Chemometrics and Intelligent Laboratory Systems 146 (2015) 457–463

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Comparison of semi-supervised and supervised approaches for classification of e-nose datasets: Case studies of tomato juices

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

[a] Department of Biosystems Engineering, Zhejiang University, 688 Yuhangtang Road, Hangzhou 310058, PR China
[b] College of Quality & Safety Engineering, China Jiliang University, Hangzhou 310018, PR China
[c] Department of Computer Science, Zhejiang University, Hangzhou 310027, PR China

ARTICLE INFO

Article history:
Received 8 May 2014
Received in revised form 23 June 2015
Accepted 2 July 2015
Available online 10 July 2015

Keywords:
Electronic nose
Semi-supervised classification
Supervised classification
Cluster-then-Label
Spectral clustering
Cherry tomato juice

ABSTRACT

Supervised classification, which is a fundamental classification approach for e-nose data, requires sufficient labeled data for training. However, sufficient labeled data requires extensive money, materials, energy and time. In this paper, a semi-supervised approach—Cluster-then-Label—that simultaneously uses labeled and unlabeled data to build a better classifier with fewer training data was introduced to deal with e-nose data for the first time. A novel clustering algorithm—spectral clustering—was also introduced to improve this semi-supervised approach. Three experiments—discriminating storage shelf life (SL), identifying pretreatments and authenticating juices, respectively—were conducted on cherry tomato juices using a PEN 2 e-nose, generating three datasets of different data structures. For each dataset, only 20% of data were selected for training. Classifications of the datasets by this semi-supervised approach and four supervised approaches (linear discriminant analysis (LDA), quadratic discriminant analysis, multi-class support vector machine and back propagation neural network) were compared. The results indicate that this spectral clustering based semi-supervised approach outperforms the supervised approaches in all cases. By using this semi-supervised approach, it is now possible to build reliable classifiers with only a few labeled data. It is also worth mentioning that this new approach takes no remarkable superiority over LDA. Thus, our next plan is to use more e-nose datasets for test.

1. Introduction

Research in the field of artificial olfaction (such as electronic noses, e-nose) has been focused on three main aspects: development of materials for sensors and sensor arrays, optimization and comparison of multiple data analysis methods, and application to various analytical tasks [1]. Successful applications of e-noses require not only sensors with excellent performances but also appropriate analytical methods.

Supervised classification is a fundamental classification approach for e-nose data [2]. In supervised classification, we are provided with a collection of labeled (pre-classified) data instance; the problem is to label a newly encountered, yet unlabeled, data instance. Typically, the given labeled data instances are used to learn the descriptions of classes which in turn are used to label a new data instance [3]. A lot of supervised classification methods, e.g., linear discriminant analysis (LDA) [4], quadratic discriminant analysis (QDA) [5], classification and regression trees (CART) [6], classification and influence matrix analysis (CAIMAN) [7], various neural networks (NNs) [8–13], support vector machine (SVM) [14] and random forest (RF) [15], have been successfully applied for e-nose data analysis. Generally, supervised classification requires sufficient labeled data to train a good classifier (sufficient usually means that the labeled data can roughly represent underlying structure of the entire data space) [16]. If the labeled data only represent part of the underlying data structure, or if the labeled data are mostly consisted of outliers, the classifier built would over fit the particular training data and thus lack generalization, i.e., it can't function well for the testing data. However, in many tasks, there is a paucity of labeled data since data labeling may require human annotators, special devices, or expensive and slow experiments. On the other hand, unlabeled data are often abundant and easy to obtain.

Semi-supervised classification, which uses unlabeled data together with labeled data to build a better classifier, has become a recent topic of interest especially in the area of computational statistics [17,18], image analysis [19–21], network traffic [22], document classification [23,24], biomedical informatics [25], etc. Some often-used approaches in the semi-supervised learning area include: self-training, co-training, transductive support vector machines, generative models, and graph-based methods [26]. Our research is inspired by Cluster-then-Label—a generative model that employs various clustering algorithms instead of probabilistic generative mixture models to identify mixing components from unlabeled data [27]. Cluster-then-Label makes use of both labeled and unlabeled data to reveal actual data space structure through clustering analysis. However, since this semi-supervised approach is

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

http://dx.doi.org/10.1016/j.chemolab.2015.07.001
0169-7439/

clustering based, it is very sensitive to its underlying assumptions, i.e., clusters coincide with decision boundaries. If the assumption is incorrect, the result can be poor. Therefore, it is very important to find a suitable clustering algorithm for a given dataset when employing this approach.

