Computers and Electronics in Agriculture 147 (2018) 64–69

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Discrimination of tea varieties using FTIR spectroscopy and allied Gustafson-Kessel clustering

Xiaohong Wu[a,b,*], Jin Zhu[a], Bin Wu[c], Jun Sun[a], Chunxia Dai[a,d]

[a] School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China
[b] Key Laboratory of Facility Agriculture Measurement and Control Technology and Equipment of Machinery Industry, Jiangsu University, Zhenjiang 212013, China
[c] Department of Information Engineering, ChuZhou Vocational Technology College, Chuzhou 239000, China
[d] School of Food and Biological Engineering, Jiangsu University, Zhenjiang 212013, China

ARTICLE INFO

Keywords:
Tea
Fourier transform infrared reflectance
Fuzzy clustering
Classification methods

ABSTRACT

For the purpose of classifying tea varieties, allied Gustafson-Kessel (AGK) clustering was proposed to cluster the Fourier transform infrared reflectance (FTIR) spectra of tea samples. As a fuzzy clustering algorithm, AGK can not only produce fuzzy membership and typicality values but also cluster various shapes of data with the help of Gustafson-Kessel (GK) clustering. After FTIR spectra were collected by FTIR-7600 infrared spectrometer, they were preprocessed with multiple scatter correction (MSC). To reduce the dimensionality of FTIR spectra and make the classification of data easily, principal component analysis (PCA) and linear discriminant analysis (LDA) were used to process the FTIR spectra. After that, fuzzy c-means (FCM) clustering, possibilistic c-means (PCM) clustering, AGK clustering and allied fuzzy c-means (AFCM) clustering were performed to cluster data, respectively. The clustering accuracy of AGK achieved 93.9% which was the highest one than other fuzzy clustering algorithms. The results obtained in experiments showed that AGK coupled with FTIR spectroscopy could provide an effective discrimination model for classification of tea varieties successfully.

1. Introduction

As a healthy beverage (Sun et al., 2017), tea is being greeted with increasing approval in the world. Tea originated in China and the custom of drinking tea is from Sichuan, China. Drinking tea often keeps you in the pink because tea contains some wholesome compounds such as catechins, cholestenone, caffeine (Li et al., 2015; Sinija and Mishra, 2009), inositol, pantothenic acid, amino acid (Li et al., 2013) and folic acid. Besides a beverage, tea can be served as the good medicine of the bowel disease, coronary heart disease, hypertension, etc. For different varieties of tea, the effectiveness is often distinct. For example, pu'er tea can lower blood fat and prevent from atherosclerosis and coronary heart disease. Tie Guanyin, a variety of oolong tea, has anti-aging and anti-cancer effect. Zhu Yeqing can relieve heat and quell thirst with the effect of detoxication and diuresis. Moreover, the factors affecting the quality of the tea, such as color and luster, sweet smell and taste, are closely related to variety, cultivation conditions and storage conditions. The consumers want to buy real tea, not the fake in the tea market. Therefore, it becomes an urgent task to set up a rapid and effective classification of tea varieties for consumers, researchers and farmers.

Human sensory evaluation is a common method for evaluating the quality of tea (Yaroshenko et al., 2014; Zhi et al., 2013). Trained sensory panels always grade tea samples according to their appearance, aroma, soup color, taste and Securinega through their vision, taste, smell and touch. However, training skilled tea personnel spends a lot of time and money. Furthermore, for the same tea sample, sensory evaluation results are not the same to different professionals because the results are influenced by experience, gender, mental state, physical condition and other factors.

Recently, some studies have been described to classify tea varieties using chemical analysis methods, such as high performance liquid chromatography (HPLC) (Wang et al., 2014a,b), gas chromatography-mass spectrometry (GC–MS) (Ding et al., 2015; Lv et al., 2015), liquid chromatography coupled with tandem mass spectrometry (LC-MS) (Zhang et al., 2014), and inductively coupled plasma optical emission spectrometry (ICP-OES) (Szymczycha-Madeja et al., 2015). Nevertheless, chemical analysis methods mentioned above are complex and time-consuming to be used in discriminating tea varieties.

In order to meet the rapid development of tea market and people's demand for higher quality of tea, it is urgent to develop a fast, safe and green detection technology for identification of tea varieties. Fourier transform infrared reflectance (FTIR) spectroscopy or near-infrared

[*] Corresponding author at: School of Electrical and Information Engineering, Jiangsu University, Xuefu Road No. 301, Zhenjiang 212013, China.
E-mail address: wxh_www@163.com (X. Wu).

https://doi.org/10.1016/j.compag.2018.02.014
Received 4 August 2017; Received in revised form 26 January 2018; Accepted 10 February 2018

spectroscopy (Sinija and Mishra, 2008), as a fast and nondestructive technology, can be served as a powerful analytical tool and it has been widely used to discriminate tea varieties (Cai et al., 2015; Wu et al., 2016), apple (Wu et al., 2016a; Wu et al., 2016b), meat (Kodogiannis et al., 2014), jujube (Yang et al., 2015), olive oil (Jiménez-Carvelo et al., 2017), fish species (Alamprese and Casiraghi, 2015), etc. Some researchers studied the feasibility of rapidly discriminating tea varieties using FTIR spectroscopy or near-infrared spectroscopy coupled with supervised pattern classification methods (Diniz et al., 2014; He et al., 2012, 2007), such as back propagation artificial neural network (BP-ANN) (Chen et al., 2007, 2009) and support vector machine (SVM) (Chen et al., 2007, 2009). BP-ANN, a kind of artificial neural network (ANN), is a nonlinear learning method with a few layers networks, and it can solve nonlinear classification problem. But BP-ANN easily gets trapped at local minima and does not converge to the global minimum point (Chandwani et al., 2015). Under the principle of structural risk minimization, SVM can solve local minima problem and carry out nonlinear classification with kernel trick. However, it is difficult to adjust the parameters of SVM such as the kernel parameter for the radial basis function (RBF) kernel, the most frequently used kernel function, and the soft-margin parameter C (Chang and Chou, 2015). Moreover, to search the proper parameters for optimizing SVM may require a large number of calculations (Liu et al., 2011). Therefore, it makes sense that simple and effective classification methods should be researched for identification of tea varieties.

