A Method for Recognition of Mixed Gas Composition Based on PCA and KNN

Wanyu Xia[1], Tingting Song[1], Zhanwei Yan[1], Kai Song[3], Deyun Chen[2], Yinsheng Chen[1,2,]*
[1]National Experimental Teaching Demonstration Center for Measurement and Control Technology and Instrumentation, Harbin University of Science and Technology, Harbin 150080, China
[2]Postdoctoral Research Station of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China
[3]School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin 150080, China
Corresponding author: Yinsheng Chen (e-mail: chenys@htbust.edu.cn).

Abstract

In the mining process of coal, oil, and natural gas, there are often a large number of toxic, hazardous and explosive gases, which pose a greater safety hazard. These gases usually exist in the form of mixtures. Accurately identifying the composition of the mixed gas is of great significance to the prevention of safety accidents. In order to improve the accuracy of the component recognition of the mixed gas, this paper proposes a mixed gas component recognition algorithm based on the combination of principal component analysis (PCA) and k-nearest neighbor algorithm (KNN). PCA is used to extract the characteristics of the sensor array signal to obtain the characteristic value of the gas, and then KNN is used to realize the recognition of the gas type. The results show that the recognition rate of the feature quantity after dimensionality reduction as input is higher than that before dimensionality reduction. Finally, PCA and KNN algorithm and PCA and support vector machine (SVM) algorithm are compared for the recognition rate of mixed gas. Experimental results show that the proposed method has a recognition rate of 96.88% for mixed gas components.

Introduction

Poisonous and harmful gases are generated in the mining production process, and lower concentrations can cause explosions, leading to threats to the lives of workers. The identification of the components of the mixed gas is of great significance to ensure safe production.

Capone S et al. used principal component analysis and principal component regression to qualitatively identify and quantitatively detect the mixed gas of CO and CH4, and obtained good results [1]. Paolesse R et al. used principal component analysis and support vector machine to identify moldy grains. The recognition rate of PCA was 85.3% and the recognition rate of SVM was 93.6% [2]. Laref Rachid et al. used two identical electronic noses to test for nitrogen dioxide. The experimental results show that, compared with the classical direct standardization method, the support vector machine regression algorithm improves the problem of output drift caused by the environmental change of electronic nose, and has better effect [3]. The above literature uses SVM for processing. SVM shows excellent performance in a small sample, but there is a problem that the parameters are difficult to determine.

In response to the above problems, this paper proposes a mixed gas composition recognition algorithm based on the combination of PCA and KNN. PCA transforms high-dimensional space problems into low-dimensional space for processing, making the problems simple and intuitive. The features are not related to each other, and can provide most of the information of the original data, eliminating redundant information and achieving dimensionality reduction. In the classification process, KNN realizes the classification of the model by setting the K value in the model, compares the characteristics of the test data with the corresponding characteristics in the training set, classifies the data with similar characteristics into one category, and has a high tolerance for outliers and noise , which is suitable for multi- classification problems.

Method

A.Principal component analysis

As a linear feature extraction method, PCA uses the idea of dimensionality reduction to convert multiple variables into a few principal components. Each of the principal components is a linear combination of the original variables, and the principal components are not related to each other, so these principal components can reflect most of the information of the original variables, and the information contained does not overlap each other. The principle of dimensionality reduction of the PCA algorithm is to calculate the covariance matrix of the original data matrix, and then reduce the correlation between features. The purpose is to remove redundant information and achieve dimensionality reduction to maximize the variance of the processed data.

The implementation steps of principal component analysis are as follows:

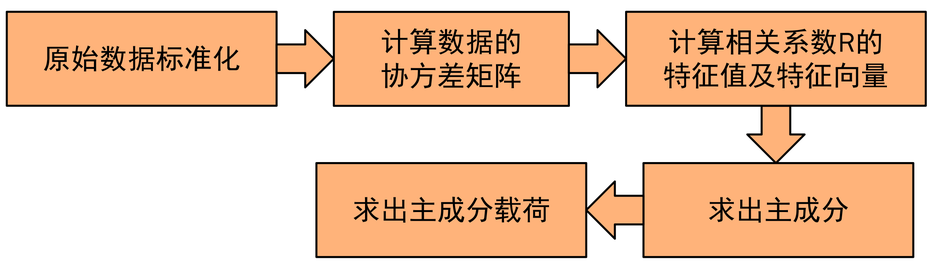

Figure 1. The implementation steps of principal component analysis

B.K nearest neighbor algorithm

K-nearest neighbor algorithm is a non-parametric classification method in machine learning. In the classification process, when the data and labels in the training set are known, after inputting the test data, the features of the test data are compared with the corresponding features in the training set. Find the top K most similar data in the training set, and the category corresponding to the test data is the category with the highest frequency among the K data. In the KNN algorithm, there are three commonly used distances, namely Manhattan distance, Euclidean distance and Minkowski distance.

(1) Euclidean Distance **(2) Manhattan distance**

$$dist(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \qquad dist(X,Y)=\sum_{i=1}^{n}|x_i-y_i|$$

Experiment and result analysis

A.Dataset

In order to verify the effectiveness of the PCA and KNN algorithms proposed in this paper in the identification of mixed gases, a verification experiment was done on the data set published by UCI. The data set was collected in a gas delivery platform facility at the ChemoSignals Laboratory in the BioCircuits Institute, University of California San Diego. The data set includes two binary mixtures: ethylene and methane mixture, ethylene and CO mixture, both mixtures are in the case of the background gas is air. The signal response data is generated by the reaction of mixed gas and 16 metal oxide sensors under different conditions. The sensor array is composed of 4 different types of gas sensors (each 4 units).

