ISEIS

Journal of Environmental Informatics 17(1) 46-53 (2011)

Journal of Environmental Informatics

www.iseis.org/jei

Comparison of Decision Tree Algorithms for Predicting Potential Air Pollutant Emissions with Data Mining Models

D. Birant*

Department of Computer Engineering, Dokuz Eylul University, Tinaztepe Campus, Izmir 35100, Turkey

Received 3 May 2010; revised 28 December 2010; accepted 18 January 2011; published online 12 March 2011

ABSTRACT. Predicting air pollutant emissions from potential industrial installations is important for controlling air pollution and future planning of air quality management. This paper proposes the classification and prediction of the emission levels of industrial air pollutant sources using decision tree technique. It presents the comparison results of many decision tree algorithms (C4.5, CART, NBTree, BFTree, LADTree, REPTree, Random Tree, Random Forest, LMT, FT and Decision Stump) in terms of running time, classification accuracy and applicability. In comparison, six performance metrics were used: classification accuracy, precision, recall, f-measure, mean absolute error and mean squared error. The aim of the study is to determine the best classifier as a data mining model for the prediction of emission levels of the industrial plants as dependent variable from known values of independent variables: the physical region of the plant, the height of the plant, working hours, the height of the stack, the diameter of the stack, the velocity of the waste in the stack, the temperature of the waste in the stack, plume rise, source classification code, control equipment type and emissions method code. In the experimental studies, all these algorithms are applied on the dataset that consists of sulphur oxide emission levels of industrial pollutants in Izmir. According to the results, while C4.5 algorithm has the highest accuracy value, Decision Stump algorithm is the fastest one. The average classification accuracy found as 82.4% empirically shows the benefits of using decision tree technique in the classification and the prediction of emission levels.

Keywords: air pollution, data mining, classification and prediction, decision support systems, artificial intelligence

1. Introduction

Industry represents one of the main sources of emissions of Sulphur Oxide (SO_x) in many metropolitan areas. SO_x causes a wide variety of health and environmental impacts because of the way it reacts with other substances in the air. It causes local air pollution and contributes to the formation of acid rain. It can harm plants and corrode buildings and monuments. Additionally, at peak levels, SO_x can cause temporary breathing difficulty for people with asthma. Long-term exposure to high levels of SO_x can cause respiratory illness and aggravate existing heart disease (Pandey et al., 2005).

In order to prevent negative effects of SO_x, important emission sources and their contributions to air pollution at specific sites should be identified as a basis for developing air quality management strategies. Furthermore, it is also important to determine how much additional contribution will be in air with new/potential industrial pollutant sources. This future prediction can be done through development of a classification model by using current industrial pollutant sources and then the usage of this model for the estimation of the emission levels of new or potential industrial pollutant sources. Classification and prediction techniques are among the popular tasks in data mining. For classification and prediction, some intelligent system techniques have been used such as Bayesian, Artificial Neural Network (ANN), Decision Tree, Genetic Algorithm (GA), Support Vector Machine (SVM), Nearest Neighbor, and Fuzzy Logic. Example studies for each method are given below.

Bayesian (Liu et al., 2008; Fasbender et al., 2009) and feed-forward neural network (Corani, 2005) techniques were used for the prediction of the emission levels of several pollutants. A fuzzy relation model and a Gaussian dispersion model were integrated for air pollution control for industrial plants (Zhou et al., 2004). Similarly, fuzzy and genetic algorithm techniques were integrated for estimating unknown pollution values (Shad et al., 2009). In their study, fuzzy prediction technique was used to determine pollution concentration (PM_{10}), while genetic algorithm made easier to choose an optimum membership function. While the studies from Huang et al. in 2010 and Anastassopoulos et al. in 2008 proposed two different modeling systems for air pollution, the work presented in the paper (Lu and Wang, 2005) examined the feasibility of applying SVM to predict air pollutant levels. In the study (Gautam et al., 2008), a new scheme is proposed to predict chaotic time series of air pollutant concentrations using nearest neighbor searching. Differently from these previous studies, this paper proposes the

* Corresponding author. Tel.: +90 232 4127418; fax: +90 232 4127402.
E-mail address: derya@cs.deu.edu.tr (D. Birant).

ISSN: 1726-2135 print/1684-8799 online
 doi:10.3808/jei.201100186

usage of decision tree technique, instead of other classification techniques such as Bayesian, ANN, GA, SVM and Fuzzy Logic.

Air pollutant concentrations are related to several local characteristics and parameters. Since the relationship among the parameters is complex and strongly nonlinear, the usage of ANN technique is particularly suitable for modeling multifactor, uncertainty and nonlinearity (Zhou et al., 2007). Therefore, ANN has been applied to predict some well-known air pollutant concentrations such as NO_2 (Shakil et al., 2009), CO_2 (Ionescu and Candau, 2007), PM_{10} (Papanastasiou et al., 2007), O_3 (Elkamel et al., 2001), and SO_2 (Abdul-Wahab and Al-Alawi, 2008; Cortina-Januchs et al., 2009). In some studies, neural network was used to predict only one indicator and in other systems, it was used to predict more than one air pollution indicators (such as SO_2, PM_{10} and CO) at the same time (Kurt et al., 2008).