As an unsupervised classification method, fuzzy clustering always shows better performance than traditional one. Fuzzy clustering has been widely used in digital image processing, computer vision and pattern recognition (Bezdek et al., 1999). Fuzzy c-means (FCM) clustering, a well-known fuzzy clustering, derives its origin from hard c-means (HCM) clustering algorithm (Bezdek, 1981). FCM has the probabilistic constraint interpreting memberships as degrees of sharing (Bezdek, 1981). However, because of the probabilistic constraint, FCM is sensitive to noise (Barni et al., 1996). How to deal with the noisy data is an important issue in designing fuzzy clustering models. To overcome the noise sensitivity drawback of FCM, Krishnapuram and Keller have presented the possibilistic c-means (PCM) clustering algorithm by abandoning the constraint of FCM and constructing the novel objective functions (Krishnapuram and Keller, 1993). PCM can deal with noisy data better than FCM, but it is very sensitive to initializations and sometimes generates coincident clusters. PCM attaches importance to the possibility (typicality) but neglects the important membership. To combine the benefits of FCM and PCM, an allied fuzzy c-means (AFCM) clustering has been proposed to produce memberships and possibilities simultaneously (Wu and Zhou, 2006). On the other hand, Gustafson and Kessel proposed Gustafson-Kessel (GK) clustering to cluster data containing different geometric shapes (Costel et al., 2007). However, there are few reports on applying fuzzy clustering algorithms for classification of tea varieties together with FTIR spectroscopy.

In this work, FTIR spectroscopy was used as a nondestructive technology in detecting tea samples with different varieties. A novel fuzzy clustering called allied Gustafson-Kessel (AGK) clustering was proposed by combination of AFCM and GK for classification of tea varieties. After FTIR spectra of tea samples were processed by multiple scatter correction (MSC), principal component analysis (PCA) and linear discriminant analysis (LDA), AGK was performed for classifying the tea varieties and its clustering accuracy and the computing time were compared with those of FCM, PCM and AFCM. Furthermore, we pointed out the noise sensitivity problem of GK and our proposed AGK can solve this problem. The main objectives of this research are: (1) to investigate the potential of FTIR for classification of tea varieties; (2) to propose allied Gustafson-Kessel (AGK) clustering based on AFCM and GK; (3) to compare the clustering accuracy and the computing time of FCM, PCM, AFCM and AGK in classifying FTIR spectra of tea samples.

2. Materials and methods

2.1. Sample preparation

Three varieties of tea samples: Emeishan Maofeng, high quality Leshan trimeresurus and low quality Leshan trimeresurus were prepared in this study. They came from two regions: Emeishan (Emeishan Maofeng) and Leshan (Leshan trimeresurus). The number in each variety of them is 32 and the total number is 96. After all samples were ground with a small mill, 0.5 g tea powder from each sample was evenly mixed with KBr according to ratio 1:100, and 1 g mixture was chosen to be pressed with film as one sample for FTIR experiments.

2.2. Spectral acquisition

The FTIR spectra of tea samples were collected using a FTIR-7600 infrared spectrometer (Lambda Scientific Pty Ltd, Edwardstown, Australia) with the high-sensitivity Deuterated Triglycine Sulphate (DTGS) detector. After the spectrometer was turned on and warmed up for one hour, each spectrum was acquired as the average of 32 scans over the range of 4001.569–401.1211 cm^{-1} with a sampling interval of 1.9285 cm^{-1}. The dimensionality of the spectrum is 1868. Each tea sample was made three separate spectral measurements and the average of the three spectra was prepared as the final datum for further data analysis. Because infrared spectrophotometer is sensitive to the change of the temperature and the humidity in laboratory, the spectral collection was operated at temperatures of around 25 °C, and on the relative humidity of 50–60%.

2.3. Spectra preprocessing method

The FTIR spectra of tea samples with the wave numbers from 4001.569 cm^{-1} to 401.1211 cm^{-1} were shown in Fig. 1. The spectra contain not just spectral absorption information related to chemical content of tea samples but light scatter information. Light scattering were influenced by physical factor such as particle size, shape and distribution, and there are possible differences in the light scatter information of different samples (Wang et al., 2014a,b). Under the influence of light scatter information, classification results are not satisfactory, and multiple scattering correction (MSC) is the commonly used method of resolving it effectively. Standard normal variate (SNV) is also the widely used pre-processing method in FTIR spectroscopy and it is similar to MSC in reducing the scattering effects of the spectrum (Zhao et al., 2016). In this study, MSC was utilized to preprocess the FTIR spectra and Fig. 2 illustrated the results.

