Chemometrics and Intelligent Laboratory Systems 133 (2014) 17–24

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Comparison of spectral clustering, *K*-clustering and hierarchical clustering on e-nose datasets: Application to the recognition of material freshness, adulteration levels and pretreatment approaches for tomato juices

Xuezhen Hong [a], Jun Wang [a,*], Guande Qi [b]

[a] Department of Biosystems Engineering, Zhejiang University, 688 Yuhangtang Road, Hangzhou 310058, PR China
[b] Department of Computer Science, Zhejiang University, Hangzhou 310027, PR China

ARTICLE INFO

Article history:
Received 24 November 2013
Received in revised form 19 January 2014
Accepted 25 January 2014
Available online 8 February 2014

Keywords:
Spectral clustering
K-clustering
Hierarchical cluster analysis
Cluster validation
Electronic nose
Tomato juice

ABSTRACT

Various clustering algorithms have been developed since conventional hierarchical cluster analysis (HCA) and partitioning clustering algorithms have their own limitations and scopes of applications. However, in the area of e-nose where clustering is applied, the conventional algorithms (mostly HCA) still play a dominant role. In addition, comparison among different clustering methods or validation of clustering results was seldom mentioned. In this paper, we present a state-of-the-art clustering method – spectral clustering – and compare it with six conventional clustering methods: *K*-clustering (ISODATA, FCM and *k*-means) and HCA (single linkage, complete linkage and Ward's). Three external validation criteria – mutual information criteria (MI), precision and rand index (RI) – were used to evaluate clustering performances on three independent e-nose datasets. The spectral clustering outperforms with statistical significance (alpha = 0.05) the performance of other methods, and the single linkage presents the worst (unacceptable) clustering result. In addition, the proposed approach – cluster validation criteria in combination with majority voting – in a way makes clustering a semi-supervised classification technique. Using this approach it is possible to compare clustering based semi-supervised methods with classification methods to find which method is better for discrimination of a certain e-nose dataset.

1. Introduction

The researches in electronic nose (e-nose) field have been focused on three main aspects: the developments of materials for sensors and sensor arrays, the optimizations and comparisons of multiple statistics and pattern recognition methods, and the combination of both sensor systems and analytical methods for various detecting tasks in food, cosmetic, and pharmaceutical industry as well as in environmental control and clinical diagnostics [1–3]. Successful applications of e-noses require not only sensors with excellent performances but also appropriate analytical methods.

Clustering is a fundamental data analysis task that groups a given collection of unlabeled data instances into meaningful clusters according to similarity (similar instances are grouped together while different instances belong to different groups). Clustering enables us to identify important relationships and structures within a dataset, thus allowing us to make predictions or discover hypotheses to account for the detected structure in the data. In addition, a more rational organization of information facilitates the subsequent step of supervised learning [4].

Various clustering algorithms have been developed [5]. In the aspect of e-nose data clustering, hierarchical cluster analysis (HCA) and partitioning clustering are mostly adopted [6]. The HCA could be further divided into the following subgroups according to the manner that the similarity measure is calculated: single linkage clustering (SL), complete linkage clustering (CL), between-groups linkage clustering, within-groups linkage clustering, centroid clustering and Ward's clustering etc. [7]. And variants of *K*-clustering such as *k*-means, ISODATA, fuzzy *c*-means (FCM) and partitioning around medoids (PAM) are the commonly used partitioning clustering methods [8].

It is widely acknowledged that the above conventional clustering methods have their own limitations and scopes of application. For example, the between-groups linkage, within-groups linkage and centroid method are sensitive to the shape and size of clusters, i.e., they can easily fail when clusters have complicated forms departing from the hyperspherical shape; and the *k*-means clustering, which is sensitive to noisy data and outliers, has linear complexity and works well on datasets having isotropic clusters [7]. Furthermore, different clustering methods – or even different configurations of the same algorithm – produce different partitions and none of them have proved to be the best in all situations [9]. A good approach would be to adopt different clustering methods and compare the results. Nevertheless, by examining the recent literature [10–33] about e-nose where CA is applied to the

* Corresponding author. Tel.: +86 571 88982178; fax: +86 571 88982191.
E-mail address: jwang@zju.edu.cn (J. Wang).

0169-7439/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2014.01.017

experimental data, we found that except Falasconi et al. [10] who compared clustering performances of five conventional clustering methods on e-nose datasets, most researchers [11–33] only adopted one conventional HCA or partitioning clustering method, with no comparison among different clustering methods being mentioned. A summary of the applications of CA for e-nose is presented in Table 1.

An important reason for the above two problems – lacking innovative clustering methods and missing comparison among different clustering methods – is the absence of cluster validation criteria. In most of the aforementioned cases, only the resulting dendrogram (represents the nested grouping of objects and similarity levels at which groupings change) was analyzed, while evaluation of clustering outcome (such as number of correctly clustered patterns) was seldom mentioned.

In this work, three cluster validation criteria were proposed for e-nose data. An innovative clustering algorithm – spectral clustering – was also employed. Recently, spectral clustering has been researched as a popular topic. By constructing an undirected weighted similarity graph on the data, spectral clustering utilizes the spectrum of the graph Laplacian to obtain a low dimensional representation of the data, and then does clustering using classical methods, such as *k*-means [34]. This graph-theoretic based clustering method is simple to implement, and it can be solved efficiently by standard linear algebra software and very often outperforms conventional clustering algorithms [35]. Applications of this method have been reported in language distinction [36], image segmentation [37], link prediction in biology and social networks [38], process monitoring [39], and tumor delineation [40] etc.

The main objectives of this research are: (1) to propose cluster validation criteria for quantification and evaluation of clustering results, (2) to compare among different clustering algorithms, and (3) to explore if the state-of-the-art spectral clustering would outperform conventional CA methods in the field of e-nose.

2. Experimental

2.1. Experimental datasets

In this work, three independent e-nose researches were taken, generating three independent e-nose datasets.

Chinese variety, *youbei* cherry tomatoes were picked three times for the experiments — tracing freshness of tomatoes that were squeezed for juice consumption, recognition of tomato juices with different adulteration levels and pretreatments, respectively. Thus, there were in total three independent e-nose datasets.

Dataset 1 (material freshness dataset) consists of six groups of juice samples. Light-red (approximately 70% of the surface, in the aggregate, shows pinkish-red or red) [41] cherry tomatoes were selected and stored in a refrigerator at 4 °C for 16 days. The e-nose measurements were conducted every three days (i.e. on days 1, 4, 7, 10, 13 and 16), resulting in six groups of e-nose data. 25 replications were prepared for each group, so the dataset 1 can be described as a 150 (25 replications × 6 groups) × 10 (e-nose sensors) matrix.

