The 2014 Biomedical Engineering International Conference (BMEiCON-2014)

Air Quality Classification in Thailand Based on Decision Tree

Kattariya Kujaroentavon, Supaporn Kiattisin, Adisorn Leelasantitham and Sotarat Thammaboosadee
Information Technology Management Program Faculty of Engineering, Mahidol University
25/25 Phutthamonthon 4Rd., Salaya Nakhon Pathom 73170, Thailand
E-mail: kat_aretee@hotmail.com, supaporn.kit@mahidol.ac.th, adisorn.lee@mahidol.ac.th, zotarat@gmail.com

***Abstract*—The paper presents a model for management classifier air quality by algorithm of decision tree using air quality index in Thailand including a pollutant's concentration e.g. O_3, NO_2, CO, SO_2, PM_{10} and levels of healthy concern. The purpose of this research is to establish rules of separated air quality classification by levels of healthy concern. The results of this study are correctly classified into instances of training set of 96.80% and testing set of 91.07%. The ROC curve shows that the training set data and testing set data are similar to such results. The algorithm of decision tree can use to become rules of separated air quality classification by levels of healthy concern.**

Keywords—air quality, Model, Classification, Levels of Healthy Concern, Decision Tree, air quality, Model, Classification, Levels of Healthy Concern, Decision Tree

I. INTRODUCTION

Air Pollution is main problem of people will met for affect health and respiratory. Almost this problem happened in downtown. People smell bad atmosphere and many dust into lungs. From statistic of respirator's patients. In 2007, patients 242,405 up to became 305,929 in 2008. In 2009, patients 363,744 up to became 365,372 in 2010. Finally In 2011 up to 381,184 following Fig. 1.

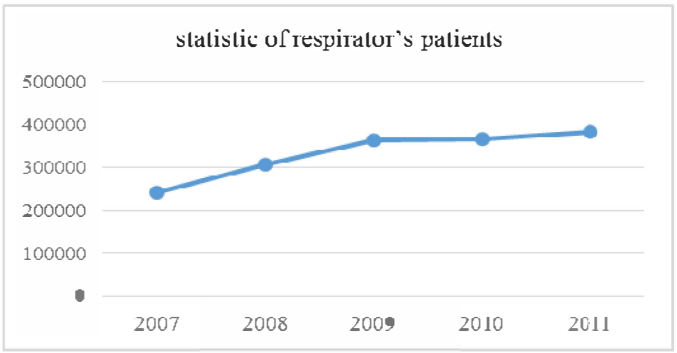

Fig. 1. Show statistic's patients

The results of statistic numbers of respirators patients were up very years. Although respirators patients someone they relative with air pollution from traffic problem which happened by directly and indirectly following Table 1. All of them from a pollutant's concentration of dusts less than 10 micron nitrogen dioxide (NO_2) carbon dioxide (CO) sulfur dioxide (SO_2) and Ozone left out to atmosphere effected to health of the people with directly [1].

Table 1 Show levels Air quality health impacts [2]

Air Quality Index	**Protect Your Health**
Good	No health impacts are expected when air quality is in this range.
Moderate	Unusually sensitive people should consider limiting prolonged outdoor exertion.
Unhealthy for Sensitive Groups	The following groups should limit prolonged outdoor exertion - People with lung disease, such as asthma - Children and older adults - People who are active outdoors
Unhealthy	The following groups should avoid prolonged outdoor exertion: - People with lung disease, such as asthma - Children and older adults - People who are active outdoors Everyone else should limit prolonged outdoor exertion.
Very Unhealthy	The following groups should avoid all outdoor exertion: - People with lung disease, such as asthma - Children and older adults - People who are active outdoors Everyone else should limit outdoor exertion.

The first air quality index, name the "Pollutant Standard Index" (PSI), was developed and introduced by United States Environmental Protection Agency, taking into consideration five major (criteria) air pollutants, namely, CO, SO_2, PM_{10}, O_3, and NO_2. In 1999, the index was further completed and replaced by the Air Quality Index or AQI. The most widely used index for air quality assessment and management. $PM_{2.5}$ and 8-hr average ozone [3].