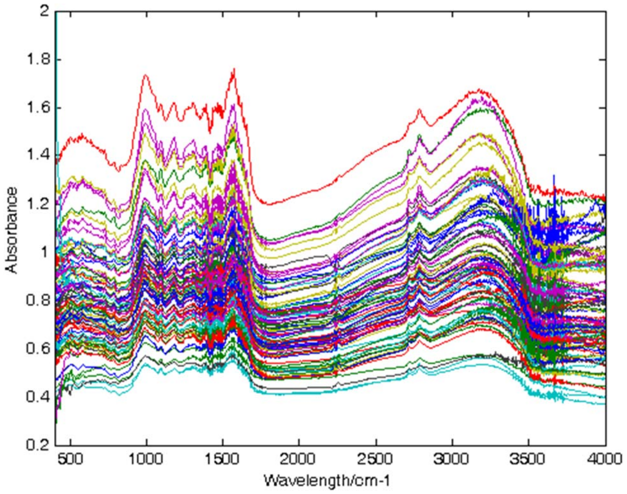

Fig. 1. The FTIR spectra of tea samples.

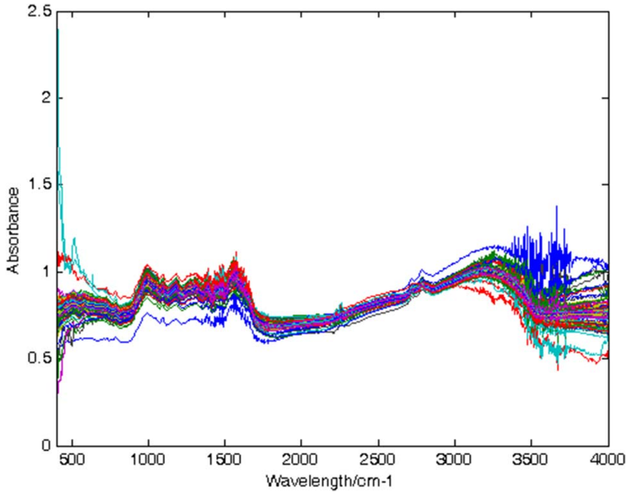

Fig. 2. The FTIR spectra preprocessed by MSC.

2.4. Clustering methods

2.4.1. Allied fuzzy c-means clustering

Given an unlabeled data set $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\} \subset \Re^d$, AFCM clustering tends to find the partition of $\boldsymbol{X}$ into $1 < c < n$ fuzzy subsets by minimizing the following objective function

$$J_{AFCM}(U,T,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik}^m + t_{ik})D_{ik}^2 + \sum_{i=1}^{c}\eta_i\sum_{k=1}^{n}(t_{ik}\log t_{ik} - t_{ik}) \tag{1}$$

Here, x_k is the sample datum; c is the number of clusters, $1 < c < n$; n is the number of data points; u_{ik} is the fuzzy membership value of x_k in class i, $U = [u_{ik}]_{c\times n}$; t_{ik} is the possibilistic (typicality) value x_k in class i, $T = [t_{ik}]_{c\times n}$; weighting exponent $m \in (1,\infty)$; $D_{ik} = \|x_k - v_i\|$; v_i is the cluster center of the i-th class, $V = [v_1,v_2,\ldots,v_c]$; The parameter η_i can be calculated as follows

$$\eta_i = \frac{\sum_{k=1}^{n} u_{ik,FCM}^m D_{ik}^2}{\sum_{k=1}^{n} u_{ik,FCM}^m} \tag{2}$$

Here, $u_{ik,FCM}$ is the terminal fuzzy membership value from FCM.

Subject to the constraints: $\sum_{i=1}^{c} u_{ik} = 1$,$\forall\ k$, and $0 \leqslant u_{ik},t_{ik} \leqslant 1$. If $D_{ik} = \|x_k - v_i\| > 0$ for all i and $k,m > 1$, and $\boldsymbol{X}$ contains at least c distinct data points, $\min_{(U,T,V)} J_{AFCM}(U,T,V)$ is optimized and the following equations are obtained

$$u_{ik} = \left[\sum_{j=1}^{c}\left(\frac{D_{ik}}{D_{jk}}\right)^{\frac{2}{m-1}}\right]^{-1} ,\forall\ i,k \tag{3}$$

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik}^m + t_{ik})x_k}{\sum_{k=1}^{n}(u_{ik}^m + t_{ik})} ,\forall\ i \tag{4}$$

$$t_{ik} = \exp\left(-\frac{D_{ik}^2}{\eta_i}\right) ,\forall\ i,k \tag{5}$$

Then the iteration algorithm of the AFCM clustering is described as follows

```
Initialization
    (1) Run FCM until termination to obtain the final cluster centers $\boldsymbol{V}$
        as the initial cluster centers $\boldsymbol{V}^{(0)}$ of AFCM, and use Eq. (2) to
        calculate $\eta_i$;
    (2) Fix $c$, $\varepsilon$ and $m$. $1 < c < n$, $m > 1$;
    (3) Set iteration counter $r = 1$ and maximum iteration $r_{\max}$;
Repeat
    Step 1 Update fuzzy membership $U^{(r)}$ by Eq. (3);
    Step 2 Update typicality $T^{(r)}$ by Eq. (4);
    Step 3 Update $V^{(r)}$ by Eq. (5);
    Step 4 Increase $r$;
Until ($\|V^{(r)} - V^{(r-1)}\| < \varepsilon$) or ($r > r_{\max}$)
```

2.4.2. Allied Gustafson-Kessel clustering

Given an unlabeled data set $\boldsymbol{X} = \{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\} \subset \Re^d$, The objective function of AGK is defined as

$$J_{AGK}(U,T,V) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik}^m + t_{ik})D_{ikA_i}^2 + \sum_{i=1}^{c}|S_{fi}|^{\frac{1}{d}}\sum_{k=1}^{n}(t_{ik}\log t_{ik} - t_{ik}) \tag{6}$$

Here, D_{ikA_i} is the distance norm between x_k and v_i.

$$D_{ikA_i}^2 = (x_k - v_i)^{\mathrm{T}} A_i (x_k - v_i), \quad A_i = |S_{fi}|^{\frac{1}{d}}(S_{fi}^{-1}), \quad S_{fi} = \frac{\sum_{k=1}^{n} u_{ik}^m (x_k - v_i)(x_k - v_i)^{\mathrm{T}}}{\sum_{k=1}^{n} u_{ik}^m} \tag{7}$$

A_i is the norm-inducing matrix in the ith cluster center; S_{fi} is the fuzzy scatter matrix in the ith cluster center.