Dataset 2 (adulteration dataset) consists of seven groups of juice samples. Juices squeezed from fresh light-red cherry tomatoes were blended with the ones squeezed from overripe and decaying cherry tomatoes at seven levels of adulteration (from 0 to 30% (w/w) in steps of 5%). The seven groups were: 0% (100% fresh tomato juice), 5% (95 g of fresh tomato juice adulterated with 5 g of overripe tomato juice), 10% (90 g of fresh tomato juice adulterated with 10 g of overripe tomato juice), 15% (85 g of fresh tomato juice adulterated with 15 g of overripe tomato juice), 20% (80 g of fresh tomato juice adulterated with 20 g of overripe tomato juice), 25% (75 g of fresh tomato juice adulterated with 25 g of overripe tomato juice) and 30% (70 g of fresh tomato juice adulterated with 30 g of overripe tomato juice). 25 replications were prepared for each adulteration group, so the dataset 2 can be described as a 175 (25 replications × 7 groups) × 10 (e-nose sensors) matrix.

Dataset 3 (pretreatment dataset) consists of six groups of juice samples. Appropriate amount of light-red cherry tomatoes were pretreated by six different processes prior to being squeezed. The six pretreatments were as follows: control (non-treatment), freezing (freezing at −18 ± 1 °C during 16 h), low temperature blanching (60 °C, 3 min), high temperature blanching (90 °C, 1 min), microwave blanching (800 W, 2450 MHz of microwave oven, 30 s) and steam blanching (steam for 30 s). 25 replications were prepared for each treatment group, so the dataset 3 can be described as a 150 (25 replications × 6 groups) × 10 (e-nose sensors) matrix.

2.2. Apparatus and sampling procedures

For each research, the cherry tomatoes were placed in a fruit squeezer and juiced for 30 s to obtain juices. A PEN 2 e-nose (Airsense Analytics,

Table 1
Summary of main applications of clustering methods in the area of e-nose.

Content of study concerning CA application	Clustering methods	Ref.
Identification of Japanese green tea samples with different contents of coumarin	Between-groups linkage	[11–13]
Characterization of 17 Chinese vinegars		
Clustering of WO_3 thin-film sensors array		
Identification of spirits with strong internal similarities	Complete linkage	[14]
Discrimination of different types of damage of rice plants	Single linkage	[15,16]
Identification of quality grade of green tea		
Identification of wine grapes taken at different drying times	Ward's method	[17–23]
Cluster analysis of control blood, post- and pre-dialysis blood		
Discrimination between dermatophyte species and strains		
Clustering consumers into homogeneous groups according to the liking of tomatoes		
Screening of antifungal agents for efficacy against dermatophyte *Trichophyton* species		
Discrimination of odors from trim plastic materials used in automobiles		
Classification of blueberry fruit disease		
Clustering eleven aged cheddar cheeses	HCA (not specified)	[24–30]
Clustering five rice extrudate samples		
Detection of fungal contamination in library paper		
Optimization of chemiresistor sensor array		
Detection of microbial and chemical contamination of potable water		
Assess the abilities of different sensing layers to distinguish between analytes		
Early detection and differentiation of spoilage of bakery products		
Identification for five days of aroma pattern emitted by an encapsulated essence	PAM[a]	[31]
Optimization of the cross-selective sensor arrays	Fuzzy partitioning	[32]
Determination of features that produce the best clustering in a 30-dimensional space	Full-dimensional CA	[33]
Discussion of cluster validity issues for e-nose data	HCA and *k*-means	[10]

[a] PAM: partitioning around medoids.

GmBH, Schwerin, Germany) based on ten different metal-oxide semiconductors (MOS) was then used to test the squeezed juices. A description of the ten MOS has been given in our previous work [42]. During the e-nose measurement, each sample (10 mL of cherry tomato juice) was placed in a 500 mL airtight glass vial that was sealed with plastic wrap. The glass vial was closed for 10 min (headspace-generation time) at a room temperature of 25 ± 1 °C while the headspace collected the volatiles from the samples. During the measurement process, the headspace gaseous compounds were pumped into the sensor arrays through Teflon tubing connected to a needle in the plastic wrap, causing the ratio of conductance of each sensor changed. The measurement phase lasted for 70 s, which was long enough for the sensors to reach stable signal values. The signal data from the sensors were collected by the computer once per second during the measurements. Conductivity ratio G/G_0 (G and G_0 are the conductivities of sensors exposed to sample gas and zero gas, respectively) was recorded as the e-nose signal. When the measurement process was complete, the acquired data were stored for later use. After each experiment, calibration procedure was carried out to reduce the influence of external parameters such as variation in the relative humidity of the air, changes in the temperatures and the drift of the sensors over time, using zero gas (air filtered by active carbon).

3. Data analysis methods

3.1. Clustering algorithms

In this study, six conventional as well as one state-of-the-art clustering algorithms were presented. The seven clustering methods are as follows: agglomerative HCA (SL, CL and Ward's), *K*-clustering (FCM, ISODATA and *k*-means) and spectral clustering. Before performing a cluster analysis, it is necessary to consider scaling or transforming the variables since variables with large variances tend to have a larger effect on the resulting clusters than variables with small variances do. Meanwhile, as we mentioned before, different definitions of distance between instances may result in different clustering results. Thus, in this paper, the three e-nose datasets were all standardized prior to cluster analysis, and the best known and mostly used Euclidean distance was employed for all the clustering algorithms. The standardization was defined as the difference between the original responding value of each sensor and the mean value, divided by the standard deviation.

3.1.1. HCA

The HCA constructs clusters by recursively partitioning the instances in either a top-down or bottom-up fashion. In agglomerative HCA, each object initially represents a cluster of its own. While in divisive HCA, all objects initially belong to one cluster. Then the merging (agglomerative HCA) or division (divisive HCA) of clusters continues until the desired cluster structure is obtained [43]. SL, CL and Ward's clustering are three mostly used HCA methods.

The SL clustering (also called the connectedness, the minimum method or the nearest neighbor method) considers the distance between two clusters to be equal to the shortest distance from any member of one cluster to any member of the other cluster [5]. This method maintains good performance on datasets containing non-isotropic clusters. However, it has a drawback known as the “chaining effect”, i.e., a few points that form a bridge between two clusters would cause the SL clustering to unify these two clusters into one [6].

The CL clustering (also called the diameter, the maximum method or the furthest neighbor method) considers the distance between two clusters to be equal to the longest distance from any member of one cluster to any member of the other cluster [5]. This method is not strongly affected by outliers, but it can break large clusters and has trouble with convex shapes [44].

The Ward's clustering (also known as method of the minimum variance) searches similarity matrix for the most similar pair of clusters and reduces the number of clusters by merging the most similar pair of clusters. Objective of this algorithm is to find at each stage those two clusters whose merger gives the minimum increase in the total within group sum of square errors (or distances between the centroids of the merged clusters) [45,46]. The Ward's clustering has been widely used. However, it may cause elongated clusters to split and portions of neighboring elongated clusters to merge. In addition, it often falls into local optimum.

The structure obtained by hierarchical clustering is often presented in the form of a dendrogram, where each linkage step in the clustering process is represented by a connection line. The main disadvantages of HCA are inability to scale well and to undo what was done previously [5].