A number of empirical studies have been done in many different areas by using many different datasets for comparing classification techniques: Bayesian, ANN, Decision Tree, GA, SVM and Nearest Neighbor. In some cases, other methods perform better than decision tree (Chen et al., 2007), in some cases, decision tree is better than other methods (Endo et al., 2008). Generally, decision tree is one of the best classifiers when considering classification accuracy. In addition, decision tree has been proven in its ability of processing very large databases faster than other many techniques such as neural network and SVM techniques. Decision tree algorithm is accepted to be among the powerful classification algorithms in artificial intelligence and decision support systems. For these reasons, in this study, decision tree technique is selected for classification and prediction.

Differently from the previous studies, this study introduces to the use of decision tree technique for the prediction of the SO_x emission levels of industrial air pollutant sources. Furthermore, this is the first time that eleven decision tree algorithms in data mining are applied and compared for the classification and prediction of the potential air pollutant emissions. Differently from air pollutants concentration models, this study proposes an artificial intelligent model that has ability to learn by examples, like a human. Decision tree sometimes named as decision learning is among supervised learning techniques that are able to correctly classify unknown/new data at the end of the initial training period.

This paper proposes the usage of decision tree technique for classification and then for the prediction of the SO_x emissions of air pollutants from industrial facilities. The main tasks are: (i) building a model using train datasets, (ii) validating on test datasets, and (iii) using the model to predict the output value of the target function for any valid new data if the model accuracy is good enough such as 85%. In order to determine which decision tree algorithm builds the best model for environmental data, many decision tree algorithms (C4.5, CART, NBTree, BFTree, LADTree, REPTree, Random Tree, Random Forest, LMT, FT and DS (Decision Stump)) are compared in terms of running time, classification accuracy and applicability.

The aim of this study is to predict emission levels of new or potential industrial plants with predictive data mining models to contribute planning process of the new emission source places. By this study, it is also possible to predict unmeasured SO_x emission levels of current industrial plants. In order to achieve these purposes, many current emission sources' impurity state and their contribution to air should be well known. The dataset used in the experimental studies contains the amount of air pollutants emitted into the atmosphere from different industrial sources in Izmir metropolitan area. Experimental studies presented in this paper empirically demonstrate the benefits of using decision tree technique in the classification and the prediction of SO_x emission levels.

2. Decision Tree Technique

2.1. Classification and Prediction with Decision Tree

Decision tree is among the commonly used learning method for classification and prediction in data mining, artificial intelligence and decision support systems.

Definition 1. Let S be a training set of samples expressed in terms of k attributes from the set $A = \{A_1, A_2, ..., A_k\}$, and n classes from the set $C = \{C_1, C_2, ..., C_n\}$. Thus each sample $s \in S$ has $k + 1$ tuples:

$$s = <V_1, V_2, .., V_k; C_j>$$

where $V_i \in \text{Range}(A_i)$ is a value in the range of the attribute $A_i \in A$ and $C_j \in C$. A *decision tree* is a tree which is constructed using an algorithm that selects an attribute A_i and a subset of its values V_i to branch on.

Definition 2. A *cutpoint* is a threshold value, T, for the attribute A which is typically discretized during decision tree generation by partitioning its range into intervals by threshold value. For an attribute A, the *best cut point* T is selected from its range of values by evaluating every *candidate cut point* in the range of values.

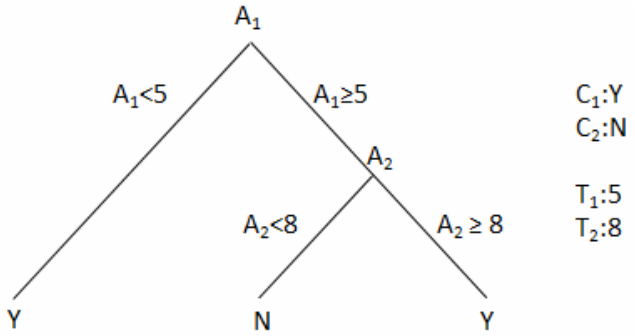

Figure 1. An example of decision tree.

Figure 1 shows an example decision tree in which internal nodes contains an attribute, each branch is an attribute value, and each leaf node specifies a class. In artificial intelligence and decision support systems, a decision tree can be constructed without complicated computations and is appropriate for exploratory knowledge discovery.

In order to determine the class label of a given new tuple $<V_1, V_2, \ldots, V_k>$, we traverse the tree top down, starting from

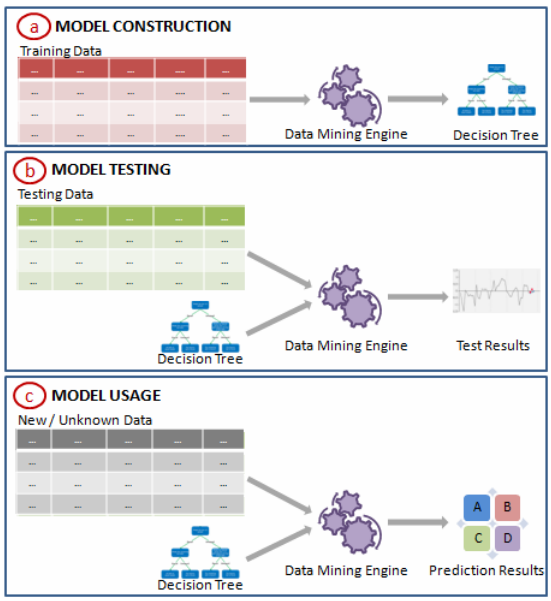

Figure 2. Classification and prediction processes: (a) model construction; (b) model testing; (c) model usage.

the root node *r* and visiting internal nodes *x* according to the attribute values of the given tuple until we reach a leaf node. The class of the leaf is the predicted class of the new tuple.