Subject to the constraints the same as AFCM, optimization of $\min_{(U,T,V)} J_{AGK}(U,T,V)$ is calculated, so we have

$$u_{ik} = \left[\sum_{j=1}^{c}\left(\frac{D_{ikA_i}}{D_{jkA_j}}\right)^{\frac{2}{m-1}}\right]^{-1} ,\forall\ i,k \tag{8}$$

$$t_{ik} = \exp\left(-\frac{D_{ikA_i}^2}{|S_{fi}|^{\frac{1}{d}}}\right) ,\forall\ i,k \tag{9}$$

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik}^m + t_{ik})x_k}{\sum_{k=1}^{n}(u_{ik}^m + t_{ik})} ,\forall\ i \tag{10}$$

Then the iteration algorithm of the AGK clustering is described as follows

```
Initialization
    (1) Run FCM until termination to obtain the final cluster centers $V$
        as the initial cluster centers $V^{(0)}$ of AGK, and use Eq. (3) to
        calculate fuzzy membership as the initial membership $U^{(0)}$ of
        AGK;
    (2) Fix $c$, $\varepsilon$ and $m$. $1 < c < n$, $m > 1$;
    (3) Set iteration counter $r = 1$ and maximum iteration $r_{\max}$;
Repeat
    Step 1 Calculate $S_{fi}$, $A_i$ and $D_{ikA_i}$ by Eq. (7);
    Step 2 Update fuzzy membership $U^{(r)}$ by Eq. (8);
    Step 3 Update typicality $T^{(r)}$ by Eq. (9);
    Step 4 Update $V^{(r)}$ by Eq. (10);
    Step 5 Increase $r$;
Until ($\|V^{(r)} - V^{(r-1)}\| < \varepsilon$) or ($r > r_{\max}$)
```

The clustering results of AGK are sensitive to the initial membership $U^{(0)}$ the same as GK clustering, so we use the terminal cluster centers and fuzzy membership of FCM as the initial $V^{(0)}$ and $U^{(0)}$ of AGK. On the other hand, GK clustering is sensitive to noisy data contained in some data sets because its memberships of a data point across classes sum to one the same as the probabilistic constraint of FCM. The membership of GK clustering is the same as AGK illustrated in the Eq. (8), and therefore

$$\sum_{i=1}^{c} u_{ik} = \sum_{i=1}^{c}\left(\left[\sum_{j=1}^{c}\left(\frac{D_{ikA_i}}{D_{jkA_j}}\right)^{\frac{2}{m-1}}\right]^{-1}\right) = 1 \tag{11}$$

Eq. (11) is identical with the probabilistic constraint of FCM. Since FCM is sensitive to noises Krishnapuram and Keller, 1993, GK clustering is also sensitive to noises. In order to solve the noise sensitivity problem of FCM, AFCM clustering integrates FCM with PCM. Inspired by this, we integrate GK clustering with PCM to present AGK. AGK solves the noise sensitivity problem of GK clustering, and furthermore AGK can provide fuzzy memberships and possibilities simultaneously.

3. Results and discussion

3.1. Spectra analysis

As shown in Fig. 1, the FTIR spectra in the range of 4001.569 cm^{-1} to 401.1211 cm^{-1} contain molecular bonds (C—H, O—H, C—O, etc.) information. To different tea variety, there remain some differences in chemical components, content and proportion. This makes it possible to use FTIR spectroscopy for classification of tea varieties. The absorption peak at 987.38 cm^{-1} stood for C—H outer bending vibration in the olefinic double bonds RCH = CH2; the absorption peak at 1575.6 cm^{-1} and 1604.5 cm^{-1} indicated C═C stretch vibration in the aromatic ring, and the band between 3500 cm^{-1} and 3100 cm^{-1} gave evidence for N—H vibration in the amides. The FTIR spectra had a very close distribution in the infrared spectral region, and most of them overlapped heavily. Therefore, classification of tea varieties will be confronted with difficulty in dealing with the overlapped spectra. Then, it is meaningful to build the effective classification mode to discriminate the overlapped spectra.

3.2. Dimension reduction using PCA

The FTIR spectra are the high-dimensional data and they contain a large amount of redundant information, so it is very difficult in classifying the original spectra. PCA is a good choice to reduce the dimensionality of the FTIR spectra and eliminate the redundant information. As shown in Fig. 3, the first two principal components (PC1 and PC2), which spanned the two-dimensional feature space where the FTIR spectra were expressed, could explain 81.48% of the total variance. Tea samples were labeled according to their varieties, i.e. Emeishan Maofeng, high quality Leshan trimeresurus and low quality Leshan trimeresurus. One variety of tea samples got mixed up with two other varieties of tea samples. Therefore, the FTIR spectra are still difficult to be classified. Because the first 14 principal components accounted for 99.51% of the total variance and resulted in the highest clustering accuracy by cross-validation, PCA reduced the dimensionality of FTIR spectra from 1868 to 14 in this study.

