Data mining-based indoor air quality monitoring system in the era of big data

Lina Yuan 1, *, Huajun Chen 1, Bo Tian 1, and Ging Gong 2

1 College of Data Science, Tongren University, Tongren, China.
2 Graduate School, Tongren University, Tongren, China.

*Corresponding author e-mail: 893422817@qq.com

Abstract. With the ever-increasing enhancement of living standards, indoor air quality has been drawing more and more attention, due to the conditions of indoor environment with significant relevance to human growth development, health and work efficiency. The quality of indoor environment has a direct impact on the quality of people's life, and even concerns the issue of human survival. In consequence, the pervasive and exponentially increasing indoor air data presents imminent challenges to how to monitor and evaluate the indoor environment parameters, and create a comfortable, quiet and clean indoor environment, which is indispensable based on data mining methods in the context of big data. This paper focuses on two methods of nonlinear feature extraction: Kernel principal component analysis (KPCA) and PCA. Experimental results demonstrate that the recognition accuracy of feature extraction-based KPCA is superior to that of PCA.

1. Introduction
In the era of Internet, with the exploding growth of indoor air data, based on data mining in the context of big data, the analysis of indoor air monitoring system has become a challenging hot topic. Specifically, depending on accurate big data, the user experience will be better. For instance, the successful development of innovative products such as air box and sky platinum air conditioner is according to the results of big data analysis and statistics, which are not so much pure intelligent air products (IAPs) as personal health air management of intelligent housekeepers. By combining the quality inspection and calibration services of humidity, temperature, dust sensors, and other big data, the accuracy of sensor detection of IAPs is sharply enhanced [1].

The indoor environment includes closed places for people to live, work, socialize and play in, such as residence, office study, medical treatment, entertainment, sports and transportation, and so on. Because most of the time in a person's life is spent in indoor, and people are exposed to indoor air pollutants for much longer than they are outside, a comfortable interior environment is absolutely vital for people's physical and mental health [2,3].

The classification of common indoor pollutants is illustrated in Fig. 1 [4]. It can be seen from the properties, sources and harm of indoor pollution gas, that most of the pollution gases have carcinogenic and toxic properties, which have certain undiscoverable characteristics. Hence, it plays a significant role in improving people's living standards to timely, accurately and reliably detect the types and concentrations of hazardous gases in the environment.
Common indoor pollutants

Physical pollution
- Electrostatic pollution
- Noise pollution
- Electromagnetic radiation pollution
- Mites pollution
- Legionella contamination
- Parasitic contamination

Biological contamination
- *Lifestyle pollutant*: natural discharge from human body, mainly including radon, ammonia and other pollution gases
- *Domestic chemical pollutants*: from cleaning and washing products, cosmetics, mainly including organic volatile VOC and other pollution gases
- *Building decoration pollutants*: from building materials and decoration materials, mainly including formaldehyde, benzene, heavy metal, etc.
- *Gas and natural gas*: etc., also produce a large number of pollution gases, such as CO and NOx.

Chemical pollution

Figure 1. The classification of common indoor pollutants.

2. Evaluation methods of indoor environmental comfort level and its influencing factors

2.1. Evaluation methods of indoor environmental comfort level

The quality of indoor environment is mainly depended on indoor microclimate, i.e., air temperature, relative humidity, air velocity, noise level, illumination, air quality, etc., when these factors comprehensively effect on the human body, and are in a state of optimal combination, it can make human body produce comfortable feeling. In this article, indoor environment can be divided into four parts: indoor thermal environment, light environment, sound environment, and indoor air quality, as shown in Fig. 2 [5].

![Evaluation-based Hierarchical diagram of indoor comprehensive environment.](image)

Thermal comfort refers to people's lukewarm subjective feeling of the microclimate. The quality of indoor thermal environment have a significant impact on people's physical health, work and learning efficiency. Research studies show that when the air temperature is about 25°C, the efficiency of mental work achieves the most; when lower than 18°C or higher than 28°C, it drops sharply [5]. Therefore, it is necessary to select thermal comfort as one of the indoor environmental evaluation indicators.

Luminous environment has been a serious threat to human health and work efficiency, causing serious harm to people. Studies indicate that light pollution can damage the cornea and iris of the human eye, inhibits the function of photoreceptor cells in the retina, and causes visual fatigue and vision loss. For this purpose, it is indispensable to pay attention to light environmental pollution, evaluate the comfort level of indoor light environment, and improve the visual environment.

Noise pollution does great harm to people's physical and mental health, such as ear discomfort, cardiovascular injury, the dysfunction of nervous system, mental disturbance, endocrine disorder, the lower work efficiency and other indoor noise sources. The main sources of indoor noise are: traffic noise, urban construction noise, noise from social life and public place, and household electrical noise. In modern society, with continuous increase of various noise, it is urgent to be kept a watchful eye on and enhanced the acoustic environment.
2.2. The influencing factors of indoor environmental comfort level

Indoor air environment is one of the most frequent contacted and the closest environments. With the growing of sources and quantity of indoor pollutants, the more the degree of building tightness increases, the worse indoor air quality seriously deteriorates. In consequence, it is of great necessity to evaluate indoor air quality, master the status and changing trend of indoor air quality, and make people's life from comfort to health development.

