# Vinegar Classification Based on Feature Extraction and Selection From Tin Oxide Gas Sensor Array Data

**Zou Xiaobo**[*]**, Zhao Jiewen, Wu Shouyi and Huang Xingyi**

School of Biological and Environmental Engineering, Jiangsu University, Zhenjiang, Jiangsu 212013, China

* Author to whom correspondence should be addressed. E-mail: zou_xiaobo@ujs.edu.cn

**Abstract:** Tin oxide gas sensor array based devices were often cited in publications dealing with food products. However, during the process of using a tin oxide gas sensor array to analysis and identify different gas, the most important and difficult was how to get useful parameters from the sensors and how to optimize the parameters. Which can make the sensor array can identify the gas rapidly and accuracy, and there was not a comfortable method. For this reason we developed a device which satisfied the gas sensor array act with the gas from vinegar. The parameters of the sensor act with gas were picked up after getting the whole acting process data. In order to assure whether the feature parameter was optimum or not, in this paper a new method called "distinguish index"(DI) has been proposed. Thus we can assure the feature parameter was useful in the later pattern recognition process. Principal component analysis (PCA) and artificial neural network (ANN) were used to combine the optimum feature parameters. Good separation among the gases with different vinegar is obtained using principal component analysis. The recognition probability of the ANN is 98 %. The new method can also be applied to other pattern recognition problems.

**Keywords:** Gas sensor array; Feature extraction; Principal component analysis; Neural network; Vinegar; Electronic nose

## Introduction

Traditionally, human sensory panels (group of people with highly trained senses of smell), gas chromatography (GC), and mass spectrometry (MS) have been used to analyze food odors. The

disadvantages of human sensory panels include subjectivity, poor reproducibility (i.e., results fluctuate depending on time of day, health of the panel members, prior odors analyzed, fatigue, etc.), time consumption, and large labor expense. Also, human panels can not be used to assess hazardous odors, work in continuous production, or remote operation. GC and GC/MS systems can require a significant amount of human intervention to perform the analysis and then relate the analysis to something useable[1]. The main motivation for tin oxide gas sensor array based devices is the development of a qualitative, low-cost, real-time, and portable method to perform reliable, objective, and reproducible measures of volatile compounds and odors. In the past these devices (electronic noses) have been developed for the classification and recognition of a large variety of foods, such as juices[2], coffee[9] meats [4,7,10], fishes[12], cheese[3], spirits[1],wines[5,6,8],and fruits[11].

In many applications for chemical sensors, information can be gained not only from a steady-state value of the sensor signal, but also from the kinetics of the response. However, using steady-state sensor value to classify different mixture gases results in losing many information of the sensor signal. Few articles mention the advantage of the transient signal when classifying flavors [1 15]. And there was not a comfortable method. Therefore, the purpose of this work is to show how to extract parameters containing information from an array of sensors (feature extraction), a good working method to determine which of the features are the most important (feature optimization). In this paper, a gas sensor system designed to perform vinegar analysis is introduced, and its application aiming at the classification of two different type vinegars named as 'Chinkang Vinegar' and 'Sanxi Vinegar', which are the most saleable vinegars in China.

## Experiments

The electronic nose (Fig.1) can identify and quantify chemical vapors. The system is composed of a 12 bit AD/DA converter, an air filter for suppressing humidity, a suction pump, and a personal computer. The chemical sensor array employs an array of five tin-oxide gas sensors, a humidity sensor and a temperature sensor to examine the environment. Although each sensor is designed for a specific chemical, each responds to a wide variety of chemical vapors. Collectively, these sensors respond with unique signatures (patterns) to different chemicals[17].
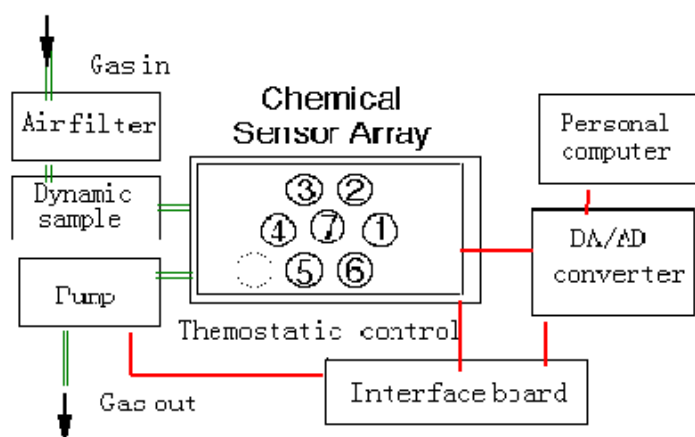