Decision tree has advantages such as it can produce a model which is simple to understand and represent interpretable rules; it is capable to deal with noisy data; and it can produce quite informative outputs. In artificial intelligence and decision support systems, this technique can be used for both continuous and categorical variables, but it is suitable for only predicting categorical outcomes.

Decision tree technique is among the popular tasks in data mining. It has been used successfully in many areas such as healthcare, finance, marketing, human resources, sport, telecommunications, and other fields. Thus, it is also a potentially useful approach for environmental studies, especially for the studies related to air pollution. For this reason, this study introduces to the investigation of the variables of air pollutants using decision tree technique.

Classification with decision tree has three main processes; the first (Figure 2a) is the learning process, where the training data is used to construct a model (classifier). The classifier is presented in the form of classification rules. In the second stage (Figure 2b), a test is done by using testing data to determine the classification accuracy of the model. If the accuracy is considered acceptable, the model can be applied to the classification of new data or unseen data. The third process (Figure 2c) is the use of the constructed (successful) model in classification to predict future data trends (García-Laencina et al., 2010).

2.2. Decision Tree Algorithms

The most well known decision tree algorithms are C4.5, CART and Naive Bayes Tree. These algorithms are greedy local search algorithms which construct trees top-down. This section presents high-level information about these algorithms and other tree based classification algorithms.

C4.5 (Quinlan, 1993) first grows an initial tree using the divide-and-conquer strategy and then prunes the tree to avoid overfitting problem. It calculates overall entropy and information gains of all attributes. The attribute with the highest information gain is chosen to make the decision. So, at each node of tree, C4.5 chooses one attribute that most effectively splits the training data into subsets with the *best cut point T* (Definition 2). According to the entropy and information gain, *Gain*(*S*, *A*), formulas, which are given in Equations (1) and (2), for the attribute *A* of the dataset *S*, where $freq(C_j, S)$ is the number of cases that belong to class C_j and $|S|$ is the number of cases in set *S* and S_i is a subset of the set *S*:

$$Entropy(S) = -\sum_{j=1}^{n} \frac{freq(C_j,S)}{|S|} \times \log_2 \frac{freq(C_j,S)}{|S|} \quad (1)$$

$$Gain(S,A) = Entropy(S) - \sum_{i \in A}^{k} \frac{|S_i|}{|S|} Entropy(S_i) \quad (2)$$

CART (Breiman et al., 1984) stands for Classification and Regression Trees. CART analysis is a form of binary recursive partitioning. The technique is aimed at finding a rule(s) which could predict the value of a dependent variable *Y* from known values of *n* explanatory variables X_i (predictors), where $i = 1, 2, 3, \ldots, n$. Initially, data contains a set of objects with known values of the dependent variable *Y* and predictors X_i. CART builds trees for recursive partitioning of all the objects into smaller subgroups by providing maximum homogeneity of the values of the dependent variable *Y*.

NBTree (Kohavi, 1996) is a hybrid technique between Naive Bayes and Decision Trees. It creates a decision tree with Naive Bayes classifiers at the leaves. After a tree is grown, a naive Bayes is constructed for each leaf using the data associated with that leaf.

BFTree (Shi, 2007) stands for Best-First decision tree learning. It expands nodes in best-first order instead of a fixed depth-first order. This classifier uses binary split for both categorical and numerical attributes.

LADTree (Holmes et al., 2002) builds multi-class alternating decision trees using logistic boosting strategy. At the each iteration of the algorithm, a single attribute test is chosen as the splitter node for the tree. The aim of the algorithm is to fit the working response to the mean value of the instances by minimizing the least-squares value between them.

REPTree (Witten and Frank, 2000) stands for Reduced Error Pruning Tree. It builds a decision tree using information gain as the splitting criterion. It also prunes the constructed tree using reduced-error pruning to correct the effects of noisy training examples and to reduce the complexity in the classification process. In order to provide optimization for speed, numeric attributes must be sorted once.

Random Tree (Fan et al., 2003) introduces different random elements to construct distributed decision trees. It consi-

ders randomly chosen attributes at each node. A chosen discrete feature on a decision path cannot be chosen again. Continuous feature can be chosen multiple times, however, with a different splitting value each time. During classification, the probabilities from each tree in the ensemble are averaged to produce the final prediction.

Random Forest (Breiman, 2001) is an ensemble of classification or regression trees, induced from bootstrap samples of the training data. It uses random feature selection strategy and grows many classification trees such that each tree depends on the values of a random vector sampled independently.

LMT (Landwehr et al., 2005) stands for Logistic Model Tree. LMT combines two popular techniques: decision tree induction and linear logistic regression. At each split, the logistic regressions of the parent node are passed to the child nodes and a leaf node accumulates all parent models to estimate a probability for each class. A pruning process is also applied to increase the generalization of the model.

FT (Gama, 2004) builds Functional Trees that could have logistic regression functions at the nodes. The algorithm can deal with binary, categorical, and numeric attributes and missing values.

DS (Decision Stump) is basically a single-level decision tree where the split at the root level is based on a specific attribute/value pair. It is usually used in conjunction with a boosting algorithm.

To the best of our knowledge, none of the previous studies uses these decision tree algorithms for the classification and prediction of the potential air pollutant emissions. In order to determine which decision tree algorithm builds the best model for environmental data, they are compared in terms of running time, classification accuracy and applicability.