Nowadays the paper about develop air quality index by used a classification is an essential technique of data mining. Such as used fuzzy inference system to separated air quality classification by used pollutant's concentration by added concentration of pollutants benzene, toluene, ethyl benzene, xylene, and 1, 3 - butadiene standards for air quality classification [4]. Used neural network Model by classification technique to forecast air quality for reduce pollution problem which population can prepare with population effect before [5] and use classification technique to make model Decision Tree. To assignment results of

978-1-4799-6801-5/14/$31.00 ©2014 IEEE

concentration of pollutants which influenced for healthy of population [6]. Used a decision tree to forecast daily dissolved oxygen rates in a lagoon along the French Mediterranean sea coast [7]. Including used a decision tree identifying controlling factors of ground-level ozone levels over southwestern Taiwan [8].

From this passage. Researcher was introduced rules of separated air quality classification which influenced for healthy. To support decision for separated air quality classification. By combined the information about concentration of pollutants.This paper introduced rule of separated air quality classification by level of healthy concern and used decision tree which technique of classification and can use the results of them to analysis factors is caused to happened the pollution problem more standard with directly.

II. METHODOLOGY

Aim of this paper is use data mining to create model by using decision tree with classifier technique. This paper separate step for use data mining in 5 steps with the following Fig. 2.

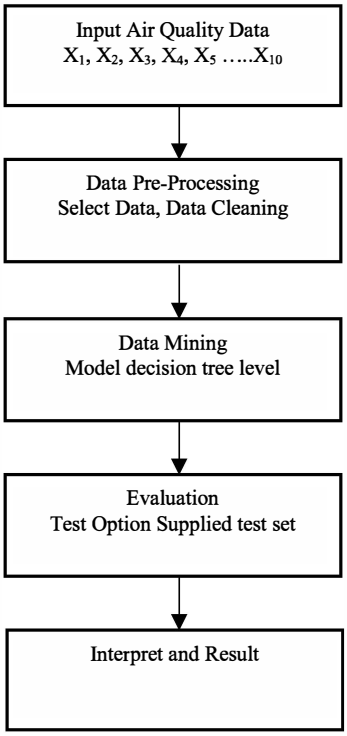

Fig. 2. Data mining in 5 steps

A. Input Air Quality Data

First, combined information about factors which influenced for levels of air Quality such as PM_{10}, PM_{25}, So_2 (1 hour), So_2 (24 hour) etc. Collect data about concentration of pollutants in each kinds in Thailand for 2012-2013.

B. Data Pre-Processing

Using the data to Pre-processing before process chose interested Attribute and repeated data out. Included missing value data noisy data and inconsistent data. After clean the data, adapted data for using in data mining step.

Form concentration of pollution. The result of air quality's pollution control department. The pollution in 2012-2013, Thailand has concentration of pollutants separated level which influenced for healthy in 4 levels are good, moderate, unhealthy for sensitive groups and unhealthy. From the standard of air quality classification in Thailand have 6 levels. In Table 2. By air quality index from 0-100 is an air quality in normal atmosphere. If air quality index more 100 is show that concentration of pollutants has over standard.

In this paper used a concentration of pollutants have kinds including Ozone NO_2 CO SO_2 PM_{10}. For created rule to separated levels of Healthy Concern Ozone classification in Thailand.

Table 2 Levels of Healthy Concern [9]

Air Quality Index (AQI) Values	**Levels of Healthy Concern**
0 to 50	Good
51 – 100	Moderate
101 – 150	Unhealthy for Sensitive Groups
151 – 200	Unhealthy
201 – 300	Very Unhealthy
301 to 500	Hazardous

Table 3 The Concerntration of pollutants [10]

Attribute	**Attribute Name**	**Average (hour)**	**Descriptions**
1	SO_2	24	Sulfur dioxide
2	NO_2	1	Nitrogen dioxide
3	CO	8	Carbon dioxide
4	Pm_{10}	24	Dust less than 10 micron
5	Ozone	1	Ozone Average
6	Level	-	Levels of Healthy Concern

C. Data Mining

A decision Tree is decision method. It consists of a root, nodes, branches and leafs (terminals) which the results will happened when the situation started, it shows in decision form and divided in each ways to decision. The Following Fig. 3.