The original data uses the concentration value as the category label. In this paper, a fixed concentration of gas (single or mixed) is completely reflected as a gas sample, which is divided into 161 samples. It is divided into single ethylene, CO, methane, ethylene and CO, ethylene and methane. Each type of gas is divided in a ratio of 3:2, that is, 97 samples are used for training and 64 samples are used for testing. The composition of the experimental sample set is shown in Table I.

TABLE I. THE COMPOSITION OF THE EXPERIMENTAL SAMPLE SET

Labels	Gas type	Number of whole samples	Number of training set	Number of testing set
1	CO	30	18	12
2	Ethylene	68	41	27
3	Methane	36	10	7
4	CO-Ethylene	17	6	4
5	Methane-Ethylene	10	22	14

B.Gas composition recognition experiment

In the KNN algorithm, this paper uses two distances for classification, among which the Euclidean distance classification can get a better classification effect. The classification results are shown in Table II.

TABLE II. ACCURACY OF MIXED GAS RECOGNITION BASED ON DIFFERENT DISTANCES OF KNN

Different distance	Accuracy
Euclidean Distance	96.9%
Manhattan Distance	90.6%

In order to illustrate the effectiveness of the PCA and KNN algorithm in the identification of mixed gases, the algorithm uses PCA to reduce the dimensionality of the feature space, and finally obtains 4 new features from the 16 original features. The original features and the dimensionality-reduced features are input into KNN classifier for analysis, and the selection of k obtained through empirical rules is generally lower than the square root of the number of training samples; this time, the value of k is selected as 9. This experiment compares whether PCA performs feature extraction on the data. Finally, the recognition rate of PCA and KNN algorithm and PCA and SVM algorithm for mixed gas is compared. The results of the experiment are shown in the figure below.

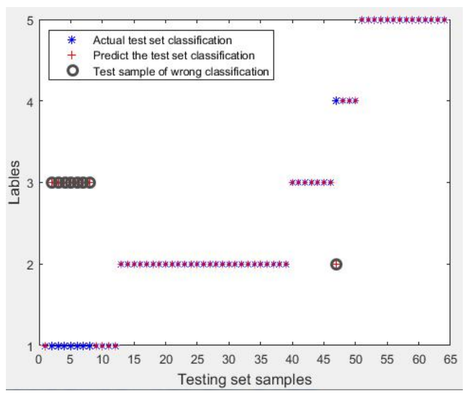

Figure 2. Recognition result of mixed gas based on KNN

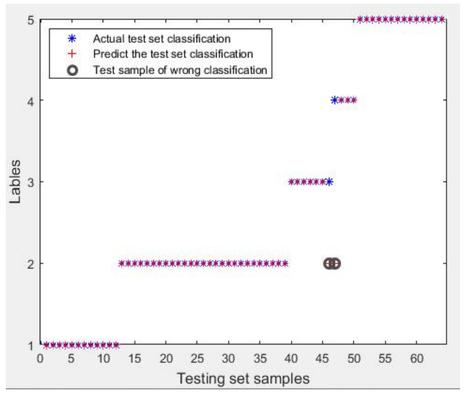

Figure 3. Recognition result of mixed gas based on PCA and KNN

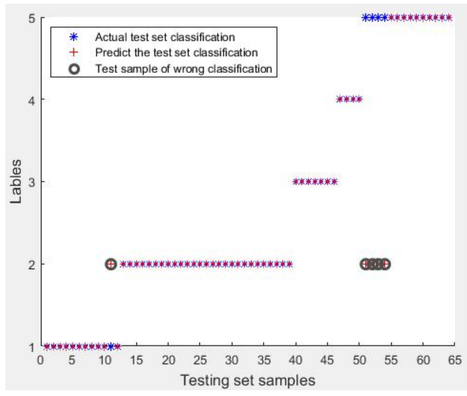

Figure 4. Recognition result of mixed gas based on PCA and SVM

In order to verify the recognition effect of the mixed gas component recognition algorithm based on PCA and KNN, this paper uses different methods to identify the mixed gas, as shown in Table III. It can be seen that the mixed gas recognition algorithm proposed in this paper has a high mixed gas recognition accuracy rate of 96.9%.

TABLE III. COMPARISON OF THE ACCURACY OF MIXED GAS RECOGNITION BY DIFFERENT METHODS

Algorithm	Accuracy
KNN	87.5%
PCA+SVM	92.2%
PCA+KNN	96.9%

Conclusion

This paper proposes a mixed gas composition recognition algorithm based on the combination of PCA and KNN. In this method, PCA is used to reduce the dimension of the original data, and then uses the KNN classifier for component recognition. The algorithm is compared with the PCA and SVM algorithms, and it is found that the algorithm proposed in this paper has a better recognition effect. The research of this algorithm continues the solution to the subsequent concentration estimation problem of the electronic nose system.

References

[1]Capone S, Siciliano P ,Barsan N,et al. 2001.Analysis of CO and CH4 gas mixtures by using a micromachined sensor array[J]. Sensors and Actuators B (Chemical), 78(1-3):40-48.

[2]Paolesse R, Alimelli A , Martinelli E ,et al. 2006. Detection of fungal contamination of cereal grain samples by an electronic nose[J]. Sensors and Actuators B (Chemical), 119(2):425-430.

[3]Laref Rachid,Losson Etienne,Sava Alexandre,Siadat Maryam. 2018.Support Vector Machine Regression for Calibration Transfer between Electronic Noses Dedicated to Air Pollution Monitoring.[J]. Sensors (Basel, Switzerland),18(11).