MJIT 2017 Malaysian Journal of Industrial Technology

ANALYSIS OF FEATURE SELECTION WITH K-NEAREST NEIGHBOUR (KNN) TO CLASSIFY INDOOR AIR POLLUTANTS

S. M. Saad[1,a], A. Y. M. Shakaff[2,b], M. Hussein[3,c],
M. Mohamad[4,d], M. A. M. Dzahir[5,e] and Z. Ahmad[6,f]

[1,3,4,5,6]Faculty of Mechanical Engineering, Universiti Teknologi Malaysia (UTM), 81310 Skudai, Johor Bahru, Malaysia.

[2]Center of Excellence for Advanced Sensor Technology (CEASTech), Universiti Malaysia Perlis (UniMAP), Taman Muhibbah, Jejawi, 02600 Arau, Perlis.

[a]shaharil@utm.my, [b]aliyeon@unimap.edu.my, [c]mohamed@utm.my, [d]maziah@utm.my, [e]azuwan@utm.my, [f]zair@utm.my

*Corresponding author's
shaharil@utm.my

Abstract

Indoor air may be polluted by various types of pollutants which may come from cleaning products, construction activities, perfumes, cigarette smoke and outdoor pollutants. This type of pollutants could emit dangerous gases such as carbon monoxide (CO), carbon dioxide (CO_2), ozone (O_3) and particulate matter. These gases are usually safe for us to breathe in if they are emitted in safe quantity but if the amount of these gases exceeded the safe level, they might be hazardous to human being especially children and people with asthmatic problem. Therefore, a smart indoor air quality monitoring system (IAQMS) is needed that able to tell the occupants about which pollutant that trigger the indoor air pollution. In this study, an IAQMS that able to classify the air pollutants has been developed. This IAQMS applies a classification method based on K-Nearest Neighbour (KNN). It is used to classify the air pollutants based on five conditions: ambient air, human activity, presence of chemical products, presence of food and beverage and presence of fragrance. In order to get good and best classification accuracy, an analysis of several feature selection based on data pre-processing method is done to discriminate among of sources. The output from each data pre-processing method has been used as the input for the classification. The result shows that KNN analysis with the data pre-processing method for most of the features obtained remarkably high classification accuracy of above 97% and able to classify the air pollutants at high classification rate.

Keywords: Indoor air quality; Air pollutants; K-nearest neighbour

1.0 INTRODUCTION

Indoor air quality (IAQ) is a complex subject due to the constantly changing interaction of complex factors that affect the types, levels, and importance of pollutants in indoor environments (EPA, 1997). Indoor air pollution (IAP), which undermines indoor air quality (IAQ), is found to contain indoor air pollutants such as harmful gases and contaminants at a concentration level up to five times higher than the concentration of these pollutants in normal air. In severe cases, the concentration level of the pollutants could rise up to 100 times than normal concentration level which lead to hazardous condition for human's health [1]. In fact, more attentions should be given to the issue of IAP because people normally spend 90 percent of their time in indoor environments [1].

Poor IAP can be due to indoor air contaminants or inadequate pollution controls despite otherwise normal or baseline rates of ventilation. Sources of indoor air pollutants are from different origins including the occupants themselves (carbon dioxide gas exhaled), inadequate materials or materials with technical defects used in the construction of the building, the work performed within (cleaning of carpet), excessive or improper use of normal products (pesticides), combustion gases (smoking), and cross-contamination coming from other poorly ventilated zones (DOSH, 2010).Poor IAQ level could lead to certain health effects while extremely poor IAQ level could be fatal. Different IAQ parameters may come from different sources and impose different health effects towards human. Some IAQ parameters such as radon are of concern because exposure to high levels of this pollutant over long periods of time increases risk of serious, life threatening illness such as lung cancer. Other contaminants, carbon monoxide (CO) specifically, can cause death within minutes if the concentration level is too high. Some pollutants can cause both short term and long term health problems. Short term exposures to tobacco smoke can cause irritation and significant respiratory problems for some people, while long term exposures may cause lung cancer (EPA, 1997). Nonetheless, people react differently even though they were exposed to the same contaminants at similar concentration. Exposure to a very low level of chemicals may be irritating to some people while some other might not have any reactions. People with asthma and pre-existing conditions are more vulnerable to certain types of exposure than other people (EPA, 1997).

