Fuzzy Systems and Data Mining IX
A.J. Tallón-Ballesteros and R. Beltrán-Barba (Eds.)

doi:10.3233/FAIA231045

Classification of Odor Drift Data Based on Several Machine Learning Algorithms

Yuhan WU and Yiqin BAO[1]
College of Information Engineering, Nanjing Xiao Zhuang University, China

Abstract. Based on the classification and recognition algorithm of machine learning, this paper analyzes and researches the odor drift data set. First of all, data visualization is used to effectively master the data distribution law, coherence, outlier noise points and other information of the data set. According to the situation, the data is normalized and dimensionality reduction preprocessing, and the training set and test set are divided. KNN model, decision tree model, random forest classifier model and MLP multi-layer perceptron model were used to test and compare the data sets. The test results show that the performance of random forest model for odor drift data classification is relatively good, up to 95%, which can be used in practice.

Keywords. Machine learning, Classification recognition, Odor drift data, Classification prediction algorithms

1. Introduction

With the continuous development of science and technology and data, artificial intelligence has made very significant and excellent progress in the past few decades, and has a wide range of applications in many fields. Among them, classification recognition is a very important technology in artificial intelligence, and it is one of the basic tasks of machine learning. Through classification recognition, machines can learn patterns and rules from a large number of data. And classify the new data[1].

Odor drift is a phenomenon in which odors from a source are transmitted through the air and spread to other locations or the surrounding environment. Odor drift has important applications in many fields, such as environmental monitoring, indoor air quality control and odor control. If it is not well controlled, it may have a negative impact on the surrounding environment and pose a threat to people's life and health. At present, there are few relevant studies on odor drift, because the scope and dimension of odor drift are large, and the discrimination of odor and the accuracy of measurement are very complicated.

Machine learning algorithm has the ability to process large-scale data and complex models[2], and can classify and identify multidimensional data features of odor drift, so as to judge the feature significance and recognizability of odor, so as to better explore and manage odor drift data later. Through experiments on odor drift, we can better

[1] Corresponding Author, Yiqin BAO, Nanjing Xiao Zhuang University, China; Email: 392335241@qq.com.
This work is supported by Natural Science Foundation Project of China (61976118), Key topics of the '13th five-year plan' for Education Science in Jiangsu Province (B-b /2020/01/18).

distinguish the accuracy of machine learning algorithms supported by more superior algorithms and understand the relevant principles.

Based on some classic and powerful machine learning classification algorithms, this paper analyzes the odor drift data set, preprocesses the data, evaluates the classification models respectively, selects appropriate models for data classification and result analysis, and clearly demonstrates the superiority and principle of some classification models while obtaining the degree of recognition of different gases.

The contributions and innovations of this paper are summarized as follows:

1) The data distribution characteristics and pre-processing of odor drift data set are studied.

2) Use models to model and predict data.

3) Compare and analyze the predicted results.

The rest of the paper is organized as follows: the second section analyzes and preprocesses the odor drift data set, the third section conducts data modeling and prediction, the fourth section analyzes the test results, and the fifth section summarizes the full text.

2. Analysis and preprocessing of odor drift data sets

2.1 Data preparation

This article uses the Pycharm platform, Python code for related analysis and modeling tests. For the analysis and implementation of odor Drift dataset, the Gas Sensor Array Drift dataset is a data set related to the gas sensor array and its drift phenomenon over time. This dataset is commonly used to study drift compensation techniques for gas sensors [3] and machine learning algorithms for gas classification.

The dataset contains 13,910 measurements from 16 chemical sensors used to simulate drift compensation in six gases at different concentration levels, 128 dimensional properties (8 features *16 sensors)[4], and six gas types as shown in Table 1.

Table 1. Six gas types

Gas	Introduction
Ethanol	A colorless, odorless liquid, also known as alcohol.
Ethylene	A colorless, flammable gas commonly used as a raw material in industrial production.
Ammonia	A colorless gas with a pungent odor. It is widely used as a fertilizer in chemical manufacturing and agriculture.
Acetaldehyde	An organic compound having a colorless liquid or gaseous state, a strong pungent odor, and a metabolite of alcohol.
Acetone	A colorless liquid that is sweet and volatile. It is a commonly used organic solvent and is widely used in industry and laboratories.
Toluene	A colorless liquid with a strong chemical odor; a component of many solvents, paints, and coatings; also commonly found in industrial processes.

