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.