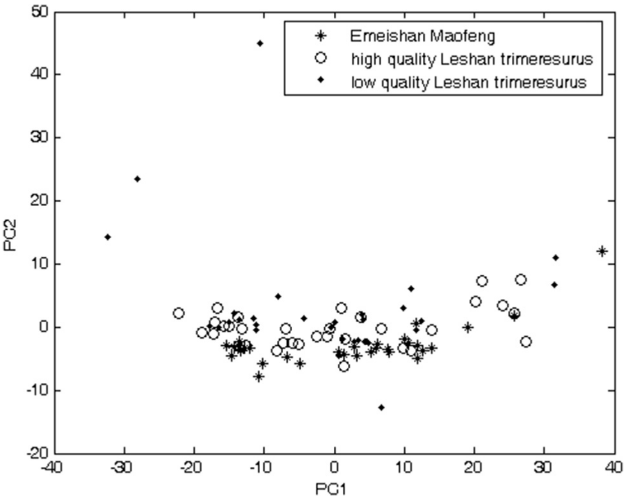

Fig. 3. PCA scores plot of vectors with PC1 and PC2.

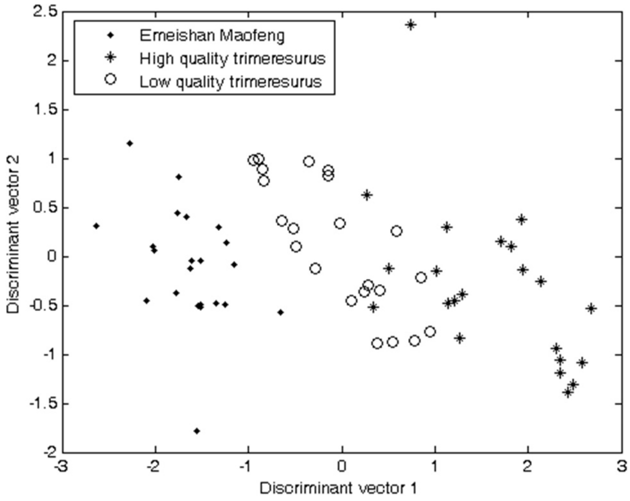

Fig. 4. LDA scores plot of vectors with DV1 and DV2.

3.3. Discriminant information extraction using LDA

As a discriminant information extraction method, linear discriminant analysis (LDA) is always used in achieving feature vectors from data. The data cannot be the high-dimensional FTIR spectra because this will lead to small sample size problem when LDA deals with data whose dimensionality outnumbers the number of samples. To solve this problem, PCA + LDA is a good strategy; that is to say LDA can extract feature vectors from 14 dimensional data from PCA. The total 96 tea samples were divided into the training set and the test set. The training set contained 30 samples (10 samples in every variety of teas), and was used to compute the feature vectors by LDA. The remaining 66 samples acted as the test set and they were projected on the feature vectors. For *c*-class classification problem, the optimal number of feature vectors is *c*-1. This study discusses three varieties of teas, and there are two feature vectors, i.e., the discriminant vector 1 (DV1) and the discriminant vector 2 (DV2). The 14 dimensional test data were projected on DV1 and DV2 and transformed into two dimensional data. Fig. 4 showed the scores plot of DV1 and DV2 on test data. As shown in Fig. 4, the data of Emeishan Maofeng were well separated from the data of high quality Leshan trimeresurus and low quality Leshan trimeresurus, and there are a few overlapping data points between high quality Leshan trimeresurus and low quality Leshan trimeresurus. Comparing with the separability of the data in Fig. 3, Fig. 4 was the better.

3.4. Classification results of FCM

After FTIR spectra of tea were compressed by PCA and discriminant information was extracted by LDA, FCM, as a classifier, was used to cluster the test data for identification of tea varieties. FCM aims to find the partition of dataset into *c* fuzzy subsets based a least-squared errors criterion. Before running FCM, we set the algorithmic parameters: threshold $\varepsilon = 0.00001$, maximum number of iterations $r_{\max} = 100$, class number $c = 3$, the number of data points $n = 66$ (the test set), weighting exponents: $m = 2$, and the initial cluster centers came from the first three data:

$$V^{(0)} = \begin{bmatrix} v_1^{(0)} \\ v_2^{(0)} \\ v_3^{(0)} \end{bmatrix} = \begin{bmatrix} 0.0814 & 0.0628 \\ 0.1209 & 0.0263 \\ 0.1278 & -0.0040 \end{bmatrix} \tag{12}$$

Through 46 iterations, the terminal fuzzy membership values of FCM were achieved and shown in Fig. 5. The terminal cluster centers were:

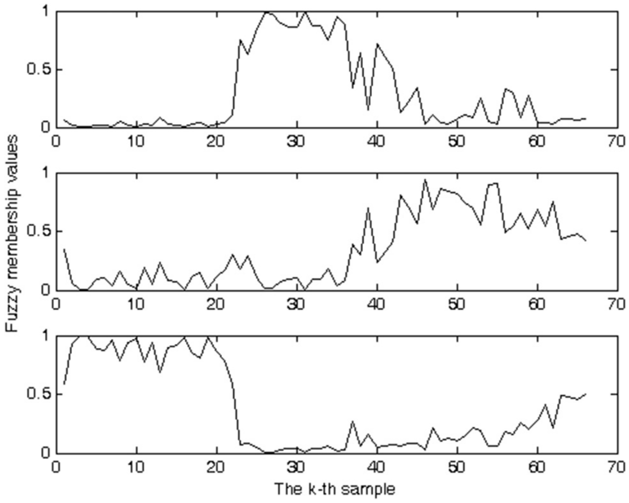

Fig. 5. The terminal fuzzy membership values of FCM.