3. Data Mining Application

There are two ways to determine the level of SO_x emission of an industrial installation in a particular region. First one is directly measurement using various technical equipments. Another method is the prediction using learning algorithms by training on a particular dataset. The aim of this application is the classification and then prediction of the SO_x emission levels of new/potential industrial air pollutant sources using decision tree technique.

3.1. Data Description

The dataset contains the measured amount of SO_x environmental impact factor emitted into the atmosphere from different industrial sources. It consists of some characteristics about 800 industrial installations in Izmir such as the *physical region* and *height* of the plant, *working hours*, the *height* and *diameter* of the stack, the *velocity* and *temperature* of the waste in the stack, *plume rise* which were calculated using Rupp's equation: plume_rise(Δh) = 1.5 $*$ V (effluent stack gas velocity) $*$ d (inside stack diameter) / μ (wind velocity as meteorological parameter), *source classification code* (a process-level code that describes the equipment and/or operation which is emitting pollutants),

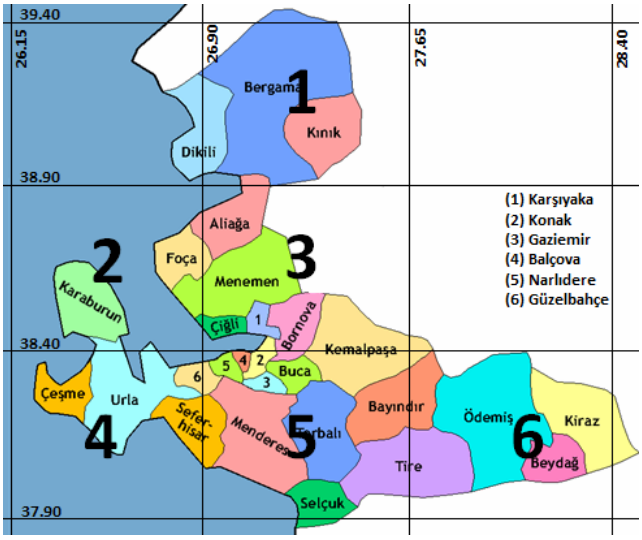

Figure 3. Izmir map separated by six regions.

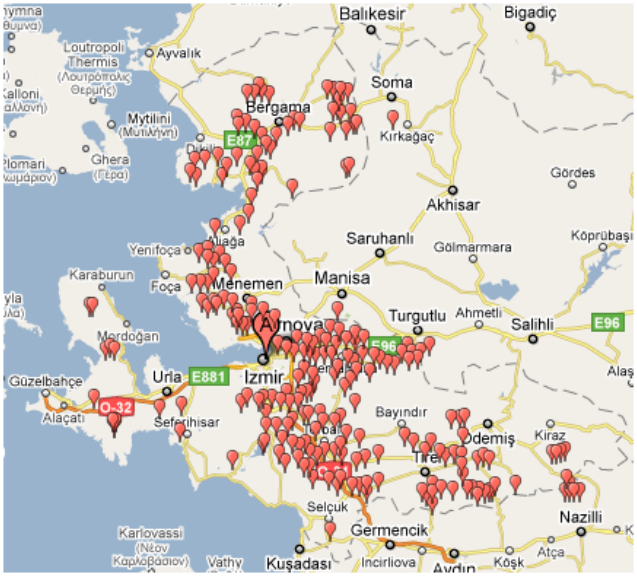

Figure 4. Izmir map with industrial installations as pushpins.

control equipment type used by a facility to regulate air emissions, *emissions method code* to identify the emission statement reporting purpose, and *SO_x emission*. Region column in the dataset contains six different regions of Izmir as shown in Figure 3. Working-hours column has values 8, 16 or 24 hours. The SO_x emission values in the dataset were acquired as 24-hour mean values and as a density (the concentration of SO_x) in µg per m^3. The target column SOx_Emission was categorized as *Low*: $0 < x < 51$, *Medium*: $50 < x < 151$, *High*: $150 < x < 501$, and *Very_High*: $500 < x$. When the application will be repeated for another city or country, instead of Izmir, these ranges may be changed according to the values in the dataset.

In this study, the six regions were determined according to the distribution of industrial installations on the map and also by considering natural environmental such as sea, lake, forest, desert, mountains or swamp. It is also possible to consider other

Table 1. Comparison of Different Decision Tree Algorithms Applied on SO_x Emission Levels Dataset

Algorithm	Tree Size	ACC (%)	MAE	MSE	PRE	REC	FME
C4.5 (J48)	53	86.25	0.1145	0.2430	0.812	0.858	0.834
DS	Single Level	85.00	0.1543	0.2658	0.722	0.850	0.781
LADTree	31	83.75	0.1225	0.2453	0.767	0.838	0.801
NBTree	3	82.50	0.1297	0.2501	0.803	0.825	0.814
CART	11	82.50	0.1311	0.2621	0.801	0.825	0.813
BFTree	17	82.50	0.1165	0.2728	0.801	0.825	0.813
FT	13	82.50	0.1081	0.2599	0.779	0.825	0.801
LMT	1	81.25	0.1319	0.2704	0.730	0.816	0.771
REPTree	33	81.25	0.1418	0.2826	0.731	0.814	0.770
Random Forest	10 Trees	80.00	0.1128	0.2787	0.747	0.807	0.776
Random Tree	438	78.75	0.1147	0.3215	0.766	0.789	0.777
Average		82.4	0.1253	0.2684	0.769	0.825	0.796

different conditions such as rain, wind, and thunderstorms. When the application will be repeated for another city or country, instead of Izmir, different region-specific decisions may be given. Figure 4 shows the physical locations of industrial installations. Each pushpin specifies the position of the industrial plants over Izmir map.