3.1.2. K-clustering

The partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning. Compared to HCA, partitioning methods are capable of back-tracking, but they require pre-set of the number of clusters by users.

k-Means and ISODATA are among the most popular, well-known “hard” partitioning methods, in which each point is assigned to only one particular cluster. The *k*-means starts with *k* cluster centers that are chosen at random or according to some heuristic procedure. In each iteration, each instance is assigned to its nearest cluster center, resulting in re-calculation of the cluster center. This process is repeated until a convergence criterion is met. The *k*-means is popular for its ease of interpretation, speed of convergence and adaptability [47]. However, this method is very sensitive to noise and outliers, and often falls into a local optimum on the sum-of-square error space. In addition, it does not guarantee unique clustering because the cluster centers are randomly chosen.

ISODATA is a modification of *k*-means that starts with a higher number of clusters. This algorithm permits splitting of clusters when a cluster variance is above a pre-specified threshold or merges them when distances between clusters are small, below another threshold [8]. However, it is difficult to find optimal parameters for ISODATA.

FCM, on the other hand, is a “soft” partitioning method that attempts to assign each instance to several clusters (depending on the degree of the fuzzy membership). The design of membership function is the most important problem for FCM [7]. Generally, FCM is better than the hard *k*-means method at avoiding local minima, but it can still converge to local minima of the squared error criterion.

3.1.3. Spectral clustering

Spectral clustering consists of two distinct stages: (a) construct an affinity graph from the data set and (b) cluster the data points through finding an optimal partition of the affinity graph [48]. The constructed affinity graph is an undirected graph $G(V, E, W)$, where $V = \{v_1, \dots, v_n\}$ represents the set of vertices, E represents the set of edges, and W is the associated affinity matrix. The edge $e_{i,j}$ between v_i and v_j carries a non-negative weight $w_{i,j}$, which represents the affinity between instance x_i and x_j. The affinity graph could be represented with a matrix:

$$W = \left[w_{i,j}\right] \tag{1}$$

where $w_{i,j}$ could be calculated using the Gaussian similarity function: $w_{i,j} = \exp(-\|x_i - x_j\|^2/(2\sigma^2))$.

Then, a graph Laplacian based on the similarity graph is defined as follows, and the eigenvector of the graph Laplacian is related to clustering:

$$L = D - W \tag{2}$$

where D is a diagonal matrix with $d_{i,i} = \sum_{j=1}^{n} w_{i,j}$.

Specifically, we use a spectral clustering algorithm according to [35]:

Input: raw data $x_1, \ldots, x_n$, number k of clusters

(1) Construction of graph: construct the k-nearest neighbor affinity graph and represent its weighted adjacency matrix as W.
(2) Computing Laplacian: compute the Laplacian L and the normalized Laplacian as $L_{sym} = D^{-1/2}LD^{-1/2}$.
(3) Finding eigenvector: compute the first k eigenvectors $u_1, \ldots u_k$ of L_{sym}. Let $U \in \boldsymbol{R}^{n \times k}$ be the matrix containing $u_1, \ldots u_k$ as the columns.
(4) Normalization: form the matrix $T \in \boldsymbol{R}^{n \times k}$ from U by normalizing the rows to norm 1: $t_{i,j} = u_{i,j}/(\sum_k u_{i,k}^2)^{1/2}$.
(5) Clustering: for $i = 1, \ldots, n$, let $y_i \in \boldsymbol{R}^k$ be the vector corresponding to the ith row of T. Cluster the points (y_i), $i = 1, \ldots, n$ into clusters $C_1, \ldots, C_k$ using the k-means algorithm.

Output: clusters $A_1, \ldots, A_k$ with $A_i = \{j | y_j \in C_i\}$.

3.2. Evaluation of clustering results — external validation criteria

Cluster validation criteria can be divided into two groups – internal and external criteria – according to whether external information is used or not. The former validate a partition by examining just the partitioned data, while the latter use the information of correct partition. Because e-noses are mostly applied for classification tasks with prior knowledge of the number of groups, in this paper, the number of clusters was set in accordance with the number of groups in the dataset, i.e., six clusters for the material freshness and the pretreatment datasets, respectively, and seven clusters for the adulteration dataset. Three external validation criteria – mutual information criteria (MI), precision and rand index (RI) – were used to examine whether the structure of clusters matches to some predefined classification of instances by comparing the actual clusters $C = \{C_1, \ldots, C_l, \ldots, C_k\}$ with the resulting $dom(y) = \{c_1, \ldots, c_h, \ldots c_k\}$ of clustering algorithm.

3.2.1. Mutual information based measure

The MI criterion is defined as follows:

$$C = \frac{2}{m} \sum_{l=1}^{g} \sum_{h=1}^{k} m_{l,h} \log_{g \times k} \left(\frac{m_{l,h} \times m}{m_{.,h} \times m_{l,.}} \right) \tag{3}$$

where $m_{l,h}$ represent the number of instances that are in cluster C_l and also in cluster c_h, m_h is the number of instances in the class c_h, and m_l is the number of instances in cluster C_l.

3.2.2. Precision

The precision criterion, which is calculated with the number of matches between C and c, is expressed as the number of correct matches M divided by number of instances n. This criterion is also known as the clustering accuracy. Calculation equation for precision is expressed as follows:

$$P = M/n = \sum_{h=1}^{k} \max_l |\{x_i | x_i \in c_h, x_i \in C_l\}|/n \tag{4}$$

where $\max_l |\{x_i | x_i \in c_h, x_i \in C_l\}|$ means that we match the clustering result c_h to actual clusters C_l by majority voting: if the majority of instances in c_h belong to C_l, then we define c_h is actually part of C_l.

3.2.3. Rand index

The rand index, which is calculated by considering each pair of instances, is defined as follows:

$$\mathrm{RI} = \frac{a + d}{a + b + c + d} \tag{5}$$

where a is the number of pairs that satisfy $x_i \in C_l, x_j \in C_l, x_i \in c_h, x_j \in c_h$, b is the number of pairs whose $x_i \in C_l$, $x_j \in C_l$, $x_i \in c_{h1}$, $x_j \in c_{h2}$, c is the number of pairs whose $x_i \in C_{l1}$, $x_j \in C_{l2}$, $x_i \in c_h$, $x_j \in c_h$, and d is the number of pairs which $x_i \in C_{l1}$, $x_j \in C_{l2}$, $x_i \in c_{h1}$, $x_j \in c_{h2}$. RI lies between 0 and 1. If the two partitions match perfectly, RI = 1; otherwise the more the partitions differ, the smaller the RI will be [10].

3.3. Software

Spectral clustering, ISODATA and FCM were performed in MATLAB R2008a. HCA (single linkage, complete linkage and centroid linkage) methods and k-means were performed in SPSS.

4. Results and discussion

4.1. A typical response curves of e-nose

A typical response mode of the PEN 2 e-nose to a tomato juice sample (from the adulteration group) as an example is presented in Fig. 1, and that to other samples is similar. The x-axis represents time, and the y-axis represents conductivity ratio G/G_0 values. Each curve represents the change of a sensor's ratio of conductance during measurement. As is shown in Fig. 1, the G/G_0 values for the ten sensors gradually changed (gradually increased or decreased) and finally reached stable equilibrium at the 70th second. The peak values (maximum or minimum) for each sensor were extracted as the original e-nose data for further analysis.