2.2.1. Select the evaluation index of indoor air quality. Objective evaluation of indoor air quality is directly evaluated by indoor pollution index, i.e., choose the representative pollutants as the evaluation index, and comprehensively and fairly reflect the status of indoor air quality. Due to the influence factors of different types of building indoor air quality is not quite the same, the chosen evaluation index should also be emphasized. This article selects four parameters, i.e., formaldehyde, total volatile organic compound (TVOC), NOX, and CO2 concentration, as the evaluation index of indoor air quality, and simulates the measurement data in the reference [6]. When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. Should authors use tables or figures from other Publications, they must ask the corresponding publishers to grant them the right to publish this material in their paper.

2.2.2. The classification of evaluation index grades for indoor air quality. The basis of evaluating indoor air quality is national standard, in this paper the current national standard of indoor air quality [7] is selected as the evaluation basis, and the evaluation index of indoor air quality is divided into four grades: best, good, moderate, and worst, as shown in Table 1.

| Contaminant | Unit | Best | Good | Moderate | Worst |
|------------|------|------|------|----------|-------|
| Formaldehyde | mg/m³ | ≤0.04 | (0.04, 0.08) | (0.08, 0.12) | >0.12 |
| TVOC       | mg/m³ | ≤0.2 | (0.2, 0.3) | (0.3, 0.6) | >0.6 |
| NOX        | mg/m³ | ≤0.02 | (0.02, 0.06) | (0.06, 0.2) | >0.2 |
| CO₂        | ppm  | ≤600 | (600, 1000) | (1000, 1500) | >1500 |

3. Data mining methods and simulation results

3.1. The kernel method

The core idea of kernel function method [8] is: first by nonlinear function the input space is mapped to high-dimensional feature space, where data is then processed. The key of this method is that the higher-dimensional eigenspace inner product operation after nonlinear transformation is transformed into the kernel function calculation of the original space by introducing kernel function, which greatly simplifies the calculation.

The kernel method realizes the nonlinear transformation among the data space, feature space and class space. Suppose xi and xj are the sample points in the data space, the mapping function from data space to feature space is Φ (·), and the basis of the kernel function is to implement the inner product transformation of a vector, i.e.

\[ K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle \quad (1) \]

Where \( \langle \rangle \) is the Inner Product, \( K(x_i, x_j) \) is the kernel function. From the formula (1), Kernel function transforms the inner product operation of high dimensional space into kernel operation of low dimensional input space, which skillfully solves the problems, such as dimension disaster etc., in high dimensional feature space, thus in order to lay a theoretical foundation for resolving complex classification or regression problems in high dimensional feature space. Generally, the complexity of the nonlinear transformation function is fairly high, while the actual kernel function in the process of calculations is relatively easier.
The Kernel function is a far more effective method to settle nonlinear problems, mapped to high-dimensional space by kernel functions, which makes the computational complexity of the high dimensional feature space independent of the space dimension. Because the kernel function-based method has many advantages, it has been widely applied in modern science and technology, for example, support vector machine (SVM), and the kernel method of pattern recognition (such as handwriting recognition (HWR) and text recognition, etc.).

The most commonly used kernel function is Gauss kernel function, and its expression is

\[ K(x, y) = \exp \left( -\frac{||x - y||^2}{\sigma^2} \right) \]  

(2)

Where \( ||x - y||^2 \) is the square of the norm. And the commonly used kernel functions are shown in Table 2 [9].

| The kernel type                   | Mathematical expression |
|-----------------------------------|-------------------------|
| Linear kernel                     | \( K(x, x_i) = x \cdot x_i \) |
| P-order polynomial kernel         | \( K(x, x_i) = \left( (x \cdot x_i) + 1 \right)^p \) |
| Gaussian kernel                   | \( K(x, x_i) = \exp \left( -\frac{||x - x_i||^2}{\sigma^2} \right) \) |
| Multi-layered perceptron kernel   | \( K(x, x_i) = \tanh\sqrt{v(x \cdot x_i) + c} \) |

3.2. Kernel principal component analysis

Kernel principal component analysis (KPCA) [10] is a kind of nonlinear method, based on the classical principal component analysis, the nonlinear problems in the original space are transformed into linear problems in the mapping space by introducing kernel function [8]. The basic idea of KPCA is that the input space is projected into the high-dimensional feature space by nonlinear mapping, where the mapping data is then acted as the PCA [11], which has strong nonlinear processing ability. Therefore, KPCA performs nonlinear PCA in the input space.

Fig. 3 and Fig. 4 shows the variation trend of the identification precision of six kinds of gas training sets and test sets, when the number of principal components increases in turn [4]. Wherein, the blue dotted line indicates the average identification rate of all target gases, and the remaining solid lines represent the identification rate of a single target gas, respectively. From Fig. 3, the average recognition rate of the training set increases steadily with the increase of the number of principal components, when added to the eighth principal components, the recognition rate begins to level off; Although there are some fluctuations in the identification rate of a single gas, the overall recognition rate presents an upward trend and finally tends to be stable. Fig. 4 demonstrates that with the gradual increase of the number of principle components, the average recognition rate of the test set of the six gases is also gradually increased, and tends to be stable. Here, the identification rate of a single gas is the same as that of the training data set, and there are some fluctuations, but the overall trend is upward.