**Figure 1.** Schematic diagram of the electronic nose.

   The five tin-oxide sensors are commercially available Taguchi-type gas sensors obtained from Figaro Co. Ltd. (Sensor 1, TGS 813; Sensor 2, TGS 880; Sensor 3, TGS 822; Sensor 4, TGS 825; Sensor 5, TGS 812). These sensors are heated to a constant temperature, holding the sensor heater voltage at 5V. The humidity sensor (Sensor 6: HS-01) and the temperature sensor (Sensor 7: Pt100) are used to monitor the conditions of the experiment. The head space sample is injected in the 1000ml thermostatically controlled measurement cell in a dynamic way. In the dynamic mode, the gas sample is conveyed to the measurement cell by a carrier gas. This gas is the atmospheric air, thermostatically controlled, filtered on active charcoal and dehydrated with silica gel. Its flow-rate is controlled at 500ml/nim, either for cleaning the measurement cell or for the dynamic injection. Exposure of a tin-oxide sensor to vapor produces a change in its electrical resistance [16].

   The system has been trained to identify the two different type vinegars named as 'Chinkang Vinegar' and 'Sanxi Vinegar', which are the most saleable vinegars in China. In order to generate one dynamic dead space, 10ml of liquid sample is drawn form one of the vinegars and injected into a 250ml thermostatically controlled cell; The headspace is generated over 10 min. Then the carrier gas conveyed the gas sample to the measurement cell. During operation the sensor array "smells" the gas from the dynamic head-spaces of one vinegar, the sensor signals are digitized and fed into computer, and the whole signal is exploited, from the absorption beginning to the stationary phase of equilibrium between reversible adsorption and desorption, the process lasts 150s. At last, we use the carrier gas cleaning the measurement cell 8 min until the sensors is recovered. Each vinegar measurement was repeated several times in order to obtain accurate and reliable data. Typical response curve for the gas sensor array reaction, the curve is smoothed and the baseline is subtracted. Here we cite the experiment of the sensors response to 'Chinkang Vinegar'. ( Fig.2).
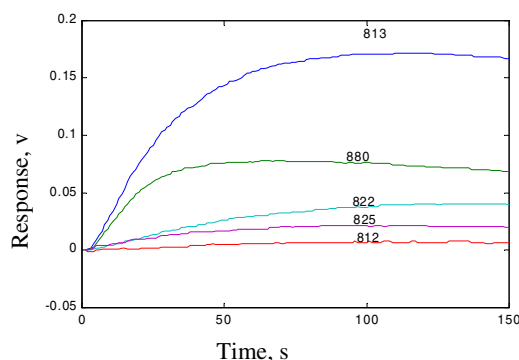


**Figure 2.** Typical response curve for the gas sensor array reaction.

## Feature Extraction and Selection

   In order to utilize all information from a time-developing system, it is possible either to use all the data points in the analysis, or to find some features (typically much fewer than the number of data points) that makes it possible to represent all the information in the measurements. The features can be picked manually[1], or by making an ordinary function approximation if the expected mathematical behavior is known[15].If too many features are used for the classification, there is a risk that the model

gets too complex, and the generalization capability of the model (i.e. the ability to correctly classify new data) can then be very poor. It is therefore useful to reduce the number of features in the model by determining which of the features contain most necessary information to distinguish between the different classes. When this is made, the problem of finding a good model for the classification is rather easy, and what model type (e.g. partial least squares or artificial neural network) to use is easy to determine. In this paper we introduce known concepts from statistics and control theory, and show their applicability to measurements with a gas sensor array in order to find a rather quick and easy way to classify different common types of vinegar.