For processing purposes, the dataset is organized into ten batches, each class and the number of measurements per month. This reorganization of the data is to ensure that there is an adequate and as evenly distributed number of experiments in each batch as possible. The data batches and corresponding months are shown in Table 2:

Table 2. Data batches and corresponding months

Batch ID	Month IDs
Batch 1	Months 1 and 2
Batch 2	Months 3, 4, 8, 9 and 10
Batch 3	Months 11, 12, and 13
Batch 4	Months 14 and 15
Batch 5	Month 16
Batch 6	Months 17, 18, 19, and 20

The data format uses the same encoding style as the libsvm format x:v, where x represents the feature number and v represents the actual value of the feature. For example:
1; 10.000000 1:15596.162100 2:1.868245 3:2.371604 4:2.803678 5:7.512213.....
The number 1 indicates the class number (in this case, ethanol), the gas concentration level is 10 PPMV, and the remaining 128 columns list the actual characteristic values for each measurement record organized as described above.

2.2 Data analysis

Read 10 data sets in turn, then convert the data features from sparse matrix to dense matrix and store them as Data Frame, convert the data labels into Series objects, access the feature list and label list in turn, and then combine the data of the feature list and label list together. The first column of the first row, 15596.1621, represents the first gas concentration data measured by the first feature, which is 15596.1621. The relevant information of the data set is shown in Figure 1.

	0	1	2	...	125	126	127
0	15596.1621	1.868245	2.371604	...	-0.545079	-0.902241	-2.654529
1	26402.0704	2.532401	5.411209	...	-0.889333	-1.323505	-1.749225
2	42103.5820	3.454189	8.198175	...	-1.334558	-1.993659	-2.348370
3	42825.9883	3.451192	12.113940	...	-1.432205	-2.146158	-2.488957
4	58151.1757	4.194839	11.455096	...	-1.930107	-2.931265	-4.088756

Figure 1. Data set related infographic

The gas classification label is added to the back of the feature data set, and the average value of each feature of the data is calculated. The histogram shows the distribution of the data, which is divided into 30 boxes. It can be seen that the average value of the data set is more concentrated in the first box. Data set indicators are unbalanced. The histogram of feature mean value is shown in Figure 2.

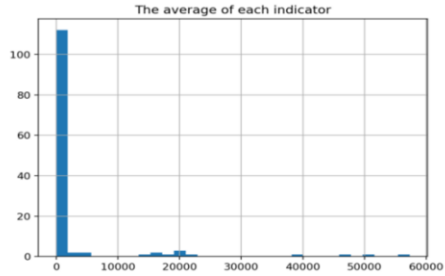

Figure 2. Feature mean histogram

Then let's analyze the average value of features of different indicators. As shown in the bar chart, it can be seen that some feature values are divided into large, and some feature values are very small and unbalanced, requiring normalization and dimensionality reduction. The bar chart of average values of different indicators is shown in Figure 3.

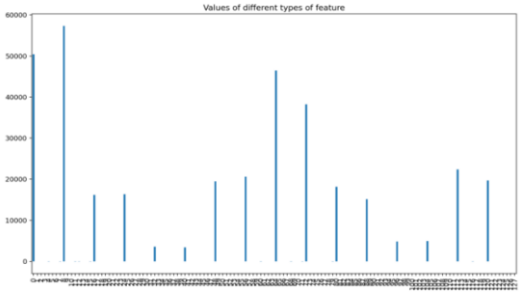

Figure 3. Bar chart of average values of different indicators

Analyze the sample number of 6 kinds of gases, add the digital label, and display it with the bar chart, you can see that the first gas has 3009 samples, the second gas has 2926 samples, and so on, each gas has a certain number of samples, relatively average, but not completely average. The column diagram for sample quantity analysis of 6 kinds of gases is shown in Figure 4.

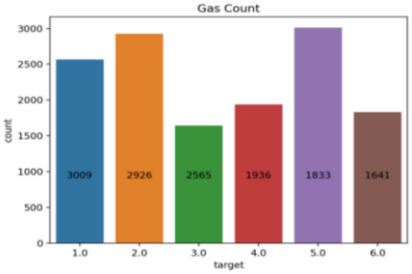

Figure 4. Gas sample number analysis column chart

The amount of data of different gases in each month is calculated, and then analyzed with a bar chart. It can be seen that around 21 months, there are more data, and the fourth gas, acetaldehyde, has the most data, and the monthly gas distribution has a large difference. The bar chart of the variation of different gas quantities over months is shown in Figure 5.

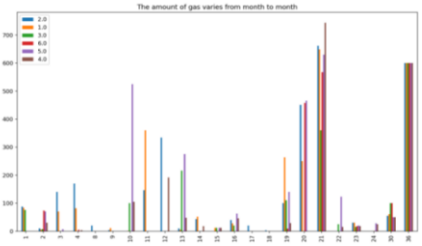

Figure 5. Bar chart of the variation of different gas quantities over months

2.3 Data preprocessing

Normalization: To preprocess the data set and normalize the feature data, we first create a MinMaxScaler object, then use the fit_transform method to scale the data to between 0 and 1, and convert the original value X to the scaled value X_scaled by the following formula:

X_scaled = (X - X.min(axis=0)) / (X.max(axis=0) - X.min(axis=0))

Where, X is the data set to be processed, and axis=0 indicates that operations are performed on each column, so the minimum and maximum values of each column are used in calculation. In this way, not only the distribution form of data can be preserved, but also outliers can be processed and differences between different features can be eliminated.[5].