4.2. Clustering of the material freshness dataset (dataset 1)

The material freshness dataset consists of six classes of juices. Fig. 2a shows the visualization of the six classes in a 2D plot (containing 150 points in total, 25 points per class, and different classes are marked by different symbols and colors), where C1 to C6 represent the classes of days 1, 4, 7, 10, 13 and 16, accordingly. As shown in Fig. 2a, the six classes are discriminable. However, some data points from the strip-type C3 class (day 7) are close to the C1 class (day 1) or the C2 class (day 4). Meanwhile, it is noticeable that the C4 class (day 10) is close to the C6 class (day 16), and the C2 class is close to the C5 class (day 13). In view of the data distribution, it is foreseeable that different clustering methods may result in different cluster structures. Fig. 2b to h demonstrates the applications of spectral clustering, FCM, ISODATA, k-means, SL, CL and Ward's linkage clustering on the six classes (containing 150 points in total, different clusters are marked by different symbols and colors, and the number of points contained in each cluster may not be identical). The spectral clustering (Fig. 2b), whose resulting clusters are almost the same as the true classes, produces the best result; while the SL (Fig. 2g), whose resulting clusters differ a lot with the true classes, is totally meaningless. Except two data points from the day 7 cluster (green) are misclassified into the day 4 cluster (red), all the data points in Fig. 2b are correctly clustered. The FCM (Fig. 2c), ISODATA (Fig. 2d) and Ward's (Fig. 2h) clustering produce similarly good results: in the case of FCM, three data points from the

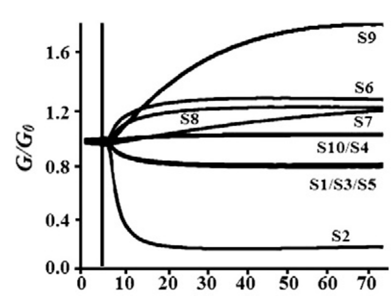

Fig. 1. A typical response of the PEN 2 e-nose to a freshly squeezed tomato juice sample (from the adulteration group).

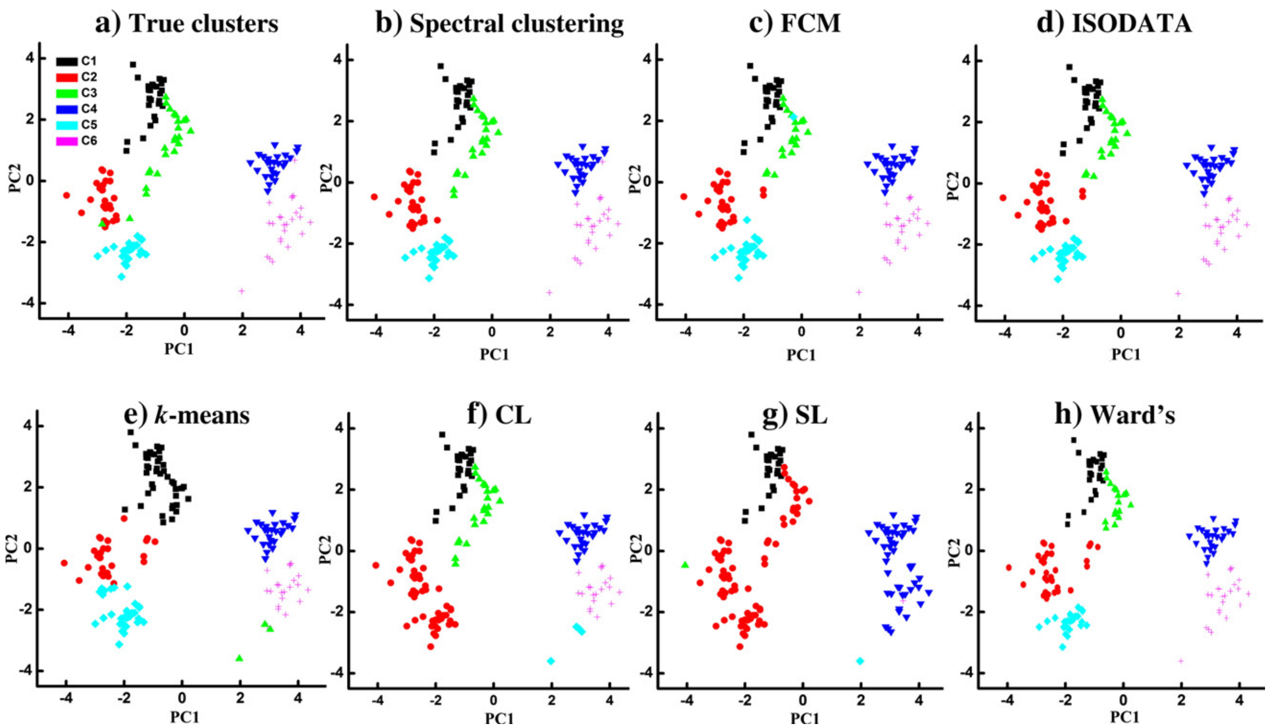

Fig. 2. Comparison among different clustering methods for clustering e-nose dataset of material freshness. (a) True clusters, (b) spectral clustering, (c) FCM, (d) ISODATA, (e) *k*-means, (f) complete linkage, (g) single linkage, and (h) Ward's linkage.

day 7 cluster (green) are misclassified into the day 4 cluster (red), one data point from the day 7 cluster is misclassified into the day 13 cluster (light-blue), and one data point from the day 16 cluster (pink) is misclassified into the day 10 cluster (dark-blue); in the case of ISODATA, four data points from the day 7 cluster (green) are misclassified into the day 4 cluster (red), and one data point from the day 16 cluster (pink) is misclassified into the day 10 cluster (dark-blue); and in the case of Ward's clustering, seven data points from the day 7 cluster (green) are misclassified into the day 4 cluster (red), and one data point from the day 16 cluster (pink) is misclassified into the day 10 cluster (dark-blue). However, the results of *k*-means (Fig. 2e) and CL (Fig. 2f) are not so good: in the case of *k*-means, the day 7 cluster is emerged into the day 1 (black) and the day 4 (red) clusters, while the day 16 cluster is divided into two clusters (marked by pink and green); in the case of CL, the day 4 and the day 13 clusters are merged together (marked by red), while the day 16 cluster is divided into two clusters (marked by pink and light-blue). Evaluation of clustering performance of the seven clustering methods using MI, precision and RI is presented in Table 2, where spectral clustering (0.5927, 0.9867 and 0.9914 for MI, precision and RI, respectively) outperforms the six conventional clustering methods and SL presents the worst performance (0.3593, 0.52 and 0.7789 for MI, precision and RI, respectively). It is noticeable that FCM and ISODATA share the same precision value. However, MI and RI of ISODATA are slightly higher. The seven clustering methods are sorted and listed sequentially from high values to low values in accordance with their quality criteria values as follows: spectral clustering, ISODATA, FCM, Ward's, CL, *k*-means and SL.