The major concern with this paper is the independent variables that are used to predict the levels of SO_x emissions. All these independent variables mentioned above (the region, the height of the plant, working hours, the height and diameter of the stack etc.) are evaluated using the formulas given in Definition 1 (entropy) and Definition 2 (information gain). These measures are used to select among the candidate variables at each step while growing the tree. These formulas score and rank each independent variable, and then select the "most informative "variable at each node in the tree. In other words, the meaningful and relevant variables are chosen automatically by the decision tree algorithm.

3.2. Training and Testing Processes

In the training process, each decision tree algorithm was applied on the 90% of the dataset with WEKA data mining toolkit and classifiers were constructed, in other words, decision rules were deduced for each algorithm. After this process, each classifier was test by using the remainder data (10% of the dataset). All the algorithms performed on a system with 2.4 GHz Core 2 Duo processor, 4 GB RAM and running on Windows Vista system. For all experiments, five trials were done and the average values were reported, since the results of the evolutionary process are somewhat sensitive to other processes.

In the testing process, six performance metrics were used to compare the algorithms (Ferri et al., 2009): classification accuracy (ACC), precision (PRE), recall (REC), F-Measure (FME), mean absolute error (MAE), and mean squared error (MSE). Given a classier and an instance, there are four possible outcomes. A *true positive* (TP) is a positive example ("i.e. high") correctly identified as a positive ("high"). A *false positive* (FP) is a negative example ("low") incorrectly identified as a positive ("high"). Also, a *true negative* (TN) is a negative example correctly identified as a negative. And a *false negative* (FN) is a positive example incorrectly identified as a negative. The "true" and "false" here can be interpreted as "correct" and "incorrect" respectively and the "positive" and "negative" can be interpreted as "labeled as positive" and "labeled as negative" respectively.

Classification accuracy (ACC) is one of the basic performance measures for classification algorithms. It is the ratio of the number of cases truly predicted by the classifier over the total number of cases in the test dataset, as formulated in Equation 3. Accuracy values can range between 0 and 100%. A perfect accuracy of 100% means that the predicted values are exactly the same as the observed ones:

$$ACC = \frac{\text{num(test examples correctly classified)}}{\text{num(total test examples)}} = \frac{TN + TP}{TN + FN + TP + FP} \quad (3)$$

Precision (PRE) is a measure of the accuracy provided that a specific class has been predicted, whereas *Recall* (REC) is a measure of the ability of a prediction model to select instances of a certain class from a data set. In other words, precision is the percentage of times that the classifier is correct in its classification of positive samples, while recall is the percentage of known positive samples that the classifier would classify as being positive. *F-Measure (FME)* that combines precision and recall is the harmonic mean of precision and recall. A perfect precision score of 1.0 for a class c_i means that every item labeled as belonging to class c_i does indeed belong to class c_i, whereas a perfect recall of 1.0 means that every item from class c_i was labeled as belonging to class c_i:

$$PRE = \frac{TP}{TP + FP} \qquad REC = \frac{TP}{TP + FN} \quad (4)$$

$$FME = \frac{2 * PRE * REC}{PRE + REC} \quad (5)$$

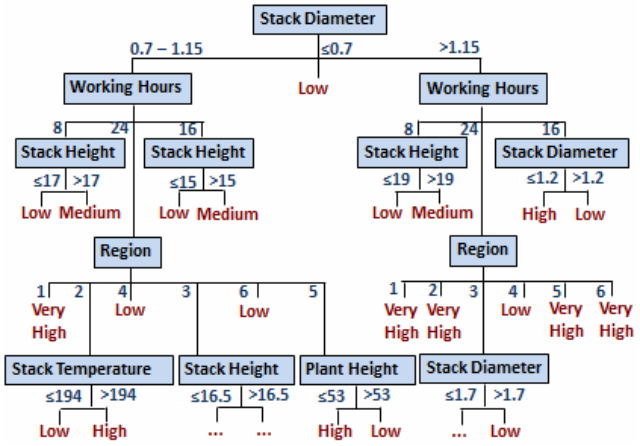

Figure 5. Decision tree constructed by C4.5 algorithm.

The *mean absolute error* (MAE) is the average of the absolute values of the prediction errors. MAE shows how much the predictions deviate from the true probability and it calculates the absolute value of the difference. Mean Squared Error (MSE) is just a quadratic version of MAE, which penalizes strong deviations from the true probability. For better classification and prediction, MSE and MAE values should be kept as low as possible, actually close to zero.

$$MAE = \frac{\sum_{j=1}^{c}\sum_{i=1}^{m}\left|o(i,j) - p(i,j)\right|}{m \times c} \quad (6a)$$

$$MSE = \frac{\sum_{j=1}^{c}\sum_{i=1}^{m}\left(o(i,j) - p(i,j)\right)^2}{m \times c} \quad (6b)$$

where m denotes the number of examples, c is the number of classes, $o(i, j)$ and $p(i, j)$ represent observed and predicted values of the example i to be of class j.