It is well known that each sensor responds to different chemical vapors at different rate and value. Therefore, from each curve, 4 features are extracted (fig.3). They are the slope max (kmax), maximum (max), average of the last 20 points (st) and the average of whole points (mean) of curve. Table1 show the represent meaning of the four extracted features. Then 20 features were extracted from 5 sensors curve. The measure of the goodness of the parameters was then used in follow performance criteria.
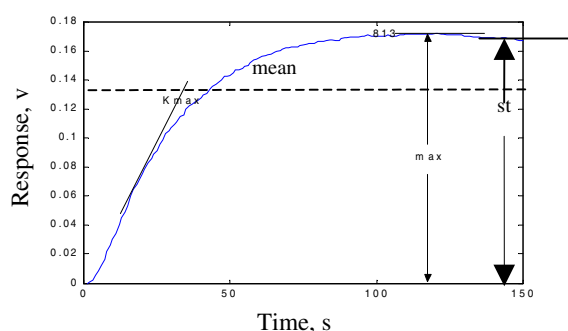


**Figure 3.** Sensor TGS813 smoothed curve and the features extracted are shown on the curve.

**Table 1.** The represent of the 4 extracted features.

| Extracted feature | Represent meaning |
|---|---|
| Max slope (kmax) | The respond rate of sensor to different vinegar gas |
| Maximum (max) | The maximum respond value |
| Average of last 20 points (st) | The stationary phase of equilibrium between reversible adsorption and desorption |
| Average of whole points (mean) | Sensor respond value during the whole process |

*Performance Criteria*

Formally, in classification processing repeatability and discriminant distance between classes are used to quantify feature or sensor performance. However, in this paper, the methodology of out put feature selection is based on calculate the distinguish index D.I. of each feature parameter. The D.I. of a feature parameter, which will be used to distinguish two states, such as 'Chinkang Vinegar' or 'Sanxi Vinegar', is derived in the following way.

For distinguishing two states (state 1 and state 2), the failure distinction ability of feature parameter can be evaluated by the "Distinction Rate (D.R.) $P_0$" [16] defined in the following formula:

$$p_0 = \int_{R_i} f_i(x)dx \qquad i = 1, 2. \tag{1}$$

Here, $f_i(x)$ is the probability density function measured in the state i, $R_i$ is decided by the following formula:

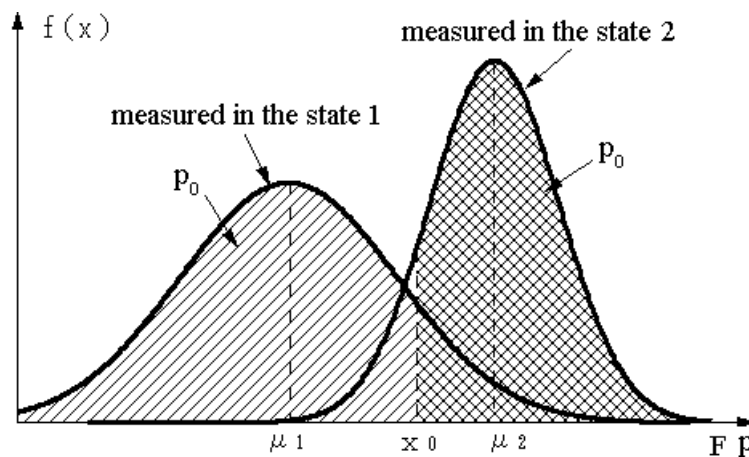$$\int_{R_1} f_1(x)dx = \int_{R_2} f_2(x)dx \tag{2}$$

For example, when $f_i(x)$ is the standard density function, $R_i$ ($-\infty \perp x_0$ $x_0 \perp \infty$) can be derived as follows:

$$\frac{1}{\sqrt{2\pi}\sigma_1} \int_{-\infty}^{x_0} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} dx = \frac{1}{\sqrt{2\pi}\sigma_2} \int_{x_0}^{\infty} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} dx \tag{3}$$