Based from health effects that have been mention before, a continuous monitoring of indoor air quality monitoring system (IAQMS) is deemed important in order to ensure that people breathe-in safe air comfortably in the indoor environments. This study proposed an enhancement of IAQMS which could classify IAP based on K-Nearest Neighbour (KNN) classification technique which has been used in many data mining applications. For the purpose of this study, the pollutants used in this study have been classified into 5 categories: ambient air, human activity, presence of chemical substances, presence of fragrances and presence of food and beverage. In order to sense and measure the pollutants, this study uses an IAQMS which was already developed for previous studies [2][3][4]. The development of IAQMS has been described in previous papers adopts an array of sensors including gas sensors, particle sensors and thermal sensors to detect multiple pollutant parameters at a relatively low cost as compared to the professional sensing devices. It uses eight sensors to measure nine indoor air pollutants which are Oxygen (O_2), Carbon Dioxide (CO_2), Carbon Monoxide (CO), Ozone (O_3), Nitrogen Dioxide (NO_2), Volatile Organic Compounds (VOCs), Particulate Matter (PM), Temperature (Temp) and Relative Humidity (RH).

2.0 EXPERIMENTAL DETAILS

There are certainly a lot of activities and conditions which could be triggered indoor air pollutant (IAP) such as carpets and furnishings, cleaning products, office machines, construction activities, water-damaged building materials, perfumes, cigarette smoke, insects and outdoor pollutants. Although these pollutants are usually safe for human, they could be hazardous to human being especially people with respiratory-related problem and children, if their amount exceeded certain limits as proposed by the US EPA [5][6]. For the purpose of this study, the pollutants of IAP are limited to five conditions that are commonly present in indoor environment: ambient air, human activity, presence of chemical products, presence of food and beverage, and presence of fragrance [1][5][7][6]. Table 1 summarized the 5 conditions for sources of indoor air pollution and their proxy that have been used in this study.

Table 1. Sources of indoor air pollutants

Condition	Proxy	Brand
Human Activity	Cigarette	Marlboro
Chemical Present	Cleaning Agent	Dettol
Fragrance Present	Air Freshness	Ambi Pur Lavender
Food & Beverage Present	Rotten Fish	Mackerel
Ambient Air	Ambient Air	Ambient

Once all the 5 conditions of sources of indoor air pollution have been identified, an experiment simulating the 5 conditions was set up for data collection purposes. The experiment was conducted in medium-size room of 4.5m x 2.4m x 2.6m which is equipped with an air-conditioner located at the center of the room at a height of 2.2m from the floor. The sensor module which is used to collect the data on indoor air was installed hanging up to the wall of the room with 1.1 meter height above the ground, a position considered as breathing zone for the occupants [8]. The sensor module was powered up using an adaptor 7.5V and was programmed to send the data to the base station every 1 minute. The data collection was conducted over 16 days between 9.00 a.m. and 5.00 p.m. with the room Temp set at 22°C. Every day, after each experiment, the air in the room was purged out by opening windows to clean the air. The process of data collection for all 5 conditions is starts from day 1 to day 16 (February 2, 2016 until February 17, 2016).

3.0 FEATURE SELECTION

One of the most important parts of intelligent system is its ability to extract useful information that is less redundant than the original one to aid fast processing on pattern recognition or classification. Before doing the data analysis, the sensor output must be processed to free itself from the drift effect, the intensity dependence and possibly from non-linearities [9]. This is called feature extraction. There are many ways or method to do the feature extraction but in this study the feature extraction can be done based on data pre-processing method. Data pre-processing is a procedure that involves on extracting certain significant characteristics from the sensor response curves or transient response in order to produce a set of numerical data or feature for further processing [10]. Choosing the correct pre-processing technique is important because it may induce the success of subsequent analysis and affect the performance of pattern recognition [11]. Most data pre-processing techniques are basically derived from a typical sensor response as shown in Figure 1 shows when the sensor is exposed to a certain odour. *Vo* is a measured value in clean ambient air or initial value called baseline while *Vs* is response value to odour or smell. Basically, data pre-processing techniques can be divided into three major categories: baseline manipulation, normalization and compression.