Decision tree model started to separate air quality classification of decision from "root" calculated information gain to used attribute in each nodes of tree attribute. Anyone has most the information gain result or less Entropy result will be attribute of node. And remaining data will calculate information gain again. Using the following formula

Entropy equation

$$\text{Entropy}(s) = \sum_{i=1}^{e} -p_1 \log_2 p_1 \quad (1)$$

By S is attribute to be measured.
P_1 is ratio of members in groups to the number all members of sample

Information Gain

$$\text{GAIN}(S, A) = \text{Entropy}(S) - \sum_{\text{value}(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (2)$$

By A is attribute A
S_v is members of attribute V valuable
S is number of samples

Split Information

$$\text{Split Information}(S, A) = -\sum_{i=1}^{n} \frac{|S_1|}{|S|} \log_2 \frac{|S_1|}{|S|} \quad (3)$$

Gain Ratio

$$\text{GAIN RATIO}(S, A) = \frac{\text{Gain}(S, A)}{\text{Split Information}(S, A)} \quad (4)$$

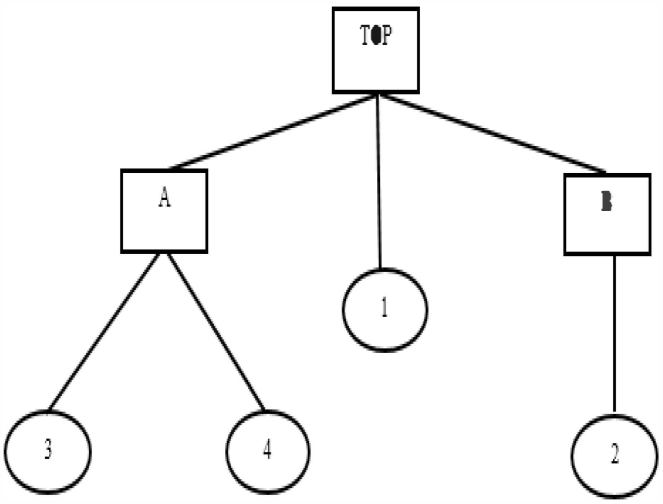

Fig. 4. Decision tree model

D. Evaluation, Interpret and Result

Evaluation interpret and result are data which processed by using attribute the following table 3. In data mining, research chose to use decision tree technique for create air quality classification model. For separate levels of air quality which influence for healthy. Using examined data in test option supplied test set. Divide in a sets first set is training set 70% and test set 30% for testing model's quality. Last step is compare efficient model in ROC curve form compare results between ROC curve form on training set and test set.