The clustering algorithms that are mostly applied in the area of e-nose [28–36] include between-groups linkage, within-groups linkage, single linkage clustering, centroid clustering, complete linkage clustering (CL), Ward's clustering, etc. However, these traditional clustering approaches have their own limitations and scopes of application. For example, between-groups linkage, within-groups linkage and centroid clustering 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; single linkage clustering maintains good performance on datasets containing non-isotropic clusters but has a drawback known as the "chaining effect" [37]; CL is not strongly affected by outliers, but it can break large clusters and has trouble with convex shapes [38]; And ward's clustering may cause elongated clusters to split and portions of neighboring elongated clusters to merge [39,40]. Recently, a state-of-the-art clustering method—spectral clustering—has become a topic of interest. By constructing an undirected weighted similarity graph, spectral clustering utilizes 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 [41]. In our previous research [42], spectral clustering was found better than six conventional clustering methods (ISODATA, FCM, *k*-means, single linkage clustering, CL and Ward's).

In this paper, Cluster-then-Label based on spectral clustering and majority voting was applied to deal with e-nose data for the first time. Three experiments—discriminating storage shelf life (SL), identifying pretreatments and authenticating juices, respectively—were conducted on cherry tomato juices using an e-nose, generating three datasets of different data structures. Classification performances based on this semi-supervised approach and various supervised approaches were compared. The main objective of this research is to explore if the proposed semi-supervised approach would outperform the supervised approaches in the case of classification with only a few labeled e-nose data.

2. Experimental

2.1. Preparation of tomato juice samples

Chinese variety, *youbei* cherry tomatoes were picked for home-made juices at the experimental orchard of Zhejiang University, Hangzhou, China. Upon arrival at the laboratory, the picked samples were rinsed with clear water and wiped dry with clean cloth prior to any experiments.

Three experiments were conducted, and detailed experimental information is given in Table 1.

The first experiment was to discriminate cherry tomato juices squeezed from tomatoes of different freshness (storage SL). For this experiment, light-red (approximately 70% of the surface, in the aggregate, shows pinkish-red or red) [43] cherry tomatoes were selected and stored in a refrigerator at 4 °C for 16 days. Every three days (i.e., on day 1, 4, 7, 10, 13 and 16), appropriate amount of cherry tomatoes were taken out and juiced by a fruit squeezer for 30 s to obtain 100% fresh cherry tomato juices. The squeezed juices were then used for e-nose measurement. 25 juice samples were prepared on each measuring day, thus, there were in total 25 samples × 6 groups (measuring day) = 150 samples.

The second experiment was to identify cherry tomato juices processed by different pretreatments. For this experiment, appropriate amount of light-red cherry tomatoes were pretreated by six different processes prior to being squeezed for 100% fresh cherry tomato juices. The six pretreatments are 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 juice samples were prepared for each pretreatment group, thus, there were in total 25 samples × 6 groups (pretreatments) = 150 samples.

The third experiment was to authenticate cherry tomato juices. For this experiment, juices squeezed form fresh light-red cherry tomatoes were blended with juices squeezed form overripe or decaying cherry tomatoes at seven levels of adulteration (from 0 to 30% (*w*/*w*) in steps of 5%). The seven adulteration levels are: 0%, 5%, 10%, 15%, 20%, 25% and 30%. 25 juice samples were prepared for each adulteration group, thus, there were in total 25 samples × 7 groups (adulteration levels) = 175 samples.

2.2. E-nose sampling procedure and data acquirement

A PEN 2 e-nose (Airsense Analytics, GmBH, Schwerin, Germany) consisting of ten metal oxide semiconductor (MOS) sensors was employed to examine the aforementioned juice samples. Description of the sensor array is given in Table 2, observed from which, the sensors are non-specific. For example, except for methane, sensor S5 is also sensitive to propane and aliphatic non-polar molecules. Meanwhile, both sensors S1 and S3 are sensitive to aromatics, and both sensors S5 and S6 are sensitive to methane. However, each sensor has different sensitivity towards the same compound.

Before e-nose detection, each juice 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) so that its headspace could collect volatiles from the sample. During the measurement process, the headspace gaseous compounds were pumped into the sensor array (400 mL/min) through a Teflon tubing connected to a needle in the plastic wrap, causing changes in conductance ratio G/G_0 (G and G_0 are conductance of the sensors exposed to sample gas and zero gas, respectively) of each sensor. The measurement phase lasted for 70 s, which was long enough for the sensors to reach stable signal values. Signal data from the sensors were collected by a computer once per second. When the measurement process was complete, the acquired data was stored for later use, and zero gas (air filtered by active carbon) was pumped into the sample gas path from the other port of the instrument for 50 s. All the experiments and measurements were carried out at a temperature of 20 °C ± 1 °C.

Table 1
Experimental design and sampling protocol employed for cherry tomato juices.

No.	Experimental content	Group information	Number of samples per group	Total samples
1	Discriminating storage shelf life (SL)	6 groups of SL: day 1, 4, 7, 10, 13 and 16	25	150
2	Identifying pretreatments	6 groups of pretreatments: control, freezing, low temperature blanching, high temperature blanching, microwave blanching and steam blanching	25	150
3	Authenticating juices	7 groups of adulterated juices: 0%, 5%, 10%, 15%, 20%, 25% and 30%)	25	175