Dimensionality reduction: Using PCA algorithm to reduce the dimensionality of data to 3 principal components, PCA can map the original high-dimensional data to a new low-dimensional space through linear transformation, and try to retain most of the variance in the original data. Since odor drift data has many features, PCA can effectively reduce the dimensions and remove redundant information, thus reducing the complexity of storage and calculation[6].

PCA(n_components=3).fit(X) library function is used to reduce the dimensionality of the data. PCA measures the correlation between features by calculating the covariance matrix of the original data. The eigenvalues of covariance matrix are decomposed to obtain the eigenvalues and corresponding eigenvectors. The feature vectors are sorted according to the size of the feature values to find the most informative feature, and the number of principal components to retain (here, 3 principal components are retained) or the percentage of the population variance is selected according to the requirements. A transformation matrix of selected principal components is used to map the original data set to the new space after dimensionality reduction[7].

Label: Add labels according to the classification, and finally store the pre-processed data into the table dataset_pca.csv, which is convenient for classification prediction through machine learning.

3. Data modeling and prediction

3.1 Correlation index

To avoid affecting the data observation, set the random seed value to 10 to ensure that the results can be reproduced. Read the data from the table dataset_pca.csv that you just stored, remove the classification labels, extract the labels, and then divide the training set and the test set. Finally, the evaluation index is set as accuracy, the fold number of cross-validation is set to 10, and the data is initialized based on stratified sampling cross-validation.

3.2 KNN model

The KNN(K-nearest neighbor) model is established. Specifying the number of KNN neighbors to be 1, which is sensitive and can more accurately adapt to small or irregular features in the data set. In the KNN algorithm, for a given new sample, it will

find the nearest K neighbors and make prediction or classification based on the labels of these neighbors [8]. KNN model makes classification prediction, calculates accuracy and generates classification report. You can see that the accuracy of the prediction is 94.16%.The relevant Precision, Recall, F1 value and Support are shown in Table 3.

Table 3. KNN model performance

No.	Precision	Recall	F1-score	Support
1	0.94	0.96	0.95	653
2	0.98	0.98	0.98	733
3	0.98	0.98	0.98	401
4	0.89	0.87	0.88	504
5	0.97	0.95	0.96	741
6	0.86	0.89	0.87	446

3.3 Decision tree model

A decision tree model (DecisionTreeClassifier) was established. To prevent overfitting, specify a depth of 3 for the tree. In the decision tree algorithm, a tree-like structure composed of nodes and directed edges was constructed, in which each internal node represented the judgment condition of a feature or attribute, and each leaf node represented a category label or predicted value [9]. The decision tree model makes classification prediction, calculates accuracy and generates classification report. You can see that the accuracy of the prediction is 91.15%.The relevant Precision, Recall, F1 value and Support are shown in Table 4.

Table 4. Decision tree model performance

No.	Precision	Recall	F1-score	Support
1	0.91	0.94	0.92	653
2	0.96	0.95	0.96	733
3	0.97	0.97	0.97	401
4	0.85	0.82	0.84	504
5	0.95	0.94	0.94	741
6	0.82	0.83	0.83	446

3.4 Random forest model

The Random Forest Classifier model is established. Set the number of decision trees to 100 to ease overfitting and improve the accuracy of the model. In the random forest classifier, the classification performance and generalization ability are improved by constructing multiple decision trees at the same time and synthesizing their prediction results [10]. The random forest classifier model is used to predict the classification, calculate the accuracy and generate the classification report. You can see that the accuracy of the prediction is 94.77%. The relevant Precision, Recall, F1 value and Support are shown in Table 5.

Table 5. Random forest performance

No.	Precision	Recall	F1-score	Support
1	0.95	0.95	0.95	653
2	0.98	0.97	0.98	733
3	0.98	0.98	0.98	401
4	0.91	0.91	0.91	504

5	0.96	0.96	0.96	741
6	0.88	0.89	0.89	446

3.5 Multi-layer perceptron model

In the multi-layer perceptron (MLP) model, the number of hidden layer neurons is set to 100, the activation function is set to ReLU, and the optimization algorithm is set to Adam. In the multi-layer perceptron, the classification problem is solved by constructing a forward propagation neural network with multiple hidden layers. Each hidden layer is composed of multiple neurons, each of which uses weighting and activation operations to make a nonlinear transformation of the input data. Through forward propagation, the input data is passed layer by layer to the output layer, and finally the prediction result or classification label is obtained. The multi-layer perceptron model makes classification prediction, calculates accuracy and generates classification report. You can see that the accuracy of the prediction is 80.04%.The relevant Precision, Recall, F1 value and Support are shown in Table 6:

