than the independent approach, as shown by experiments involving both $PM_{2.5}$ and ozone. The multilabel classifier computes more accurate posterior probabilities which better support the decision maker.

Novel experiments can be designed in which for instance the multilabel classifier is used to jointly predict multiple pollutants whose concentrations are thought to be stochastically dependent.

Multilabel classification could be applied also in many other areas of environmental modelling in which there is the need for predicting multiple dependent discrete variables, without being limited to the analysis of air pollution problems.

8. Acknowledgments

Works partially funded by the Swiss NSF grant n. 200021_146606 / 1. We thank Uwe Schlink for valuable discussions and for providing us with the data of the Berlin case study. We thank Alexandra Grancharova for providing us with the data of the Bourgas case study, obtained from the Executive Environment Agency of Bulgaria.

References

- U. Schlink, S. Dorling, E. Pelikan, G. Nunnari, G. Cawley, H. Junninen, A. Greig, R. Foxall, K. Eben, T. Chatterton, J. Vondracek, M. Richter, M. Dostal, L. Bertuccio, M. Kolehmainen, M. Doyle, A rigorous inter-comparison of ground-level ozone predictions, *Atmospheric Environment* 37 (23) (2003) 3237 – 3253.
- P. Perez, Combined model for PM10 forecasting in a large city, *Atmospheric Environment* 60 (2012) 271 – 276.

- D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.
- C. P. de Campos, G. Corani, M. Scanagatta, M. Cuccu, M. Zaffalon, Learning extended tree augmented naive structures, *International Journal of Approximate Reasoning* 68 (2016) 153–163.
- N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, *Machine learning* 29 (2-3) (1997) 131–163.
- J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, *Machine learning* 85 (3) (2011) 333–359.
- G. De’Ath, Multivariate regression trees: a new technique for modeling species-environment relationships, *Ecology* 83 (4) (2002) 1105–1117.
- D. S. Chapman, B. V. Purse, Community versus single-species distribution models for British plants, *Journal of biogeography* 38 (8) (2011) 1524–1535.
- D. J. Hand, K. Yu, Idiot’s Bayes: not so stupid after all?, *International Statistical Review* 69 (3) (2001) 385–398.
- M. Scanagatta, C. P. de Campos, G. Corani, M. Zaffalon, Learning Bayesian Networks with Thousands of Variables, in: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett, R. Garnett (Eds.), *Proceedings of the Twenty-Ninth Annual Conference on Neural Information Processing Systems (NIPS 2015)*, 1855–1863, 2015.
- J. Cussens, Bayesian network learning with cutting planes, in: *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*., 153–160, 2011.
- C. X. Ling, J. Huang, H. Zhang, AUC: a statistically consistent and more discriminating measure than accuracy, in: *Proceedings of the 18th International Joint Conference on Artificial intelligence*, 519–524, 2003.
- A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern recognition* 30 (7) (1997) 1145–1159.
- S. R. Dorling, R. J. Foxall, D. P. Mandic, G. C. Cawley, Maximum likelihood cost functions for neural network models of air quality data, *Atmospheric Environment* 37 (24) (2003) 3435–3443.

- C. Elkan, The foundations of cost-sensitive learning, in: Proceedings of the 17th International Joint Conference on Artificial Intelligence, vol. 2, 973–978, 2001.
- I. H. Witten, E. Frank, Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, 2005.
- Y. Zheng, F. Liu, H.-P. Hsieh, U-Air: When urban air quality inference meets big data, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1436–1444, 2013.
- D. Petelin, A. Grancharova, J. Kocijan, Evolving Gaussian process models for prediction of ozone concentration in the air, Simulation modelling practice and theory 33 (2013) 68–80.