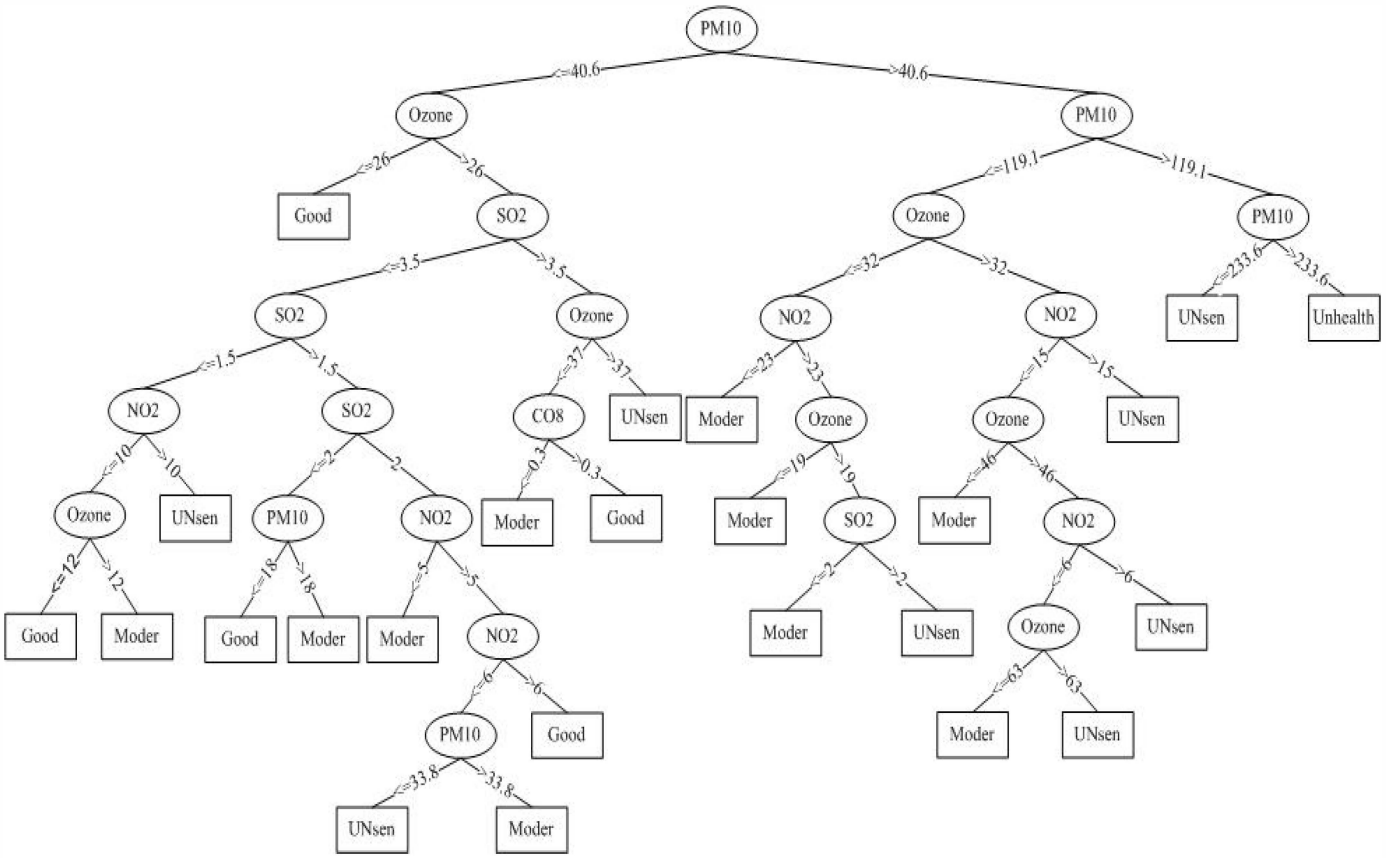

Fig. 3. Shows the tree from the classification of the Air Quality Index to Levels of Health Concern with the decision tree.

III. RESULT AND DISCUSSION

In classification, separate data in levels which influence for healthy include Good, Moderate, Unhealthy for sensitive groups and Unhealthy by using algorithm decision tree following Fig. 4. Used evaluation test option supplied test set which divided air quality data in 2 sets are 70% for training set and 30% for test set.

Result of correctly classified instances 's training set can predict data with correctly 96.8% has Incorrectly Classified Instances 3.55% and result correctly classified instances of test set is 91.07% incorrectly classified instances8.93 % following Table 4.

Table 4 show result of correctly classified Training set data and Test set data

Data	**correctly Instances**
Training Set	96.8%
Test Set	91.07%

From result of training set and test set can create receiver aerator characteristic or ROC curve to make relative graph between true positive rate with false positive rate by cut – off point Following Fig. 5.

Compared efficient for process result's algorithm between training set data and test set data. ROC curve result and cut point of Training set is X (0.79), Y (0.997) and cut point of test set X (0.121), Y (0.999) that show training set data and test set data have algorithm nearby results.

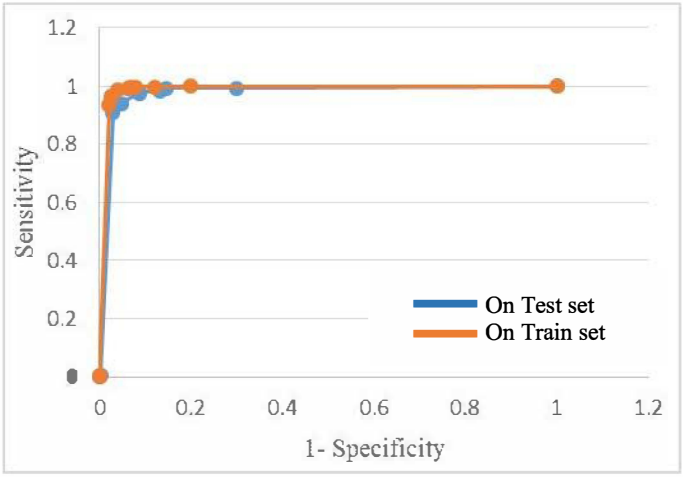

Fig. 5. Shows result of ROC curve on Trainnig set and Test Set

This paper used concentration of pollutants in Thailand based, include concentration of pollutants in air 5 kinds are Ozone, NO_2, CO, SO_2 and PM_{10} in each provinces. To create model with decision tree for use rules to separate air quality which levels to influence healthy.