Table 6. Multi-layer perceptron performance

No.	Precision	Recall	F1-score	Support
1	0.81	0.75	0.78	653
2	0.87	0.89	0.88	733
3	0.98	0.96	0.97	401
4	0.65	0.68	0.66	504
5	0.88	0.87	0.87	741
6	058	0.61	0.60	446

4. Test and analysis

Through testing, the odor drift data set was analyzed to master the 6 gas categories of the data set and the distribution of sample number. It is found that the number of data sets is large, the sample distribution is relatively average, there is a certain correlation between some features, the quantity changes of each gas are similar, and the monthly distribution of characteristic data is relatively regular. However, the numerical difference between different data features is too large, and there are some outliers in the data set, so normalization and dimensionality reduction are needed.

The data is normalized, the value range is mapped to between 0 and 1, and then the principal components are reduced to 3 by PCA dimensionality reduction, eliminating the differences between different features and preprocessing for better classification prediction.

Through KNN model, decision tree model, random forest classifier model and MLP multi-layer perceptron model, the model performance was evaluated by cross-validation of data sets, and the model was trained by training sets. The accuracy of prediction was shown in Table 7.

Table 7. Comparison of accuracy of several models

No.	Model name	Accuracy rate (%)
1	KNN	94.16
2	Decision tree	91.15
3	Random forest	94.77
4	MLP multi-layer perceptron	80.04

Through the analysis of test results in Table 7, it is found that the prediction accuracy of random forest model is the highest, reaching about 95%, which also indicates that the preprocessing process of odor drift data has the best effect. The advantage of random forest is to make predictions by integrating multiple decision trees, each of which is trained on different random subsamples and takes into account different features, which enables the model to make full use of the information of the data, reduce overfitting problems, handle large-scale and high-dimensional data, and be insensitive to missing values and outliers[11]. In the future, the parameters and algorithms of random forest will be further studied to make it more suitable for gas drift data sets.

5. Conclusions

In this paper, relevant machine learning classification algorithms for odor drift data are studied. First, the relevant contents of the data set are understood, such as sample quantity, type, characteristic indicators, etc., and the data distribution rule, coherence, outlier noise points and other information of the data set are mastered. The normalization and dimensionality reduction of the data are preprocessed, the training set and the test set are divided, and the KNN model, decision tree model, random forest classifier model and MLP multi-layer perceptron model are used to conduct relevant classification tests on the data set. The results show that the random forest model has a relatively good performance for odor drift data classification. This paper can further study and compare more reasonable algorithms to achieve more accurate classification prediction.

References

[1] R. Zhuo, Y. Guo and B. Guo, A Hyperspectral Image Classification Method Based on 2-D Compact Variational Mode Decomposition. *IEEE Geoscience and Remote Sensing Letters*, 2023, 20(19):1-5.
[2] Shah, S.-C. Design of a Machine Learning-Based Intelligent Middleware Platform for a Heterogeneous Private Edge Cloud System. *Sensors*, 2021, 21(20):7701.
[3] Luo Yu Gas sensor drift compensation based on deep belief networks [D]. *Chongqing University*, 2017.
[4] Ziyatdinov A, Marco S, Chaudry A, et al. Drift compensation of gas sensor array data by common principal component analysis[J]. *Sensors & Actuators: B. Chemical*, 2011,146(2):460-465.
[5] Yang Hanyu, Zhao Xiaoyong, Wang Lei Overview of Data Normalization Methods [J]. *Computer Engineering and Applications*, 2023,59 (03): 13-22.
[6] Guo Husheng, LIU Yanjie, WANG Wenjian. Concept drift processing of streaming Data based on Mixed Feature Extraction [J]. *Computer Research and Development,2023,8(7)*:12-19.
[7] Luo Jun Research on Image Recognition Algorithms Based on PCA and LDA [D]. *Huazhong University of Science and Technology*, 2020.
[8] Chemicals and Chemistry - Chemical Sensors; Research from University of Brescia Has Provided New Data on Chemical Sensors (k-NN and k-NN-ANN Combined Classifier to Assess MOX Gas Sensors Performances Affected by Drift Caused by Early Life Aging)[J]. *Chemicals & Chemistry*, 2020, 12(11):1-7.
[9] Ge Peng, Peng Mengjing Application and Optimization of Decision Tree Algorithm in Classification Prediction [J]. *Journal of Shandong Agricultural University* (*Natural Science Edition*), 2016,47 (06): 936-939.
[10] Research on Cost-sensitive random forest classification algorithm for High dimensional unbalanced data [D]. *Xidian University*, 2020.
[11] Xu Shaocheng. Research on High-dimensional unbalanced data classification based on Random forest [D]. *Taiyuan University of Technology*, 2018.