4.3. Clustering of adulteration dataset (dataset 2)

The adulteration dataset consists of seven classes of juices. Fig. 3a shows the visualization of the seven classes in a 2D plot (containing 175 points in total, 25 points per class, and different classes are marked by different symbols and colors), where C1 to C7 represent the seven classes of adulteration levels: 0, 5%, 10%, 15%, 20%, 25% and 30%, accordingly. As shown in Fig. 3a, the seven classes distribute sequentially from left to right in accordance with their adulteration levels. Though the seven classes are discriminable, each class is close to its neighboring classes. Meanwhile, four data points – one data point each from the C5 class (20%) and C6 class (25%) as well as two data points from the C7 class (30%) – are away from their own class centers. In view of the data distribution, it is foreseeable that different clustering methods may result in different cluster structures. Fig. 3b to h demonstrates the comparison of spectral clustering, FCM, ISODATA, *k*-means, SL, CL and Ward's linkage clustering on the seven classes (containing 175 points in total, different clusters are marked by different symbols and colors, and the number of points contained in each cluster may not be identical), where resulting clusters of spectral clustering (Fig. 3b) and FCM (Fig. 3c) are the most similar to the true classes. As shown in Fig. 3d to h, the results of ISODATA (Fig. 3d) and Ward's (Fig. 3h) are not very good, where the 0 and 5% classes are clustered similar to their true clusters but the 10%–30% classes are clustered a little different with their true clusters; the results of *k*-means (Fig. 3e) and CL (Fig. 3f) are even worse, where the aforementioned four far-away data points are clustered into one cluster while the 10%–30% clusters (five classes) are clustered into four clusters; the results of SL (Fig. 3g) is meaningless,

Table 2
Evaluation of spectral clustering, *K*-clustering and hierarchical clustering for material freshness dataset using three external validation criteria.

Clustering methods	MI[a]	Precision	RI[b]
Spectral clustering	0.5927	0.9867	0.9914
FCM	0.5693	0.9667	0.9789
ISODATA	0.5724	0.9667	0.9792
k-means	0.4471	0.7867	0.8988
CL[c]	0.5005	0.8133	0.9195
SL[d]	0.3593	0.52	0.7789
Ward's	0.5587	0.9467	0.9687

[a] MI: mutual information criterion.
[b] RI: rand index.
[c] CL: complete linkage.
[d] SL: single linkage.

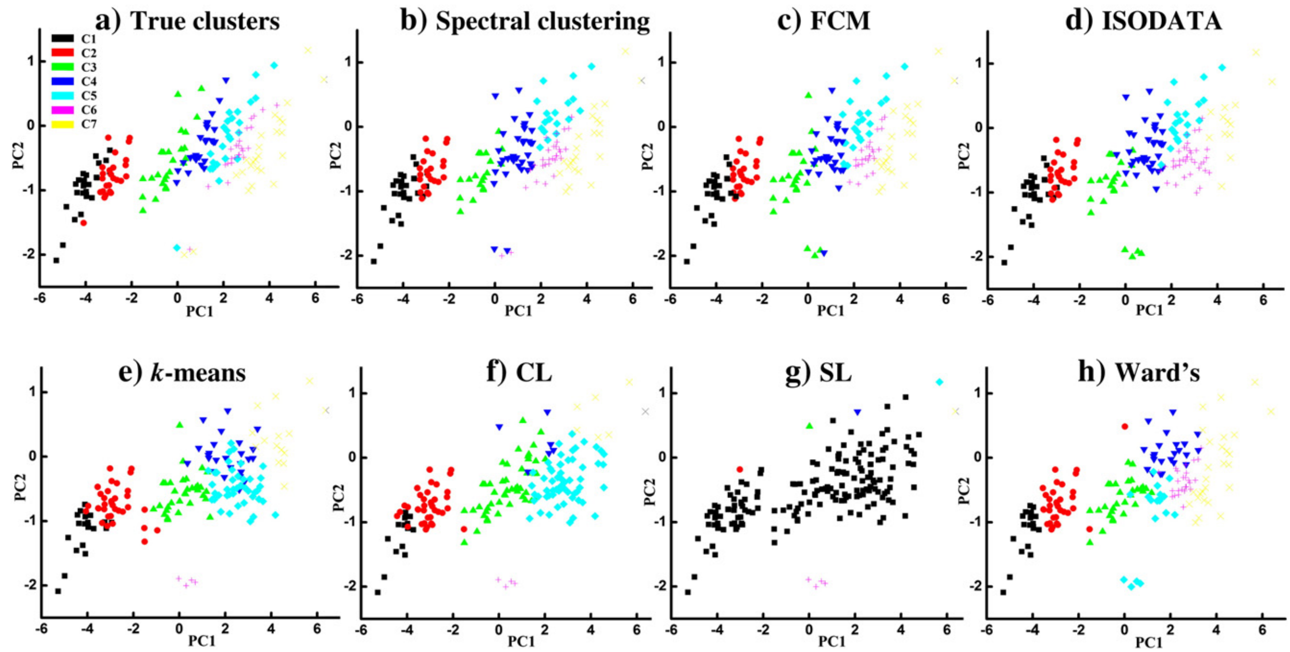

Fig. 3. Comparison among different clustering methods for clustering e-nose dataset of adulteration. (a) True clusters, (b) spectral clustering, (c) FCM, (d) ISODATA, (e) *k*-means, (f) complete linkage, (g) single linkage, and (h) Ward's linkage.

where the seven classes are mainly clustered into one cluster. Unlike in the case of storage shelf life dataset, incorrectly clustered data points here are not easy to be calculated intuitively. Evaluation of clustering performance of the seven clustering methods using MI, PR and RI is presented in Table 3, where spectral clustering (0.4167, 0.8171 and 0.9101 for MI, precision and RI, respectively) and FCM (0.4079, 0.8171 and 0.9113 for MI, precision and RI, respectively) have equally good performances that outperform the other five clustering methods. Again, the SL presents the worst performance (0.0066, 0.1829 and 0.2106 for MI, precision and RI, respectively). The seven clustering methods are listed sequentially from high values to low values in accordance with their quality criteria values as follows: spectral clustering/FCM, ISODATA, Ward's, CL, *k*-means and SL.