4. Results and Discussion

Table 1 shows the comparison of various decision tree algorithms which were applied over SO_x emissions dataset. The average classification accuracy (ACC) is 82.4%. Classification accuracy results are near to each other, but the difference among them can increase or the order can change over bigger datasets. C4.5 decision tree algorithm has a little more accuracy value when SO_x emission dataset with 800 instances are applied. Accuracy ratio of this classification algorithm is about 86.25 %. So, 69 of 80 testing instances were classified correctly. This result shows that SO_x emission levels of Industrial installations can be classified with an acceptable ratio and new/unknown SO_x emission levels of potential industrial plants can be predicted successfully with 86.25 percentages.

Precision (PRE) values found in the experimental study are close to the perfect precision score which is 1. For example, the precision value was found as 0.812 when C4.5 algorithm was used. This means that most of the items predicted as belonging to class c_i do indeed belong to class c_i. Recall (REC) values are also close to the perfect recall score which is also 1. This means that most of the items from class c_i were predicted as belonging to class c_i.

According to the results, mean absolute error (MAE) and mean squared error (MSE) values are close to 0. This means that the difference between observed and predicted values is small. When we consider these measures, Functional Trees (FT) algorithm produces the best result with 0.1081 and 0.2599 values for MAE and MSE respectively, then next Random Forest algorithm with 0.1128 for MAE value.

According to the constructed decision trees; the diameter and height of stack, the region of the industrial installation and working hours are among most effective attributes to determine SO_x emission levels. The similarity between predictions and real-world measurements was better for the region 4 when compared with the other regions. The main reason of this result is that this region fully has low emission levels.

According to the size of the trees, each decision tree algorithm builds different trees with different sizes. DS algorithm builds single-level decision tree and its classification accuracy is higher than the most of the other algorithms. However, Random Tree algorithm builds a very large tree and its classification accuracy is the lowest.

When decision tree algorithms were compared for each SO_x emission level (*low*, *medium*, *high*, and *very high*) individually, C4.5 obtained the best results for *low* level with 91.2% accuracy. The reason of this result is that training set contains more samples related to *low* level category and so the algorithm can construct better trees in terms of this category by considering more cases when calculating the entropy and information gain values.

The overall classification accuracy performance of decision tree algorithms was found good with an average accuracy of about 82.4%. So, we can say that decision tree learning has a good ability to predict SO_x emission levels and can be used for estimating unknown pollution values. Thus, we can also say that experiments, which were carried out using the environmental dataset collected from the region Izmir, empirically demonstrate the benefits of using decision tree technique in the classification and prediction of the potential air pollutant emissions.

Figure 5 shows a part of the decision tree constructed by C4.5 algorithm which has the highest accuracy value in Table 1. According to the C4.5 tree, stack diameter is the first attribute when predicting SO_x emission levels, then, working-hours at the second level follows it. Stack height, the region and the height of the industrial installation and stack temperature start to appear after the second level of the tree. For instance, tree shows that the installations, with larger than 1.15 stack diameter, which work all day in the south-western part (bottom left, region 4) of the city are generally classified as *low* SO_x emission level. However, the similar installations in the south-eastern part (bottom right, region 6) of the city are generally classified as *very_high* SO_x emission level. It is also possible to generate the decision rules by traversing all the paths from the root to the leaf node in the decision tree. For instance the following decision rules can be generated:

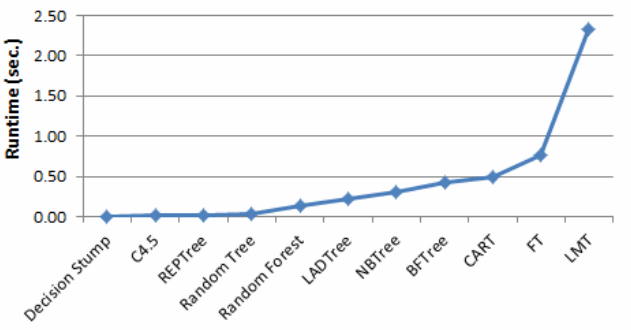

Figure 6. Runtime comparison of decision tree algorithms.

R1: If Stack_Diameter ≤ 0.7, then SOx_Level=low

R2: If 0.7 < Stack_Diameter ≤ 1.15 and Working_Hours = 8 and Stack_Height ≤ 17, then SOx_Level = low

R3: If 0.7 < Stack_Diameter ≤ 1.15 and Working_Hours = 8 and Stack_Height > 17, then SOx_Level = medium

R4: If 0.7 < Stack_Diameter ≤ 1.15 and Working_Hours = 24 and Region = 1, then SOx_Level = very high

R5: If 0.7 < Stack_Diameter ≤ 1.15 and Working_Hours = 24 and Region = 2 and Stack_Temperature > 194, then SOx_Level = high

Graph in Figure 6 shows the comparison of decision tree algorithms in terms of runtime. While Decision Stump, C4.5 and REPTree algorithms are faster than others, LMT algorithm takes very long time even with small dataset.

Table 2 shows the results obtained by using Decision Stump (DS) algorithm which is the fastest one among other decision tree algorithms. A DS is always a single-level decision tree and consists of a single node (the root) which is immediately connected to the terminal nodes. For this reason, DS tree in Table 2 is limited with only stack height attribute. Because some unknown stack height values exist in the dataset, these values are named as "missing values" and generally classified as very high. The figure also shows class distributions to be able to evaluate the data covered or not covered by the rule.