$$V^{(46)}=\begin{bmatrix} v_1^{(46)}\\ v_2^{(46)}\\ v_3^{(46)}\end{bmatrix}=\begin{bmatrix} -0.15803 & 0.040329\\ -0.001958 & 0.0049377\\ 0.11944 & -0.0055577\end{bmatrix} \tag{13}$$

The method of discrimination of tea varieties by using FCM includes two steps. One is to determine which variety the terminal cluster centers belong to and the other is to confirm which variety the test data belong to. The average values of training samples were: Emeishan Maofeng $\overline{x}_{train1}=[0.13090\ \ 0.055517]$, high quality Leshan trimeresurus $\overline{x}_{train2}=[-0.16344\ \ 0.02097]$ and low quality Leshan trimeresurus $\overline{x}_{train3}=[0.029847\ \ -0.12712]$. For the terminal cluster center $v_1^{(46)}$, it belonged to the variety of tea whose average value of training samples was closest to $v_1^{(46)}$ in Euclidean distance. $\|v_1^{(46)}-\overline{x}_{train1}\|=0.2893$, $\|v_1^{(46)}-\overline{x}_{train2}\|=0.0201$ and $\|v_1^{(46)}-\overline{x}_{train1}\|=0.2517$. Therefore, $v_1^{(46)}$ belonged to high quality Leshan trimeresurus. With the same method, we could confirm that $v_2^{(46)}$ was affiliated with low quality Leshan trimeresurus and $v_3^{(46)}$ was attached to Emeishan Maofeng. For the kth test sample x_k, if the fuzzy membership value u_{ik} is bigger than 0.5, x_k belongs to v_i. For example, the terminal fuzzy membership values of the test sample x_1 were: $u_{11}=0.0623$, $u_{21}=0.3497$ and $u_{31}=0.5880$, and $u_{31}>0.5$, so x_1 belonged to $v_3^{(46)}$. Because $v_3^{(46)}$ was attached to Emeishan Maofeng, x_1 also belonged to Emeishan Maofeng. As shown in Table 1, the clustering accuracy of FCM was 89.4%.

3.5. Classification results of PCM

After FTIR spectra of tea were processed by PCA and LDA, they were ready for clustering with FCM and PCM. Because PCM is sensitive to initializations, FCM was performed to termination firstly and the terminal cluster centers of FCM served as the initial cluster centers of PCM. After PCM was operated and came to termination, the terminal possibilistic membership values of PCM were utilized to cluster test data and its clustering accuracy was just 56%.

3.6. Classification results of AFCM

FCM can provide fuzzy membership values and PCM can produce possibilistic membership values. Different from FCM and PCM, AFCM can produce both fuzzy membership values and possibilistic membership values simultaneously. The initial cluster centers of AFCM came from the terminal cluster centers of FCM as Eq. (13). The values of parameters, such as ε, $r_{\max}$ and c, were the same as those of FCM in Section 3.4. After 40 iterations, the terminal cluster centers of AFCM were:

$$V^{(40)}=\begin{bmatrix} v_1^{(40)}\\ v_2^{(40)}\\ v_3^{(40)}\end{bmatrix}=\begin{bmatrix} -0.14411 & 0.035243\\ 0.031651 & -0.012067\\ 0.11725 & 0.0035684\end{bmatrix} \tag{14}$$

With the same method described in Section 3.4, we could decide the variety that the terminal cluster centers of AFCM belonged to. As a result, $v_1^{(40)}$, $v_2^{(40)}$ and $v_3^{(40)}$ belonged to high quality Leshan trimeresurus, low quality Leshan trimeresurus and Emeishan Maofeng, respectively. Similarly, the terminal fuzzy membership values could be used to confirm the variety of the kth sample x_k. On the other hand, the terminal possibilistic membership values could also be applied to differentiate the variety of the kth sample x_k. For the kth test sample x_k, there are three terminal possibilistic membership values, and if the possibilistic membership value t_{ik} is the biggest one, x_k belongs to v_i. For example, the terminal possibilistic membership values of x_7 were: $t_{17}=0.01$, $t_{27}=0.10$ and $t_{37}=0.89$, so x_7 belonged to $v_3^{(40)}$; that is to say, x_7 belonged to Emeishan Maofeng. The clustering accuracies of AFCM were: U/T = 89.4%/90.9%.

Table 1
Clustering accuracy of FCM, PCM, AFCM and AGK.

FCM(U)	PCM(T)	AFCM(U/T)	AGK(U/T)
89.4%	56%	89.4%/90.9%	93.9%/93.9%

3.7. Classification results of AGK

Like AFCM, AGK clustering also can offer both fuzzy membership values and possibilistic membership values simultaneously. However, the distance measure in AFCM is the Euclidean distance while that in AGK is the weighted inner product induced distance using a fuzzy covariance matrix. The computational protocols of AGK were set: termination criterion $\varepsilon=0.00001$, maximum number of iterations $r_{\max}=100$, class number $c=3$, the number of data points $n=60$ (the test set), weighting exponents: $m=2$, the dimensionality of data $d=2$ and the initial cluster centers were the terminal cluster centers of FCM as Eq. (13). Then AGK clustered the test data which were the same as FCM and AFCM dealt with. The terminal fuzzy membership values of AGK were illustrated in Fig. 6. The clustering accuracies of AGK were: U/T = 93.9%/93.9%. Therefore, the clustering accuracies of AGK were higher than those of AFCM.