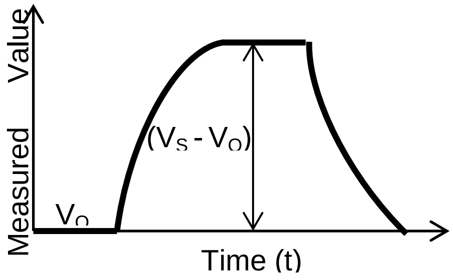

Fig. 1. Typical sensor response

The study selects 5 data pre-processing techniques which are frequently used in odour pattern recognition as summarized in Table 2. Raw data is also chosen as one of the features.

Table 2. Data pre-processing techniques selected

No	Technique	Abbreviation	Formula	References
1	RAW	RW	$X_{ij} = V_{ij}$	[12]
2	Differential	DIFF	$X_{ij} = V_{ij} - V_{bj}$	[11][12] [13]
3	Relative	REL	$X_{ij} = \frac{V_{ij}}{V_{bj}}$	[11][12][13]
4	Fractional	FRACT	$X_{ij} = \frac{V_{ij} - V_{bj}}{V_{bj}}$	[11][12][13]
5	Sensor Normalization	SN	$X_{ij} = \frac{V_{ij} - V_{ij}^{min}}{V_{ij}^{max} - V_{ij}^{min}}$	[11][12][13]
6	Vector Array Normalization	VAN	$X_{ij} = \frac{V_{ij}}{\sqrt{\sum_{q=1}^{N}(V_{ij})^2}}$	[11][12][13]

4.0 RESULTS AND DISCUSSION

The KNN is a method for classifying samples based on their closeness or distance to the training stored dataset. It is a simple method but an effective method for classification. It is type of supervised machine learning algorithms where the class of the sample is assigned to the closest class of training group. One advantage of KNN method is no training involved in KNN. The training dataset is stored without processing and used directly as the 'knowledge base'. In order to use the KNN method, an appropriate value of *K* need to be chosen since the success of classification is very much dependent on this *K* value [14]. In a sense, the KNN method is biased by *K*. There are many ways of choosing the *K* value, but a simple one is to run the algorithm many times with different *K* values and choose the one with the best performance. For purpose of classification, KNN also used two datasets; training dataset (to learn the patterns) and test dataset (to evaluate the learned or trained system). In this study, training dataset contains the input value of sensor including gas and physical parameters and source of pollutant as output class. Test dataset is similar to the training data set, except that it does not have the output class information. In order to classify a class of the sample from test dataset, KNN calculates the distance between the sample and each training points in that have been stored. Then, the distances values are sorted and nearest neighbors which dependent on *K* are determined. The labels or output classes of these neighbors are gathered and a majority vote is used for classification.

In this section, the result of developed KNN models along with its classification performance of six feature databases, namely, RW, DIFF, REL, FRACT, SN and VAN are discussed. For each feature, a separate KNN model with adjusted *K* values was formulated. Each KNN model used two datasets; training dataset with 2880 training samples (60% of 4800) and was tested with the remaining 1920 samples (40% of 4800). The classification results for the RW database is shown in Table 3.

Table 3. Performance of KNN for RW feature

KNN Model	K Factor	Classification Accuracy (%)
1	1	99.06
2	2	99.06
3	3	98.75
4	4	98.65
5	5	98.54
6	6	98.54
7	7	98.33
8	8	98.33
9	9	98.13
10	10	98.13

From Table 3, it is observed that the maximum classification accuracies for the RW feature network model is 99.06% (K factor is 2). In this study, the KNN classifier used the library provided by MATLAB software. The result of classifier which depends on K value is depending on type of data used and formula applied in KNN library.

The same procedure was repeated for the other features for KNN which are DIFF, REL, FRACT, SN and VAN. The classification rate for these features is shown in Table 4 along with the classification rate for RW feature. From Figure 2, it can be seen that all features obtained remarkably high classification accuracy of above 97% except for SN feature. SN feature gained the lowest classification accuracy at 88.33% (K factor is 4). The highest accuracy is gained by VAN feature again with 100% (K factor is 2) classification accuracy. The classification accuracy for other features such as RW, DIFF, REL and FRACT obtained classification accuracy of 99.06% (K factor is 2), 98.96% (K factor is 6), 98.13% (K factor is 2) and 98.13% (K factor is 2) respectively. It is interesting to note that REL and FRACT share the same classification accuracy.