Here $\mu_1$ and $\mu_2$ are the mean values of the feature parameters calculated by the signals measured in state 1 and state 2. $\sigma_1$ and $\sigma_2$ are their standard deviations. $x_0$ can be worked out as follows:

$$x_0 = \frac{\mu_1 \sigma_2 + \mu_2 \sigma_1}{\sigma_1 + \sigma_2} \tag{4}$$



**Figure 4.** An example of $x_0$ and $p_0$.

Fig. 4 shows $p_0$ and $x_0$. With the substitution z= x-$\mu_1$ /$\sigma_1$ or z= x-$\mu_2$ /$\sigma_2$ to fo r m u lae (3), (4), the "Distinction Rate $P_0$" can be obtained in following way:

$$p_0 = \frac{1}{2\pi} \int_{-D.I.}^{\infty} e^{-\frac{z^2}{2}} dz \tag{5}$$

or

$$p_0 = \frac{1}{2\pi} \int_{-\infty}^{D.I.} e^{-\frac{z^2}{2}} dz \tag{6}$$

Here, D.I. is called "Distinction Index" and calculated by the following formula:

$$D.I. = \frac{\mu_2 - \mu_1}{\sigma_2 + \sigma_1}; \qquad \mu_2 > \mu_1 \tag{7}$$

It is obvious that the larger the value of D.I., the larger the value of "Distinction Rate, $P_0$", and therefore, the better the feature parameter will be. So D.I. can be used as the performance criteria of feature parameter selection.

## Results and Discussion

The new method discussed here has been used to distinguish between 'Chinkang Vinegar' and 'Sanxi Vinegar'. The D.I. and D.R.($P_0$) of each FP defined as sensor's value are shown in Table2. Table 2 shows that the D.I.'s are less than 1.6 and D.R'($P_0$'s) are less than 92%. Consequently each FP is not good enough to distinguish between 'Chinkang Vinegar (CV)' and 'Sanxi Vinegar (SV)'. The sensitivity levels vary from one sensor to another, the feature is quite similar, whatever the sensor. Table2 shows that the feature maximum points (*max*) shows the best and the max slope (*Kmax*) shows the poorest for each sensor. We selected 10 optimum feature according to their D.I.: *max1, max2, max3, st1, st2, mean2, max4, max5, mean1* and *st3*. Fig.5 exhibits the results of principal component analysis (PCA) for the two vinegars with these 10 features. PCA is a simple method to project data from several FP to a three-dimensional space. The values of 86.66% of 1-axis (Fig.3 / x-axis), 5.65% of 2-axis (Fig.3 / y-axis) and 1.39% of 3-axis (Fig.3 / z-axis) indicate contribution rate to pattern separation. It shows that the pattern separation is not sharp.

**Table 2.** D.I. and D.R. of 20 feature parameters.

| Feature parameter | TGS813 | | | | TGS880 | | | |
|---|---|---|---|---|---|---|---|---|
| | Max 1 | St1 | Mean 1 | Kmax 1 | Max2 | St2 | Mean 2 | Kmax 2 |
| D I | 1.531 | 1.416 | 1.217 | 0.791 | 1.501 | 1.384 | 1.373 | 0.816 |
| D R | 91.8 | 88.5 | 85.0 | 75.1 | 91.5 | 88.1 | 87.5 | 77.3 |
| Feature parameter | TGS822 | | | | TGS825 | | | |
| | Max 3 | St3 | Mean 3 | Kmax 3 | Max4 | St4 | Mean 4 | Kmax 4 |
| D I | 1.424 | 1.158 | 0.976 | 0.074 | 1.352 | 0.930 | 0.859 | 0.002 |
| D R | 89.3 | 84.9 | 83.7 | 54.6 | 86.9 | 82.6 | 80.5 | 50.5 |
| Feature parameter | TGS812 | | | | | | | |
| | Max 5 | St5 | Mean 5 | Kmax 5 | | | | |
| D I | 1.321 | 1.110 | 0.622 | 0.005 | | | | |
| D R | 86.1 | 84.5 | 73.2 | 52.0 | | | | |