4.4. Clustering of pretreatment dataset (dataset 3)

The pretreatment dataset consists of six classes of juices. Fig. 4a shows the visualization of the seven classes in a 2D plot (containing 150 points in total, 25 points per class, and different classes are marked by different symbols and colors), where C1 to C6 represent the six classes of pretreatments: freezing, low temperature blanching, high temperature blanching, control, steam blanching and microwave blanching. As shown in Fig. 4a, the six classes are generally discriminable. However, four data points – one data point each from the C3 class (high temperature blanching) and C4 class (control) as well as two data points from the C2 class (low temperature blanching) – are away from their own class centers; meanwhile, the C1 (freezing), C3, C5 (steam blanching) and C6 (microwave blanching) classes are close to each other. In view of the data distribution, it is foreseeable that different clustering methods may result in different cluster structures. Fig. 4b to h demonstrates the applications of spectral clustering, FCM, ISODATA, *k*-means, SL, CL and Ward's linkage clustering on the seven classes (containing 150 points in total, different clusters are marked by different symbols and colors, and the number of points contained in each cluster may not be identical), where spectral clustering (Fig. 4b) and FCM (Fig. 4c) produce better results (resulting clusters of them are almost the same as the true classes) than the other five clustering methods. As shown in Fig. 4d and h, the results of ISODATA (Fig. 4d) and Ward's (Fig. 4h) are also acceptable: in the case of ISODATA, the C2 class is partitioned into two clusters while the C3 and C6 classes are merged into one cluster; in the case of Ward's, the aforementioned four far-away data points are clustered into the C2 class while the C3 and C6 classes are agglomerated into one cluster. However, the results of *k*-means (Fig. 4e), CL (Fig. 4f) and SL (Fig. 4g) are not so good: in the case of *k*-means, the C2, C3 and C6 classes are merged into one cluster, the C1 and C5 clusters are merged into one cluster, the aforementioned four far-away data points are clustered into one cluster, and the C4 class is partitioned into two clusters; in the case of CL, the C1, C3, C5 and C6 classes are merged into one cluster, and the aforementioned four far-away data points are clustered into two clusters; in the case of SL, except the C2 class, most of the other five classes are merged into one cluster. Evaluation of clustering performance of the seven clustering methods using MI, precision and RI is presented in Table 4, where spectral clustering (0.4901, 0.8933 and 0.9366 for MI, precision and RI, respectively) outperforms the other six clustering methods, and SL presents the worst performance (0.0126, 0.22 and 0.242 for MI, precision and RI, respectively). The seven clustering methods are listed sequentially from high values to low values in accordance with their quality criteria values as follows: spectral clustering, FCM, Ward's, ISODATA, CL, *k*-means and SL.

Table 3
Evaluation of spectral clustering, *K*-clustering and hierarchical clustering for adulteration dataset using three external validation criteria.

Clustering methods	MI	Precision	RI
Spectral clustering	0.4167	0.8171	0.9101
FCM	0.4079	0.8171	0.9113
ISODATA	0.3648	0.7371	0.8802
k-means	0.2688	0.5029	0.8206
CL	0.3075	0.52	0.8157
SL	0.0066	0.1829	0.2106
Ward's	0.3636	0.7029	0.8890

4.5. Summary of three datasets

In order to comprehensively and credibly compare the performances of spectral clustering, *K*-clustering and hierarchical clustering, the bootstrap resampling technique [49] was investigated to estimate

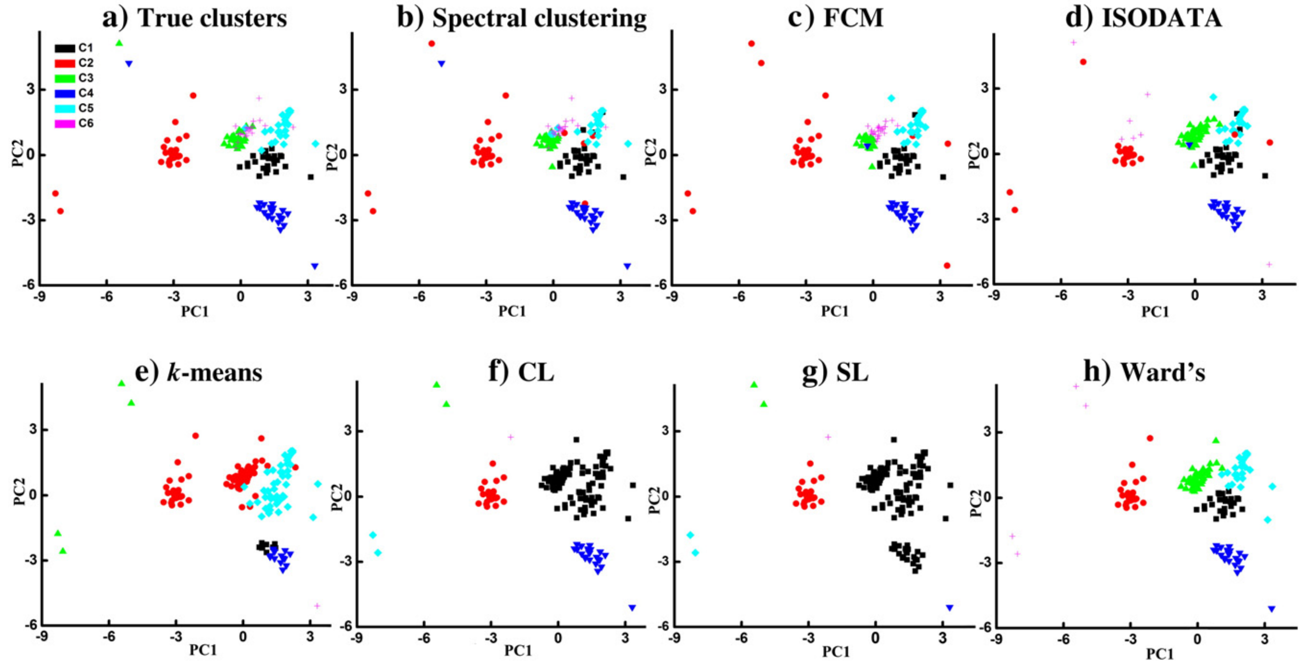

Fig. 4. Comparison among different clustering methods for clustering e-nose dataset of pretreatments. (a) True clusters, (b) spectral clustering, (c) FCM, (d) ISODATA, (e) *k*-means, (f) complete linkage, (g) single linkage, and (h) Ward's linkage.

the means of clustering accuracy for the three considered datasets. For each dataset, 200 of bootstrap samples were used, generating 200 of clustering accuracy for each clustering method. One-way analysis of variance (ANOVA) was employed to investigate if there was significant difference in clustering accuracy provided by different clustering methods. The result showed that the significance level p was <0.001, representing there was significant difference among the mean values of clustering accuracy based on different clustering methods. Tukey's multiple comparison was then applied to compare the mean values between any two of the seven clustering methods, and the results were listed in Table 5. In the case of material freshness, the spectral clustering outperforms with statistical significance (alpha = 0.05) the performances of other methods, and the performances of FCM, ISODATA and Ward's are not significantly different. In the case of adulteration, though the average clustering accuracy based on spectral clustering is the highest, the performances of spectral clustering and FCM are not significantly different. In the case of pretreatments, the spectral clustering outperforms with statistical significance (alpha =0.05) the performances of other methods, and there is significant difference in the means of clustering accuracy based on different clustering methods. In general, spectral clustering presents the highest accuracy while SL presents the lowest accuracy for all cases. Meanwhile, FCM and ISODATA are generally better than the HCA methods and the *k*-means.