3.3. Principal component analysis (PCA)

Principal component analysis (PCA) [11] is a classical multivariate statistical analysis technology. PCA technique can not only be used for reducing data dimension, removing redundant information in the data and the correlation between variables, and lowering the computational complexity, but also provide important information data such as data mean level, the largest direction of data average, etc. In addition, PCA can be leveraged for feature extraction.

PCA is a kind of principal component data processing method through the orthogonal transformation, which converts a set of probabilistic variables into linearly independent new variables.
Its basic idea is to translate and rotate the original coordinate system, and make the coordinate origin of the new coordinate system be coincident with the center of gravity of the data group points, so in the new coordinate system, the maximum direction of the original data projection variation corresponds to the first principal axis direction, while the second principal axis is orthogonal (independent) to the first one, and the variation of projection data is the second largest direction. In a similar fashion, all of the new coordinate axes can be got. These new axes are respectively called first principal axes, second principal axes and so on, which is named the first and second principal component direction. The number of principal components is always less than or equal to the dimension of the original data. If the principal axis with small variation is removed, the variables of remaining principal axis direction can contain the main information of the original data, then the original multidimensional space has been realized dimensional reduction. Thus, the original data can be projected to the new principal component of the coordinate axis direction, and the original data group points are expressed by the data of the projection space.

So the essence of PCA is to project multiple related variables into another set of orthogonal spaces, in order to get a new set of variables, which makes the new variables have the greatest variance (It is thought that signals have the greatest variance in signal processing, noise has the smallest variance, and the direction with the largest variance corresponds to the direction with the most abundant signal information). After PCA, the dimension of the signal can be reduced and the computational complexity of the problem can be reduced on the basis of retaining the main information.

First, extract the effective features on analyzing the training data set using PCA. Table 3 provides the characteristic value of each component and the cumulative contribution rate [4]. It can be observed from Table 3, the cumulative contribution rate about the data sets of first five principal components reaches 98.81%, which indicates that the information reflected by the original 6 variables is basically retained. As a result, we can extract the first 5 principal components as input characteristic variables of the support vector machine (SVM) [9] classifier.
4. Conclusion

Indoor environment is closely related to people's lives, and has become one of the environmental issues concerned by plenty of countries. Particularly, the concerns, such as the temperature in the office or indoor, humidity, illumination, noise level, air quality, etc., are the center of attention. A problem of concern. As a consequence, a growing number of portable, real-time and accurate air quality monitoring systems will be designed.

Acknowledgments

This work was financially supported by the science and technology cooperation program project in Guizhou in 2015 (Project Number: [2015]7251 Marked by the word LH in Guizhou Province), the project of platform talent in Guizhou in 2016 (Project Number: [2016]5611), and partly supported by the National Natural Science Foundation of China (NO.61562703).

References

[1] A. Altayeva, B. Omarov, Y. I. Cho, Towards Smart City Platform Intelligence: PI Decoupling Math Model for Temperature and Humidity Control, 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Jan. 2018, pp. 693-696.
[2] R. Pitarma, G. Marques, and B. R. Ferreira, Monitoring indoor air quality for enhanced occupational health, Journal of Medical Systems, 2017, 41(2): 1-8.
[3] Y. Wang and N. Hajli, Exploring the path to big data analytics success in healthcare, Journal of Business Research, 2017, 70: 287-299.
[4] Dang Lijun, Study on the Multiple Classifiers Ensemble Technology in Air Quality Monitoring System, Chongqing University, April 2014 (In Chinese).
[5] Zhang Fangfang, Study on Multi-information Fusion Algorithm of Indoor Environment, Chongqing University of Technology, April 2010 (In Chinese).
[6] Yang Hongzhu, Research on Residence Indoor Air Quality Evaluation Based on the GA-BP Algorithm, Huazhong University of Science and Technology, June 2007 (In Chinese).
[7] The quality supervision and inspection center of national indoor environment and indoor environmental protection product, GB/T 18883-2002, Indoor Air Quality Standard, China Standards Press, 1 March 2009 (In Chinese).
[8] T. Hofmann, B. Scholkopf, A. J. Smola, Kernal Methods in Machine Learning, The Annals of Statistics, 2008, 36(3):1171-1220.
[9] Zhou Zhihua. Machine Learning. Tsinghua University Press, (2016).
[10] João B. O. Souza Filho, Paulo S. R. Diniz, Improving KPCA Online Extraction by Orthonormalization in the Feature Space, IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(4): 1382-1387.
[11] Xiaogang Deng; Xuemin Tian; Sheng Chen; Chris J. Harris, Nonlinear Process Fault Diagnosis Based on Serial Principal Component Analysis, IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(3): 560-572.