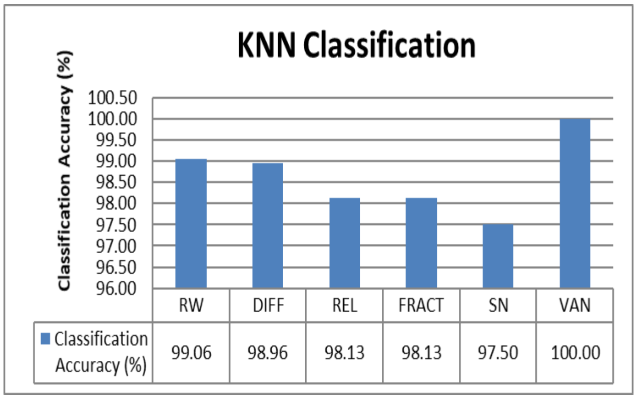

Fig. 2. KNN classification rate for each feature

The confusion matrix for the features giving the lowest and the highest classification accuracy are shown in Table 5 and Table 6.

Table 5. Confusion matrix of KNN for SN feature

	Predicted						Confusion Level (%)
	Sources of IAP	Ambient	Chemical	Food & Beverages	Fragrance	Human Activity	
Actual	Ambient	376	0	2	6	0	2.08
	Chemical	2	378	0	2	2	1.56
	Food & Beverages	4	4	354	8	14	7.81
	Fragrance	20	2	28	312	22	18.75
	Human Activity	22	6	46	34	276	28.13

Table 6. Confusion matrix of KNN for VAN feature

	Predicted						Confusion Level (%)
	Sources of IAP	Ambient	Chemical	Food & Beverages	Fragrance	Human Activity	
Actual	Ambient	384	0	0	0	0	0
	Chemical	0	384	0	0	0	0
	Food & Beverages	0	0	384	0	0	0
	Fragrance	0	0	0	384	0	0
	Human Activity	0	0	0	0	384	0

Rows and columns represent actual and predicted values, respectively. Rows and columns represent actual and predicted values, respectively. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. Table 5 shows the confusion matrix for feature SN since it gave the lowest classification accuracy. Based on the table, it can be observed that every condition contributes to the confusion level with human activity having the highest confusion level at 28.13%. Table 6 presents confusion matrix for VAN feature which has the highest classification accuracy. Compared to confusion matrix of SN in Table 4.13, KNN does not have any confusion in classifying all the 5 conditions. It means that it can classify all of the five conditions correctly. This confusion matrix validates the classification rate for VAN which 100%.

To prove that VAN feature really gave 100% classification accuracy, another analysis has been done. The principal component analysis (PCA) visualization for the VAN feature is done in 3D plot as shown in Figure 3. From Figure 3, it can be seen that none of the five conditions coincide with each other, and therefore they are mutually exclusive.

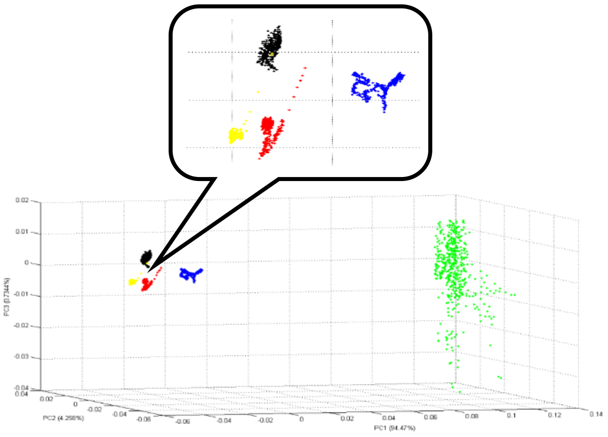

Fig. 3. 3D plot of PCA visualization for VAN feature

5.0 CONCLUSION

In this paper, an enhancement of indoor air quality monitoring system (IAQMS) has been developed which could classify IAP based on five different environments such as ambient air, chemical presence, fragrance presence, foods and beverages, and human activity. The enhancement of IAQMS applies a classification technique based on K-Nearest Neighbour (KNN) classification technique which has been used in many data mining applications. In order to get good and best classification accuracy, an analysis of several feature selection based on data pre-processing method is done to discriminate among of pollutants. The output from each data pre-processing method has been used as the input for the classifier. The result shows that the KNN model with data pre-processing method for most of the features give high classification accuracy above 97% classification accuracy . The result also shows that the five pollutants are successfully classified by KNN using VAN feature which give 100% classification accuracy. To prove that VAN feature really gave 100% classification accuracy, principal component analysis (PCA) visualization for the VAN feature is done in 3D plot. The result showed that none of the five conditions coincide with each other, and therefore they are mutually exclusive. Overall, it can be concluded that the system delivered a high classification rate based feature selection and KNN analysis.