From Fig. 6, use rules of decision tree with classification technique processed to show that levels of healthy concern in each provinces in Thailand. The concentration of pollutants in 5 kinds are Ozone, NO_2, CO, SO_2 and PM_{10} to show that levels of healthy concern in colors each that provinces. They have 6 levels in air quality in Thailand based. First level good is green, Second level moderate is yellow, third level unhealthy for sensitive groups is orange, Forth level unhealthy is red, Fifth level very unhealthy is purple and Sixth level hazardous is maroon [11] which the colors will change with input data in each areas Following Fig. 7.

IV. CONCUSSION

For classification of air quality that influence to healthy population in the future. It can use in difference places. In Air quality paper that influenced to healthy including concentration of pollutants in each station of Thailand to fix problems in each points more. That shows about factors were relative or change result of concentration of pollutants in each kind concentration of pollutants may be change away. This paper, researcher use 5 variants are Ozone, NO_2, CO, SO_2 and PM_{10} which others variants with air quality. And can use to analysis in same ways.

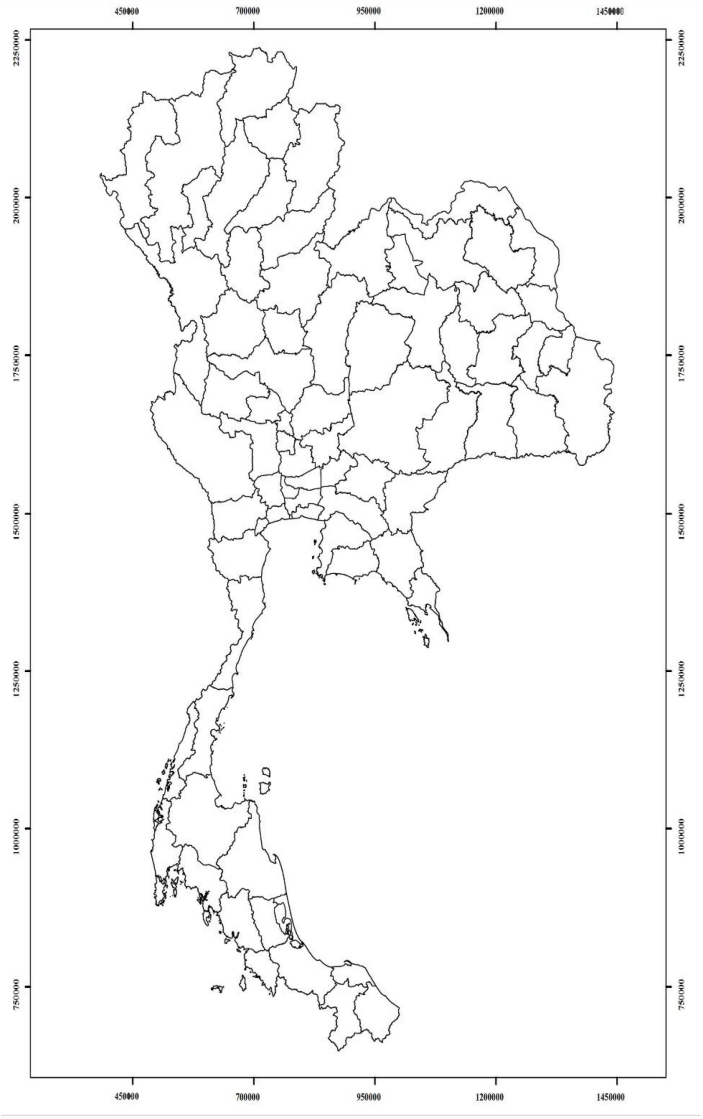

Fig. 6. The map of Thailand

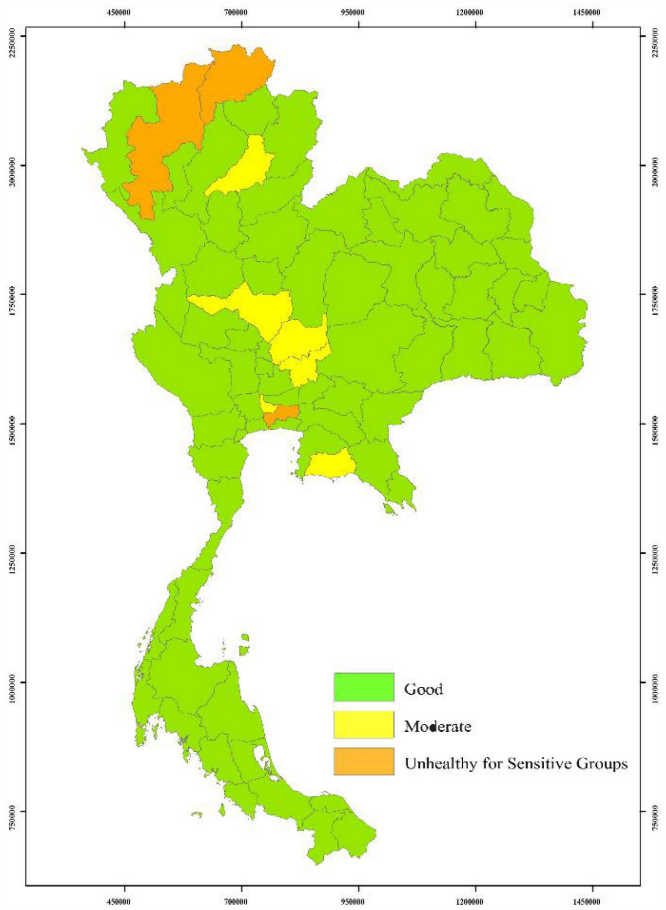

Fig. 7. Shows color level of Healthy Concern on map of Thailand

REFERENCES

[1] Ioannis N. Athanasiadis, Kostas D. Karatzas, Pericles A. Mitkas3, "Classification techniques for air quality forecasting," *Conference*, pp.1-7, August. 2006.