The success of spectral clustering is mainly based on the fact that it does not make any assumptions on the form of the clusters. Once the similarity graph is chosen, we just have to solve a linear problem, and there are no issues of getting stuck in local minima or restarting the algorithms for several times with different initializations. In addition, it is noticeable that cluster validation criteria in combination with majority voting in a way make clustering a semi-supervised technique. Using this approach it is possible to compare clustering methods with classification methods to find which method has better discrimination ability for a certain e-nose dataset. Actually, semi-supervised classification has emerged as an exciting new direction in the field of classification, and one of the semi-supervised approaches named Cluster-then-Label is exactly based on clustering and supervised learner such as the majority voting [50].

5. Conclusions

In the area of e-nose where clustering is applied, conventional clustering algorithms still play a dominant role and the search of the optimum clustering method for a certain dataset is often missing. An important reason explaining this is the absence of cluster validation criteria. This paper presents a state-of-the-art clustering method (spectral clustering) and three external cluster validation criteria (MI, precision and RI). Clustering of three e-nose datasets based on the spectral clustering, *K*-clustering (ISODATA, FCM and *k*-means) and HCA (SL,

Table 4
Evaluation of spectral clustering, *K*-clustering and hierarchical clustering for pretreatment dataset using three external validation criteria.

Clustering methods	MI	Precision	RI
Spectral clustering	0.4901	0.8933	0.9366
FCM	0.4699	0.86	0.9231
ISODATA	0.4482	0.7733	0.8984
k-means	0.2142	0.4933	0.5407
CL	0.2239	0.5067	0.5552
SL	0.0126	0.22	0.242
Ward's	0.4331	0.7933	0.8655

Table 5
Comparison of spectral clustering, *K*-clustering and hierarchical clustering for three e-nose datasets: material freshness dataset, adulteration dataset and pretreatment dataset.

Clustering methods	Material freshness	Adulteration	Pretreatments
	Accuracy[a]	Accuracy[a]	Accuracy[a]
Spectral clustering	98.12% ± 0.63% a	79.53% ± 3.56% a	89.45% ± 2.55% a
FCM	95.15% ± 2.89% b	77.97% ± 5.32% a	86.24% ± 3.70% b
ISODATA	94.55% ± 3.90% b	69.56% ± 4.21% b	82.30% ± 5.26% c
k-means	67.94% ± 9.10% d	50.02% ± 3.30% e	49.94% ± 3.30% f
CL	81.94% ± 8.10% c	52.88% ± 5.38% d	52.76% ± 5.41% e
SL	63.46% ± 6.56% e	23.07% ± 2.58% f	23.40% ± 2.14% g
Ward's	94.61% ± 1.84% b	59.79% ± 5.56% c	60.73% ± 7.74% d

[a] Accuracy is the bootstrapped mean of clustering accuracy, and means with the same letter are not significantly different at the 99.95% confidence level.

CL and Ward's) were compared. The results demonstrate that the spectral clustering outperforms with statistical significance (alpha = 0.05) the performances of other methods in all the three cases.

Though spectral clustering outperforms the conventional clustering methods in this paper and other works, it can be quite sensitive to changes in the similarity graph and to the choice of the parameters for the neighborhood graphs. In general, spectral clustering can be considered as a powerful tool which can produce extremely good results if applied with care. Meanwhile, it should also be noted that the cluster validation criteria in combination with majority voting in a way make clustering a semi-supervised classification technique.

Acknowledgments

The authors acknowledge the financial support of the National Key Technology R&D Program 2012BAD29B02-4 and the Chinese National Foundation of Nature and Science through project 31071548.