Table 2. The Results Obtained by Decision Stump Algorithm

Decision Stump	Class Distributions			
	Low	Medium	High	Very High
Stack_height <= 12.5 : Low	0.9071	0.0601	0.0273	0.0054
Stack_height > 12.5 and <= 32 : Medium	0.1958	0.4742	0.1340	0.1958
Stack_height > 32 : High	0.0926	0.2539	0.4852	0.1682
Stack_height is missing : Very High	0.0187	0.0305	0.0328	0.9178
Average	0.3036	0.2047	0.1698	0.3218

Applications based on the same decision tree algorithms have also been proposed for different areas such as for astronomy (Zhao and Zhang, 2008), for medical diagnosis (Daud and Corne, 2010) and for flood/standing water detection (Sun et al., 2010). The results of the astronomical study shows that ADTree is the best only in terms of accuracy, Decision Stump is the best only considering speed, J48 (which is an implementation of C4.5) is the optimal choice considering both accuracy and speed. In the medical study, 10 medical datasets were used to compare algorithms and the average values indicate that ADTree is the most successful algorithm according to the classification accuracy. In the water identification study, J48 (C4.5) is the best one, like this study, in terms of classification accuracy. So, C4.5 can be accepted to be among the powerful decision tree algorithms in artificial intelligence and decision support systems. According to all comparison results obtained from different datasets and obtained from different application areas, the best decision tree algorithm may change. Therefore, different decision tree algorithms should be applied and compared on each case study to determine the best model.

5. Conclusions

This paper proposes the usage of decision tree technique for classification and prediction of the emissions levels of environmental impact factors from industrial installations. Since dense and irregular industrialization is one of the leading causes of air pollution, this study is important to be able to estimate the waste gas emission levels of new/potential pollutant sources.

In order to determine the best classifier for the prediction of waste gas emission levels, existing decision tree algorithms are compared in terms of running time, classification accuracy and applicability. Experimental studies were carried out with measured waste gas emission data collected from industrial installations in the six different regions of Izmir. The experimental results show that SO_x emission levels of industrial installations can be classified and predicted successfully with an acceptable (averagely) 82.4 ratio. According to the decision trees constructed in experiments, by the order of importance; stack diameter, working hours, stack height, stack temperature and the height of the industrial installation are among most effective attributes to predict SO_x emission levels.

In future, the prediction of waste gas emission of pollutant sources can help and improve future planning in air quality management. Predicted emission levels can be used for air quality prediction with meteorological data.

Acknowledgments. This study is applied over dataset of waste gas emission levels of industrial installations in Izmir which is provided by the Department of Environmental Engineering at Dokuz Eylul University.

References

Abdul-Wahab, S.A., and Al-Alawi, S.M. (2008). Prediction of sulfur dioxide (SO_2) concentration levels from the Mina Al-Fahal Refinery in Oman using artificial neural networks. *Am. J. Environ. Sci.*, 4(5), 473-481.

Anastassopoulos, A., Nguyen, S., and Xu, X. (2008). An assessment

of meteorological effects on air quality in Windsor, Ontario, Canada - Sensitivity to temporal modeling resolution, *J. Env. Inform.*, 11(2), 45-50. doi:10.3808/jei.200800110

Breiman, L. (2001). Random forests. *Mach. Learning*, 45(1), 5-32. doi: 10.1023/A:1010933404324

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA. (Since 1993 this book has been published by Chapman and Hall, New York.)

Chen, J., Xing, Y., Xi, G., Chen, J., Yi, J., Zhao, D., and Wang, J. (2007). A comparison of four data mining models: bayes, neural network, SVM and decision trees in identifying syndromes in coronary heart disease. ISNN 2007, *Lecture Notes Elect. Eng.*, 4491, 1274-1279. doi: 10.1007/978-3-540-72383-7_148

Corani, G. (2005). Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. *Ecol. Model.*, 185, 513-529. doi:10.1016/j.ecolmodel.2005.01.008

Cortina-Januchs, M.G., Barron-Adame, J.M., Vega-Corona, A., and Andina, D. (2009). Prevision of industrial SO_2 pollutant concentration applying ANNs. *Proc. of 7th IEEE International Conference on Industrial Informatics (INDIN 2009)*, Wales, UK, 510-515.

Daud, N.R., and Corne, D.W. (2009). Human Readable Rule Induction in Medical Data Mining, *Lecture Notes Elect. Eng.*, 27(7), 787-798. doi: 10.1007/978-0-387-84814-3_79

Elkamel, A., Abdul-Wahab, S., Bouhamra, W., and Alper, E. (2001). Measurement and prediction of ozone levels around a heavily industrialized area: a neural network approach. *Adv. Environ. Res.*, 5(1), 47-59. doi:10.1016/S1093-0191(00)00042-3

Endo, A., Shibata, T., and Tanaka, H. (2008). Comparison of seven algorithms to predict breast cancer survival. *Biomedical Soft Comput. Hum. Sci.*, 13(2), 11-16.