[2] komchadluek) .2013, May 28(. People's problems! Air pollution [Online]. URL http://www.komchadluek.net

[3] airnow .Air Quality Index (AQI) - A Guide to Air Quality and Your Health [Online]. URL http://airnow.gov

[4] Pollution Control Department .Air Quality [Online]. URL http://www.pcd.go.th

[5] Mohammad Hossein Sowlat, Hamed Gharibi, Masud Yunesian, Maryam Tayefeh Mahmoudi, Saeedeh Lotfi. "A novel, fuzzy-based air quality index (FAQI) for air quality assessment," Volume 45, Issue 12, pp. 2050–2059, Apr. 2011.

[6] KAŠPAROVÁ MILOSLAVA, KŘUPKA JIŘÍ, "Air Quality Modelling by Decision Trees in the Czech Republic Locality," *Conference*, pp. 1-6, August 2008.

[7] Hand, David J, "Measuring classifier performance: A coherent alternative to the area under the ROC curve " Machine Learning, Volume 77, pp 103-123, Jun 2009.

[8] M. Amarnath, V. Sugumaran, Hemantha Kumar, "Exploiting sound signals for fault diagnosis of bearings using decision tree," *Journal*, vol. 46, pp. 1250–1256, Apr. 2013.

[9] Mevlut Ture, Fusun Tokatlib, Imran Kurtc "Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients," *Journal*, vol. 36, pp. 2017-2016, Mar. 2009.

[10] Hone-Jay Chu, Chuan-Yao Lin, Churn-Jung Liau, Yi-Ming Kuo, "Identifying controlling factors of ground-level ozone levels over southwestern Taiwan using a decision tree," *Journal*, vol. 60, pp. 142-152, Dec. 2012.

[11] D. Nerini, J.P. Durbec, C. Mante, "Analysis of oxygen rate time series in a strongly polluted lagoon using a regression tree method," *Journal*, pp. 95–105, 2000.