References

[1] M. Peris, L. Escuder-Gilabert, A 21st century technique for food control: electronic noses, Anal. Chim. Acta. 638 (2009) 1–15.
[2] M. Pardo, G. Sberveglieri, Random forests and nearest shrunken centroids for the classification of sensor array data, Sensors Actuators B Chem. 131 (2008) 93–99.
[3] M. Liu, M.J. Wang, J. Wang, D. Li, Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: application to the recognition of orange beverage and Chinese vinegar, Sensors Actuators B Chem. 177 (2012) 970–980.
[4] S.M. Scott, D. James, Z. Ali, Data analysis for electronic nose systems, Microchim. Acta 156 (2006) 183–207.
[5] L. Rokach, O. Maimon, Clustering Methods, Data Mining and Knowledge Discovery Handbook, Springer, 2005. 321–352.
[6] J. Almeida, L. Barbosa, A. Pais, S. Formosinho, Improving hierarchical cluster analysis: a new method with outlier detection and automatic clustering, Chemometr. Intell. Lab. 87 (2007) 208–217.
[7] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (1999) 264–323.
[8] T.N. Tran, R. Wehrens, L. Buydens, SpaRef: a clustering algorithm for multispectral images, Anal. Chim. Acta. 490 (2003) 303–312.
[9] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J.M. Pérez, I. Perona, An extensive comparative study of cluster validity indices, Pattern Recognit. 46 (2012) 243–256.
[10] M. Falasconi, M. Pardo, M. Vezzoli, G. Sberveglieri, Cluster validation for electronic nose data, Sensors Actuators B Chem. 125 (2007) 596–606.
[11] M. Penza, G. Cassano, F. Tortorella, G. Zaccaria, Classification of food, beverages and perfumes by WO_3 thin-film sensors array and pattern recognition techniques, Sensors Actuators B Chem. 73 (2001) 76–87.
[12] Q. Zhang, S. Zhang, C. Xie, D. Zeng, C. Fan, D. Li, Z. Bai, Characterization of Chinese vinegars by electronic nose, Sensors Actuators B Chem. 119 (2006) 538–546.
[13] Z. Yang, F. Dong, K. Shimizu, T. Kinoshita, M. Kanamori, A. Morita, N. Watanabe, Identification of coumarin-enriched Japanese green teas and their particular flavor using electronic nose, J. Food Eng. 92 (2009) 312–316.
[14] M. Liu, X. Han, K. Tu, L. Pan, J. Tu, L. Tang, P. Liu, G. Zhan, Q. Zhong, Z. Xiong, Application of electronic nose in Chinese spirits quality control and flavour assessment, Food Control 26 (2012) 564–570.
[15] B. Zhou, J. Wang, Discrimination of different types damage of rice plants by electronic nose, Biosyst. Eng. 109 (2011) 250–257.
[16] H. Yu, J. Wang, C. Yao, H. Zhang, Y. Yu, Quality grade identification of green tea using E-nose by CA and ANN, LWT—Food Sci. Technol. 41 (2008) 1268–1273.
[17] N. Lopez de Lerma, A. Bellincontro, F. Mencarelli, J. Moreno, R.A. Peinado, Use of electronic nose, validated by GC–MS, to establish the optimum off-vine dehydration time of wine grapes, Food Chem. 130 (2012) 447–452.
[18] R. Fend, C. Bessant, A.J. Williams, A.C. Woodman, Monitoring haemodialysis using electronic nose and chemometrics, Biosens. Bioelectron. 19 (2004) 1581–1590.
[19] N. Sahgal, N. Magan, Fungal volatile fingerprints: discrimination between dermatophyte species and strains by means of an electronic nose, Sensors Actuators B Chem. 131 (2008) 117–120.
[20] A.Z. Berna, J. Lammertyn, S. Buysens, C. Di Natale, B.M. Nicolaï, Mapping consumer liking of tomatoes with fast aroma profiling techniques, Postharvest Biol. Tecnol. 38 (2005) 115–127.
[21] K. Naraghi, N. Sahgal, B. Adriaans, H. Barr, N. Magan, Use of volatile fingerprints for rapid screening of antifungal agents for efficacy against dermatophyte *Trichophyton* species, Sensors Actuators B Chem. 146 (2010) 521–526.
[22] A. Guadarrama, M. Rodríguez-Méndez, J. De Saja, Conducting polymer-based array for the discrimination of odours from trim plastic materials used in automobiles, Anal. Chim. Acta. 455 (2002) 41–47.
[23] C. Li, G.W. Krewer, P. Ji, H. Scherm, S.J. Kays, Gas sensor array for blueberry fruit disease detection and classification, Postharvest Biol. Technol. 55 (2010) 144–149.
[24] M. Drake, P. Gerard, J. Kleinhenz, W. Harper, Application of an electronic nose to correlate with descriptive sensory analysis of aged cheddar cheese, LWT—Food Sci. Technol. 36 (2003) 13–20.
[25] T. Feng, H. Zhuang, R. Ye, Z. Jin, X. Xu, Z. Xie, Analysis of volatile compounds of Mesona Blumes gum/rice extrudates via GC–MS and electronic nose, Sensors Actuators B Chem. 160 (2011) 964–973.
[26] O. Canhoto, F. Pinzari, C. Fanelli, N. Magan, Application of electronic nose technology for the detection of fungal contamination in library paper, Int. Biodeterior. Biodegrad. 54 (2004) 303–309.
[27] T. Alizadeh, Chemiresistor sensors array optimization by using the method of coupled statistical techniques and its application as an electronic nose for some organic vapors recognition, Sensors Actuators B Chem. 143 (2010) 740–749.
[28] O. Canhoto, N. Magan, Electronic nose technology for the detection of microbial and chemical contamination of potable water, Sensors Actuators B Chem. 106 (2005) 3–6.
[29] J.P. Mensing, A. Wisitsoraat, A. Tuantranont, T. Kerdcharoen, Inkjet-printed sol–gel films containing metal phthalocyanines/porphyrins for opto-electronic nose applications, Sensors Actuators B Chem. 176 (2013) 428–436.
[30] R. Needham, J. Williams, N. Beales, P. Voysey, N. Magan, Early detection and differentiation of spoilage of bakery products, Sensors Actuators B Chem. 106 (2005) 20–23.
[31] S.D. Rodriguez, M.E. Monge, A.C. Olivieri, R.M. Negri, D.L. Bernik, Time dependence of the aroma pattern emitted by an encapsulated essence studied by means of electronic noses and chemometric analysis, Food Res. Int. 43 (2010) 797–804.
[32] B. Snopok, I. Kruglenko, Nonexponential relaxations in sensor arrays: forecasting strategy for electronic nose performance, Sensors Actuators B Chem. 106 (2005) 101–113.
[33] D.M. Wilson, K. Dunman, T. Roppel, R. Kalim, Rank extraction in tin-oxide sensor arrays, Sensors Actuators B Chem. 62 (2000) 199–210.
[34] C. Joseph, Q. Mu, B. Wei, T. Dacheng, 3D human posture segmentation by spectral clustering with surface normal constraint, Signal Process. 91 (2011) 2204–2212.
[35] U. Von Luxburg, A tutorial on spectral clustering, Stat. Comput. 17 (2007) 395–416.
[36] R. Nock, P. Vaillant, C. Henry, F. Nielsen, Soft memberships for spectral clustering, with application to permeable language distinction, Pattern Recognit. 42 (2009) 43–53.
[37] A. Ducournau, A. Bretto, S. Rital, B. Laget, A reductive approach to hypergraph clustering: an application to image segmentation, Pattern Recognit. 45 (2012) 2788–2803.
[38] P. Symeonidis, N. Iakovidou, N. Mantas, Y. Manolopoulos, From biological to social networks: link prediction based on multi-way spectral clustering, Data Knowl. Eng. 87 (2013) 226–242.
[39] K. Fujiwara, M. Kano, S. Hasebe, Correlation-based spectral clustering for flexible process monitoring, J. Process Control 21 (2011) 1438–1448.
[40] F. Yang, P.W. Grigsby, Delineation of FDG-PET tumors from heterogeneous background using spectral clustering, Eur. J. Radiol. 81 (2012) 3535–3541.
[41] USDA, United States Standards for Grades of Fresh Tomatoes, USDA, Agriculture Marketing Service, Washington, DC, USA, 1991.
[42] X.Z. Hong, J. Wang, Z. Hai, Discrimination and prediction of multiple beef freshness indexes based on electronic nose, Sensors Actuators B Chem. 161 (2012) 381–389.
[43] A. Smoliński, B. Walczak, J. Einax, Hierarchical clustering extended with visual complements of environmental data set, Chemometr. Intell. Lab. 64 (2002) 45–54.
[44] M. Steinbach, L. Ertöz, V. Kumar, The Challenges of Clustering High Dimensional Data, New Directions in Statistical Physics, Springer, 2004. 273–309.
[45] J.H. Ward Jr., Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. 58 (1963) 236–244.
[46] C. Hervada-Sala, E. Jarauta-Bragulat, A program to perform Ward's clustering method on several regionalized variables, Comput. Geosci. 30 (2004) 881–886.
[47] I.S. Dhillon, D.S. Modha, Concept decompositions for large sparse text data using clustering, Mach. Learn. 42 (2001) 143–175.
[48] T. Xia, J. Cao, Y.-d. Zhang, J.-t. Li, On defining affinity graph for spectral clustering through ranking on manifolds, Neurocomputing 72 (2009) 3203–3211.
[49] R. Wehrens, H. Putter, L. Buydens, The bootstrap: a tutorial, Chemometr. Intell. Lab. 54 (2000) 35–52.
[50] X. Zhu, A.B. Goldberg, Introduction to semi-supervised learning, Synth. Lect. Artif. Intell. Mach. Learn. 3 (2009) 1–130.

Xuezhen Hong received her Eng. Degree in Biosystems Engineering, in 2009, from Zhejiang University, China. She has been a PhD student of Zhejiang University since 2009. Her current interest includes data analysis based on electronic noses and tongues.

Jun Wang received his B. Eng. Degree and M. Eng. Degree in Agricultural Engineering in 1986 and 1988, respectively, and Dr. Eng. Degree in 1991 from Zhejiang Agricultural University. He has been a Professor at Zhejiang University since 1999. His current interests include electronic noses and measurement automation for evaluation of food quality.

Guande Qi received his BSc in Life Science from Zhejiang University, Hangzhou, China, in 2008. He has been a PhD student of the Department of Computer Science at Zhejiang University since 2008. His research interests include machine learning and data mining.