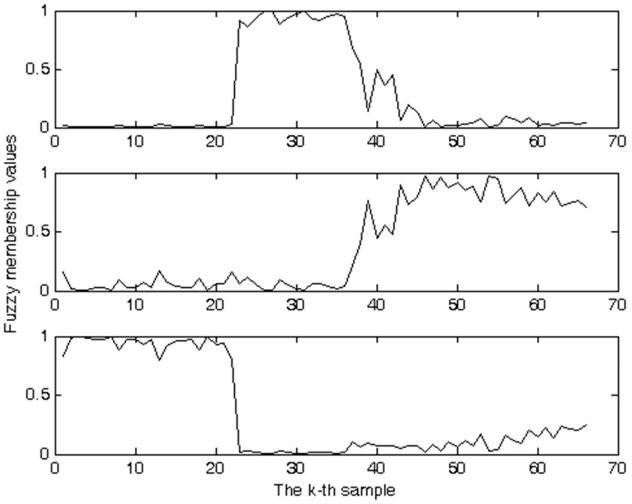

Fig. 6. The terminal fuzzy membership values of AGK.

Table 2
The computing time (s) of FCM, PCM, AFCM and AGK.

FCM	PCM	AFCM	AGK
0.4219	0.6719	0.7187	0.7969

3.8. Computing time

To calculate the computing time, we wrote programs for FCM, PCM, AFCM and AGK with Matlab 7.1 and ran them in the computer: Intel Pentium Dual T2330 CPU 1.6 GHz and 1G RAM. After the FTIR spectra of tea samples were transformed into two-dimensional data by PCA and LDA, four fuzzy clustering algorithms, i.e. FCM, PCM, AFCM and AGK, were performed to cluster the data, and the clustering time of them was shown in Table 2. Before running PCM, AFCM and AGK, FCM should be performed to compute the terminal cluster centers served as the initial cluster centers of PCM, AFCM and AGK. As a result, PCM, AFCM and AGK consumed more time than FCM. Because AGK need to compute the distance norm D_{ikA_i} which is complicated than Euclidean distance in FCM, PCM and AFCM, AGK spent more time than them.

4. Conclusions

In order to discriminate tea varieties correctly, allied Gustafson-Kessel (AGK) clustering was proposed by combining allied fuzzy c-means (AFCM) clustering with Gustafson-Kessel (GK) clustering. FTIR spectra of tea samples were collected by FTIR-7600 infrared spectrometer. After the spectra were processed by MSC, PCA and LDA, they were classified by four fuzzy clustering algorithms: FCM, PCM, AFCM and AGK. The experimental results showed that AGK had the highest clustering accuracies than the others.

Acknowledgements

This research was financially supported by the priority academic program development of Jiangsu Higher Education Institutions, National Science Foundation of China (No. 31471413), Anhui Province Higher Education revitalization plan talent project of 2014 for excellent young talents supporting plan in universities and the key project of Education Department of Sichuan Province (No. 12ZA070).