The ten optimum features that were selected by the D.I. were also used in an artificial neural net network.. But before the ten features were transmitted into the input layer, they need to be normalized. Because the ten features were coming from the same example, we can use the method that can change general normal distribution into standardized normal distribution [18,19,20]. Such as two different feature distributions, we separately noted them as follows:

$$T_1 \subset N(\mu_1, \sigma_1) \qquad T_1 \subset N(\mu_2, \sigma_2) \qquad (8)$$

$$T_1^{'} = \frac{T_1 - \mu_1}{\sigma_1} \qquad T_2^{'} = \frac{T_2 - \mu_2}{\sigma_2} \qquad (9)$$
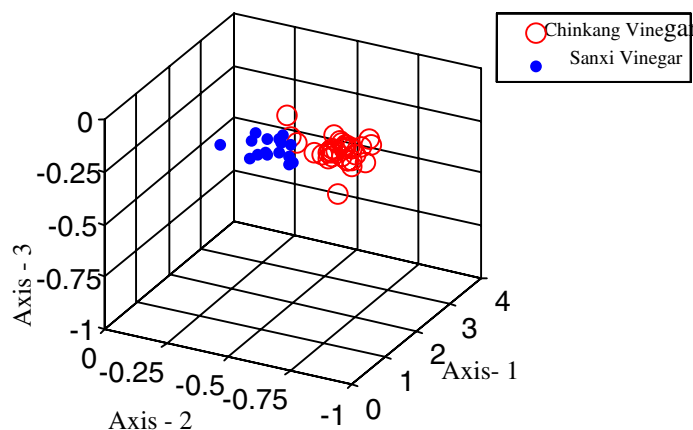


**Figure 5.** Results of the PCA of the gas sensor array for 40 vinegar gas samples.

Then $T_1^{'}$ and $T_2^{'}$ both are stochastic variable with $N(0,1)$ distribution. Then we can use formula (9) to make unified normalization for the ten different features. Therefore, we got *max1', max2', max3', st1', st2', mean2', max4', max5', mean1'* and *st3'* substitute for the ten optimum features transmitted into the input layer. The structure and parameters of neural network have been described in Fig.6 and Table 3. Two different vinegars were used as the output layer. The network was trained using data so that the desired outputs could be obtained. The connections between hidden and both input and output layers were optimized after 15,000 times training for the two vinegar samples. Fig.6 illustrates both the ten normalized optimum features and the ANN classification of the system for the two of test vinegars presented to the system. The recognition probability of the neural network analysis, defined as the ratio of the number of right answers to that of total trials was 98%.

**Table 3.** ANN training parameters.

| Type: | Backpropagation in batch mode |
|---|---|
| Architecture: | 10-8-2 feedforward |
| Activation: | Logistic |
| Learning Rate: | 0.01 |
| Momentum: | 0.2 |
| No. of Epochs: | 15000 |

As a comparison, a neural network was also trained using all the 20 features extracted from 5 sensors curve. The recognition probability, using the same validation method as for the network above, was then 90%, which is significantly lower than for the network with only the parameters chosen by the D.I. method. This is probably due to the fact that a network with many parameters requires a much larger sample data set to be able to fit the network parameters without over-training.
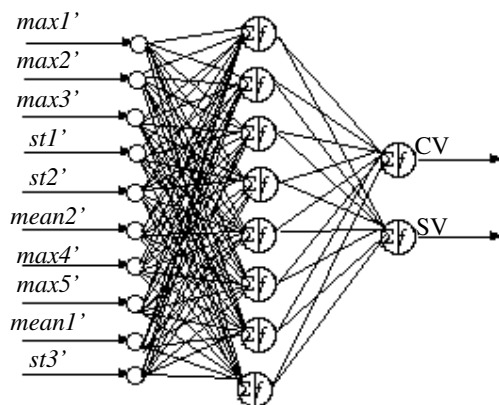


**Figure 6.** Vinegar classification with ANN.

## Conclusions

This stud presents a methodology of feature extraction and selection from the out puts of a tin oxide gas multi-sensor array. Transient signals were acquired in dynamic experimental conditions. From these curves, 20 features were extracted and sorted by D.I. 10 better features were selected in the later pattern recognition process. Principal component analysis (PCA) and artificial neural network (ANN) were used to combine the optimum feature parameters. Good separation among the gases with different vinegar is obtained using principal component analysis. The recognition probability of the ANN is 98%. The new method can also be applied to other pattern recognition problems.