Acknowledgement

The equipment used in this project was pre-developed by Universiti Malaysia Perlis (UniMAP) for previous studies. The authors would like to offer a special thanks to the members of the Centre of Excellence for Advanced Sensor Technology (CEATech), UniMAP, Malaysia for their critical advice and warm cooperation. The authors also would like to acknowledge the financial sponsorship provided by Malaysian Technical University Network (MTUN) grant and the Malaysian Ministry of Higher Education (MOHE) under myBrain15 scheme. Lastly, the authors would like to thanks to Universiti Teknologi Malaysia (UTM) for supporting this research funding and facilities through PAS grant (Project No: 02K56).

References

[1] EPA, "Buildings and their impact on the environment: A statistical summary," *U.S. Environmental Protection Agency Green Building Workgroup*, 2009. [Online]. Available: http://www.epa.gov/greenbuilding/pubs/gbstatpdf. [Accessed: 26-Nov-2014].

[2] S. M. Saad, A. Y. M. Shakaff, A. R. M. Saad, and A. M. Y. Kamarudin, "Implementation of index for real-time monitoring indoor air quality system," *2014 2nd Int. Conf. Electron. Des. ICED 2014*, pp. 53–57, 2011.

[3] S. M. Saad, A. R. M. Saad, A. M. Y. Kamarudin, A. Zakaria, and A. Y. M. Shakaff, "Indoor air quality monitoring system using wireless sensor network (WSN) with web interface," *2013 Int. Conf. Electr. Electron. Syst. Eng.*, pp. 60–64, 2013.

[4] S. M. Saad, A. M. Andrew, A. Y. M. Shakaff, A. R. M. Saad, A. M. Y. Kamarudin, and A. Zakaria, "Classifying sources influencing indoor air quality (IAQ) using artificial neural network (ANN)." *Sensors (Basel).*, vol. 15, no. 5, pp. 11665–84, 2015.

[5] R. B. G Invernizzi, A Ruprecht, R Mazza, E Rossetti, A Sasco, S Nardini, "Particulate matter from tobacco versus diesel car exhaust: an educational perspective," 2004.

[6] Loomis, D., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa., "The carcinogenicity of outdoor air pollution," *Lancet Oncol.*, vol. 14, pp. 1262–1263, 2013.

[7] K. Meena, "Indoor air pollution: sources, health effects and mitigation strategies," 2009.

[8] DOSH, "Industry code of practice on indoor air quality," 2010.

[9] A. C. Romain, J. Nicolas, V. Wiertz, J. Maternova, and P. Andre, "Use of a simple tin oxide sensor array to identify five malodours collected in the field," *Sensors Actuators, B Chem.*, vol. 62, pp. 73–79, 2000.

[10] C. Distante, M. Leo, P. Siciliano, and K. C. Persaud, "On the study of feature extraction methods for an electronic nose," *Sensors Actuators, B Chem.*, vol. 87, no. 2, pp. 274–288, 2002.

[11] R. Gutierrez-Osuna and H. T. Nagle, "A method for evaluating data-preprocessing techniques for odor classification with an array of gas sensors," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 29, no. 5, pp. 626–632, 1999.

[12] J. Nicolas, a. C. Romain, V. Wiertz, J. Maternova, and P. André, "Using the classification model of an electronic nose to assign unknown malodours to environmental sources and to monitor them continuously," *Sensors Actuators, B Chem.*, vol. 69, pp. 366–371, 2000.

[13] J. W. Gardner and P. N. Bartlett, "A brief history of electronic noses," *Sensors Actuators B Chem.*, vol. 18, no. 1–3, pp. 210–211, 1994.

[14] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," *Move to Meaningful Internet Syst. 2003 CoopIS, DOA, ODBASE*, pp. 986–996, 2003.