References

Alamprese, C., Casiraghi, E., 2015. Application of FT-NIR and FT-IR spectroscopy to fish fillet authentication. LWT - Food Sci. Technol. 63, 720–725.
Barni, M., Cappellini, V., Mecocci, A., 1996. Comments on a possibilistic approach to clustering. IEEE Trans. Fuzz. Syst. 4, 393–396.
Bezdek, J.C., Keller, J., Krisnapuram, R., Pal, N., 1999. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer Academic Publishers.
Bezdek, J.C., 1981. Pattern Recognition With Fuzzy Objective Function Algorithms. Plenum Press.
Cai, J.X., Wang, Y.F., Xi, X.G., Li, H., Wei, X.L., 2015. Using FTIR spectra and pattern recognition for discrimination of tea varieties. Int. J. Biol. Macromol. 78, 439–446.
Chandwani, V., Agrawal, V., Nagar, R., 2015. Modeling slump of ready mix concrete using genetic algorithms assisted training of Artificial Neural Networks. Expert Syst. Appl. 42, 885–893.
Chang, C.C., Chou, S.H., 2015. Tuning of the hyperparameters for L2-loss SVMs with the RBF kernel by the maximum-margin principle and the jackknife technique. Pattern Recogn. 48, 3983–3992.
Chen, Q.S., Zhao, J.W., Fang, C.H., Wang, D.M., 2007. Feasibility study on identification of green, black and Oolong teas using near-infrared reflectance spectroscopy based on support vector machine (SVM). Spectrochim. Acta A 66, 568–574.
Chen, Q.S., Zhao, J.W., Lin, H., 2009. Study on discrimination of Roast green tea (Camellia sinensis L.) according to geographical origin by FT-NIR spectroscopy and supervised pattern recognition. Spectrochim. Acta A 72, 845–850.
Costel, S., Katharina, Z., Jürgen, W.E., 2007. Fuzzy divisive hierarchical clustering of soil data using Gustafson-Kessel algorithm. Chemometr. Intell. Lab. 86, 121–129.
Ding, X.X., Ni, Y.N., Kokot, S., 2015. Analysis of different Flos Chrysanthemum tea samples with the use of two-dimensional chromatographic fingerprints, which were interpreted by different multivariate methods. Anal. Method. 7, 961–969.
Diniz, P.H.G.D., Gomes, A.A., Pistonesi, M.F., Band, B.S.F., de Araújo, M.C.U., 2014. Simultaneous classification of teas according to their varieties and geographical origins by using NIR spectroscopy and SPA-LDA. Food Anal. Method. 7, 1712–1718.
He, W., Zhou, J., Cheng, H., Wang, L.Y., Wei, K., Wang, W.F., Li, X.H., 2012. Validation of origins of tea samples using partial least squares analysis and Euclidean distance method with near-infrared spectroscopy data. Spectrochim. Acta A 86, 399–404.
He, Y., Li, X.L., Deng, X.F., 2007. Discrimination of varieties of tea using near infrared spectroscopy by principal component analysis and BP model. J. Food Eng. 79, 1238–1242.
Jiménez-Carvelo, A.M., María Teresa Osorio, M.T., Koidis, A., González-Casado, A., Cuadros-Rodríguez, L., 2017. Chemometric classification and quantification of olive oil in blends with any edible vegetable oils using FTIR-ATR and Raman spectroscopy. LWT-Food Sci. Technol. 86, 174–184.
Kodogiannis, V.S., Kontogianni, E., Lygouras, J.N., 2014. Neural network based identification of meat spoilage using Fourier-transform infrared spectra. J. Food Eng. 142, 118–131.
Krishnapuram, R., Keller, J., 1993. A possibilistic approach to clustering. IEEE Trans. Fuzz. Syst. 1 (2), 98–110.
Li, X.L., Sun, C.J., Luo, L.B., He, Y., 2015. Determination of tea polyphenols content by infrared spectroscopy coupled with iPLS and random frog techniques. Comput. Electron. Agric. 112, 28–35.
Li, X.L., Luo, L.B., He, Y., Xu, N., 2013. Determination of dry matter content of tea by near and middle infrared spectroscopy coupled with wavelet-based data mining algorithms. Comput. Electron. Agric. 98, 46–53.
Liu, Q., Chen, C., Zhang, Y., Hu, Z., 2011. Feature selection for support vector machines with RBF kernel. Artif. Intell. Rev. 36, 99–115.
Lv, S.D., Wu, Y.S., Song, Y.Z., Zhou, J.S., Lian, M., Wang, C., Liu, L., Meng, Q.X., 2015. Multivariate analysis based on GC-MS fingerprint and volatile composition for the quality evaluation of Pu-Erh green tea. Food Anal. Method. 8, 321–333.
Sinija, V.R., Mishra, H.N., 2008. FT-NIR spectro-photometric method for determination of moisture content in green tea granules. Food Bioprocess Tech. 4, 136–141.
Sinija, V.R., Mishra, H.N., 2009. FT-NIR spectroscopy for determination of caffeine in green instant tea powder and tea granules. LWT-Food Sci. Technol. 42, 998–1002.
Sun, Y.B., Wang, J., Cheng, S.M., 2017. Discrimination among tea plants either with different invasive severities or different invasive times using MOS electronic nose combined with a new feature extraction method. Comput. Electron. Agric. 143, 293–301.
Szymczycha-Madeja, A., Welna, M., Pohl, P., 2015. Determination of essential and non-essential elements in green and black teas by FAAS and ICP OES simplified-multi-variate classification of different tea products. Microchem. J. 121, 122–129.
Wang, D.M., Ji, J.M., Gao, H.Z., 2014a. The effect of MSC spectral pretreatment regions on near infrared spectroscopy calibration results. Spectrosc. Spectr. Anal. 34 (9), 2387–2390.
Wang, L.Y., Wei, K., Cheng, H., He, W., Li, X.H., Gong, W.Y., 2014b. Geographical tracing of Xihu Longjing tea using high performance liquid chromatography. Food Chem. 146, 98–103.
Wu, B., Cui, Y.H., Wu, X.H., Jia, H.W., Li, M., 2016. Discrimination of tea varieties using infrared spectroscopy with a novel generalized noise clustering. Spectrosc. Spectr. Anal. 36 (7), 2094–2097.
Wu, X.H., Zhou, J.J., 2006. Allied fuzzy c-means clustering model. Trans. Nanjing Univ. Aeronaut. Astronaut. 23, 208–213.
Wu, X.H., Wu, B., Sun, J., Yang, N., 2016a. Classification of apple varieties using near infrared reflectance spectroscopy and fuzzy discriminant C-means clustering model. J. Food Process Eng. http://dx.doi.org/10.1111/jfpe.12355.
Wu, X.H., Wu, B., Sun, J., Li, M., Du, H., 2016b. Discrimination of apples using near infrared spectroscopy and sorting discriminant analysis. Int. J. Food Prop. 19, 1016–1028.
Yang, Y., Zhang, S.J., He, Y., 2015. Dynamic detection of fresh jujube based on ELM and visible/near infrared spectra. Spectrosc. Spectr. Anal. 35, 1870–1874.
Yaroshenko, I., Kirsanov, D., Kartsova, L., Bhattacharyya, N., Sarkar, S., Legin, A., 2014. On the application of simple matrix methods for electronic tongue data processing: Case study with black tea samples. Sensor. Actuator. B 191, 67–74.
Zhang, L., Deng, W.W., Wan, X.C., 2014. Advantage of LC-MS metabolomics to identify marker compounds in two types of Chinese dark tea after different post-fermentation processes. Food Sci. Biotech. 23, 355–360.
Zhao, N., Wu, Z.S., Cheng, Y.Q., Shi, X.Y., Qiao, Y.J., 2016. MDL and RMSEP assessment of spectral pretreatments by adding different noises in calibration/validation datasets. Spectrochim. Acta A 163, 20–27.
Zhi, R.C., Zhao, L., Shi, B.L., Wang, H.Y., Li, Z., Zhang, J., Xi, X.J., Jin, Y., 2013. Predicting sensory quality of Longjing tea on the basis of physiochemical data. Sensor. Mater. 25, 269–284.