## Acknowledgements

## References

1. Gardner, J.W.; Bartelet, P.N. *Electronic nose: Principles and Applications*. Oxford University Press 1999. 1-4, 185-207.
2. Christophe, S. et al. Potential of semiconductor sensor arrays for the origin Authentication of pure valencia orange Juices. *Journal of Agricultural & Food Chemistry* **2001**, *49*, 3151-3160.
3. Giuseppe, Z. etal., Determination of organic acids, sugars, diacetyl, and action in cheese by high performance liquid chromatographic. *Journal of Agricultural & Food Chemistry* **2001**, *49*, 2722-2726.

4.  Vernat-Rossi, V.; Garcia, C.; Talon, R. et al., Rapid discrimination of meat products and bacterial strains using semi-conductor gas sensors. *Sensors and Actuators B* **1996**, *37*, 43-48.

5.  Natale, C.D.; Davide, F.A. Amico, M.A. et al. An electronic nose for the recognition of the vineyard of a red wine. *Sensors and Actuators B* **1996**, *33*, 83-88.

6.  Nanto, H.; Tsubakino, S.; Ikeda, M. et al. Identification of aromas from wine using quartz-resonator gas sensors in conju-ction with neural-network analysis. *Sensors and Actuators B* **1995**, *24-25*, 794-796.

7.  Funazaki, N.; Hemmi, A.; Ito, S. et al. Application of semiconductor gas sensor to quality control of meat freshness in food industry. *Sensors and Actuators B* **1995**, *24-25*, 797-800.

8.  Natale, C.D.; Davide, F.A.; Amico, M.A.D. et al. Complex chemical pattern recognition with sensory array. The discrimination of vintage years of wine. *Sensors and Actuators B* **1995**, *24-25*, 801-804.

9.  Singh, S.; Hines, E.L.; Gardner, J.W. et al. Fuzzy neural computing of coffee and tainted-water from an electronic nose. *Sensors and Actuators B* **1996**, *30*, 185-190.

10. Bourrounet, B.; Talou, T.; Gaset, A**.** Application of a multi-gas-sensor device in the meat industry for boar-taint detection. *Sensors and Actuators B* **1995**, *26-27*, 250-254.

11. Molto, E.; Selfa, E. An Aroma Sensor for Assessing Peach Quality. *Journal of Engineering Research* **1999**, *72*, 311-316.

12. Gardner, J.W.; Bartlett, N. A brief history of electronic nose. *Sensors and Actuators B* **1994**, *18-19*, 211-220.

13. Huyberechts, G. Simultaneous quantification of carbon monoxide and methane in humid air using a sensor array and an artificial neural networks. *Sensors and Actuators B* **1997**, *45*, 123-130.

14. Weiping, Y.; Chengfen, D. et al.The study of gas sensor array signal processing with improved BP algorithm. *Sensors and Actuators B* **2000**, *66*, 283-285.

15. Holmberg, M.; Gustafsson, F. et al. Bacteria classification based on feature extraction from sensor data. *Biotechnology Techniques* **1998**, *4*, 319-324.

16. Zou, X.B.; Zhao, J.W. The study of sensor array signal processing with new genetic algorithms. *Sensors and Actuators B* **2002**, *87*, 437-441.

17. Zou, X.B.; Wu, S. Evaluating the quality of cigarettes by an electronic nose system. *Journal of Testing and Evaluation* **2002**, *30(12)*, 6.

18. Zou, X.B. Study of electronic nose on simulation inspecting the process of ferment in the brewage of vinegar (in Chinese). Master's degree thesis, Jiangsu University, Zhenjiang, Jiangsu , China, 2001, 3.

19. Laurent, R.; Tekin, K.; Thomas, M. etc. A comparative study of signal processing technique for clustering microsensor data (a first step towards an artificial nose). *Sensors and Actuators B* **1997**, *41*, 105-120.

20. Wang P. Artificial nose and Artificial tongue (in Chinese). Science book concern. Beijing, China, 2000, 12.

*Sample Availability:* Available from the authors.