Fan, W., Wang, H., Yu, P.S., Ma, S. (2003). Is random model better? On its accuracy and efficiency. Proceedings of Third *IEEE International Conference on Data Mining (ICDM'03)*, 51. doi:10.1109/ICDM.2003.1250902

Fasbender, D., Brasseur, O., and Bogaert, P. (2009). Bayesian data fusion for space-time prediction of air pollutants: The case of NO_2 in Belgium. *Atmos. Environ.,* 43, 4632-4645. doi:10.1016/j.atmosenv. 2009.05.036

Ferri, C., Hernández-Orallo, J., and Modroiu, R. (2009). An experimental comparison of performance measures for classifycation. *Pattern Recognit. Lett.*, 30(1), 27-38. doi:10.1016/j. patrec.2008.08.010

Gama, J. (2004). Functional trees. *Mach. Learning*, 55(3), 219-250. doi:10.1023/B:MACH.0000027782.67192.13

García-Laencina, P.J., Sancho-Gómez, J-L., and Figueiras-Vidal, A.R. (2010). Pattern classification with missing data: a review. *Neural Comput. Appl.*, 19(2), 263-282. doi:10.1007/s00521-009-0295-6

Gautam, A.K., Chelani, A.B., Jain, V.K., and Devotta, S. (2008). A new scheme to predict chaotic time series of air pollutant concentrations using artificial neural network and nearest neighbor searching. *Atmos. Environ.*, 42 (18), 4409-4417.doi:10.1016/j.atmosenv. 2008.01.005

Holmes, G., Pfahringer, B., Kirkby, R., Frank, E., and Hall M. (2002). Multiclass alternating decision trees. in: Elomaa, T., Mannila, H., Toivonen, H. (Eds.), ECML 2002, *Lecture Notes Comput. Sci.*, 2430, 161-172. doi:10.1007/3-540-36755-1_14

Huang, Q., Cheng, S.Y., Li, Y.P., Li, J.B., Chen, D.S., and Wang, H.Y. (2010). An integrated MM5-CAMx modeling approach for assessing PM10 contribution from different sources in Beijing, China, *J. Env. Inform.*, 15(2), 47-61. doi:10.3808/jei.201000166

Ionescu, A., and Candau, Y. (2007). Air pollutant emissions prediction by process modelling - Application in the iron and steel industry in the case of a re-heating furnace. *Environ. Model. Software*, 22(9), 1362-1371. doi:10.1016/j.envsoft.2006.09.008

Kohavi, R. (1996). Scaling up the accuracy of naive-bayes classifiers: a decisiontree hybrid. *Proc. of the Second International conference on Knowledge Discovery and Data Mining*, Portland, Oregon, 202-207.

Kurt, A., Gulbagci, B., Karaca, F., and Alagha, O. (2008). An online air pollution forecasting system using neural networks. *Environ. Int.*, 34(5), 592-598. doi:10.1016/j.envint.2007.12.020

Landwehr, N., Hall, M., Frank, E. (2005). Logistic model trees. *Mach. Learning*, 59(1-2), 161-205. doi:10.1007/s10994-005-0466-3

Liua, Y., Guoa, H., Maob, G., and Yanga, P. (2008). A Bayesian hierarchical model for urban air quality prediction under uncertainty. *Atmos. Environ.,* 42(36), 8464-8469. doi:10.1016/j.atmosenv.2008.08.018

Lu, W.Z., and Wang, W.J. (2005). Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends. *Chemosphere*, 59(5), 693-701. doi:10.1016/j. chemosphere. 2004.10.032

Pandey, J.S., Kumar, R., and Devotta, S. (2005). Health risks of NO_2, SPM and SO_2 in Delhi (India). *Atmos. Environ.*, 39(36), 6868-6874. doi:10.1016/j.atmosenv.2005.08.004

Papanastasiou, D.K., Melas D., and Kioutsioukis, I. (2007). Development and assessment of neural network and multiple regression models in order to predict PM10 levels in a medium-sized Mediterranean city, *Water, Air, Soil Pollut.*, 182(1-4), 325-334. doi:10.1007/s11270-007-9341-0

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Shad, R., Mesgari, M.S., Abkar, A., and Shad, A. (2009). Predicting air pollution using fuzzy genetic linear membership kriging in GIS. *Comput., Environ. Urban Syst.*, 33(6), 472-481. doi:10.1016/j.compenvurbsys.2009.10.004

Shakil, M., Elshafei, M., Habib, M.A., and Maleki, F.A. (2009). Soft sensor for NO_x and O_2 using dynamic neural networks. *Comput. Electrical Eng.*, 35(4), 578-586. doi:10.1016/j.compeleceng. 008.08 .007

Shi, H. (2007). *Best-first decision tree learning*. MSc Thesis, University of Waikato, Hamilton, NZ.

Sun, D., Zhang, R., and Yu, B. (2010). *Flood and Standing Water Detection*, 2010 AWG Annual Meeting, June 7-11, 2010, Madison, WI.

Witten, I., and Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Mateo, CA.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z-H., Steinbach, M., Hand, D.J., and Steinberg, D. (2008). Top 10 algorithms in data mining, *Knowl. Inf. Syst.*, 14(1), 1-37. doi:10.1007/s10115-007-0114-2

Zhao, Y., and Zhang, Y. (2008). Comparison of decision tree methods for finding active objects, *Adv. Space Res.*, 41(12), 1955-1959. doi:10.1016/j.asr.2007.07.020

Zhou, J., Er M.J., and Zurada, J.M. (2007). Adaptive neural network control of uncertain nonlinear systems with nonsmooth actuator nonlinearities. *Neurocomputing*, 70(4-6), 1062-1070. doi:10.1016/j.neucom.2006.09.009

Zhou, Q., Huang, G.H., and Chan, C.W. (2004). Development of an intelligent decision support system for air pollution control at coal-fired power plants. *Expert Syst. Appl.*, 26(3), 335-356. doi:10.1016/j.eswa.2003.09.005