Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 194 (2018) 14–20

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

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

Recognition of genetically modified product based on affinity propagation clustering and terahertz spectroscopy

Jianjun Liu [a,b], Jianquan Kan [a,*]

[a] College of Food Science, Southwest University, Chongqing 400715, China
[b] Guangxi Key Laboratory of Automatic Detecting Technology and Instruments (Guilin University of Electronic Technology), China

ARTICLE INFO

Article history:
Received 7 November 2017
Received in revised form 21 December 2017
Accepted 31 December 2017
Available online 03 January 2018

Keywords:
Terahertz
Affinity propagation clustering
SVM
Genetically modified

ABSTRACT

In this paper, based on the terahertz spectrum, a new identification method of genetically modified material by support vector machine (SVM) based on affinity propagation clustering is proposed. This algorithm mainly uses affinity propagation clustering algorithm to make cluster analysis and labeling on unlabeled training samples, and in the iterative process, the existing SVM training data are continuously updated, when establishing the identification model, it does not need to manually label the training samples, thus, the error caused by the human labeled samples is reduced, and the identification accuracy of the model is greatly improved.

1. Introduction

Terahertz usually refers to electromagnetic wave [1–4] with a frequency between 0.1 Tz–10 THz (wave length 30–3), its band is between microwave and infrared, belonging to far infrared band. Theoretical study shows that the vibrational and rotational energy levels of a large number of biomolecules (DNA, proteins, etc.) are just in the THz band, the THz time domain spectroscopy (THz-TDS) is used for detection of biological samples can produce resonance absorption peak; therefore, it is possible to identify the biological samples by terahertz spectroscopy.

With the popularization of genetically modified technology and genetically modified products, safety detection and safety evaluation of genetically modified products have been paid more and more attention to. Molecular identification methods for genetically modified food at the present stage include PCR [5–10], the method has the characteristics of high sensitivity, but this method requires international standards samples for detection, which may bring inconvenience to the detection of genetically modified products, it is obvious that far infrared spectrum may result in problems of difficult to determine the optimal parameters and the large amount of spectral data when in the detection of genetically modified products. Therefore, it is necessary to develop new detection methods as a supplement. At present, the application of near infrared spectroscopy in trans-gene is quite extensive, Jianjun Liu et al. [11] reported the application of terahertz spectroscopy in detection and identification of transgenic corn oil from corn edible oil, Literature [12] reported the application of terahertz spectroscopy in the detection of identify adulterated dairy products, Literature [13–18] reported the identification of transgenic products by using visible/near infrared spectroscopy. However, the identification of genetically modified food by terahertz is still in its infancy all over the world, therefore, it is of great theoretical and practical significance to identify genetically modified food by terahertz spectroscopy.

2. Materials and Methods

2.1. Experimental Samples

Experimental samples are selected from 230 samples of 5 varieties in Gossypium hirsutism L approved by the country and the province of Shandong, including 2 Non GM varieties, 3 genetically modified varieties (Table 1), all Gossypium hirsutism L cotton lines selected in the experiment had participated in Shandong and national variety exhibition. Details of the sample are shown in Table 1.

2.2. Experimental Equipment

The THz-TDS system used in this paper is a transmission terahertz time-domain spectroscopy system, as shown in Fig. 1. The central wavelength of the laser is 800 nm, in order to ensure the accuracy of the experiment; the system is filled with dry air until the internal relative humidity reaches below 2%. The indoor relative humidity is 25%, constant temperature is 292 K.

* Corresponding author.
E-mail address: liujianjun8888@hotmail.com (J. Liu).

https://doi.org/10.1016/j.saa.2017.12.074
1386-1425/

Table 1
Details of the experimental sample information.

Sample name	Sample type	Diameter	Shape	Thickness	Number of samples
GM LMY28	Genetically modified	1.2 cm	Round slices	1.2 mm	46
GM XQ107	Genetically modified	1.2 cm	Round slices	1.2 mm	46
NGM XLZ6	Non-genetically modified	1.2 cm	Round slices	1.2 mm	46
NGM ZM28	Non-genetically modified	1.2 cm	Round slices	1.2 mm	46
GM YM8	Genetically modified	1.2 cm	Round slices	1.2 mm	46

3. Data Processing Method

3.1. Spectral Analysis

Terahertz spectra of 5 different samples are given in Fig. 2. As seen from the Fig. 2, there is a big difference in terahertz characteristic spectra between genetically modified cotton varieties (GM LMY28, GM XQ107, GM YM8) and non genetically modified cotton varieties (NGM XLZ6, NGM ZM28), which can be easily distinguished from the THz characteristic spectrum, but the comparison of terahertz spectral characteristics between genetically modified cotton varieties (or non genetically modified cotton varieties) is similar, it is difficult to directly distinguish from the THz spectra, pattern recognition methods must be used to identify.

3.2. Reflection Propagation Clustering Supporting Vector Machine

The main idea of the affinity propagation clustering support vector machine is: affinity propagation clustering is carried out on unlabeled sample data, all cluster centers of each loop cluster and representative points are updated, until the termination condition s is satisfied (S = 0 unlabeled sample points in the samples, S = 1 no unlabeled sample points in the sample), finally, the support vector machine (SVM) is trained by the training samples set, and the discrimination model is established and the discriminant analysis is performed. The main features of the improved algorithm are SVM cannot only learn the most informative class representative points, but also mark the unlabeled samples with a certain amount of information, its principle frame is shown in Fig. 3.

In Fig. 3, L and U represent labeled and unlabeled samples respectively. When the algorithm is initial, L is the initial training set, and U is all unlabeled samples.

Implementation steps of support vector machine algorithm based on affinity propagation clustering:

① The training set L, the sample point set U to be clustered, the class representative point K, the number of cluster categories Num = K and the termination condition S (S = 0 unlabeled sample points in the samples, S = 1 no unlabeled sample points in the sample) are initialized.

② The unlabeled sample set U is added into reflection propagation clustering algorithm, and adaptive affine clustering algorithm is adopted to make cluster analysis of unlabeled samples;

③ Update the training set L and unlabeled sample set U by using clustering results;

④ Determine whether the sample set U contains unlabeled sample points, if any, making S = 0, select the most valuable class representative point M and execute the step ②; on the contrary, making S = 1 and execute the step ⑤;

⑤ Output optimal training set, training SVM to establish the identification model.

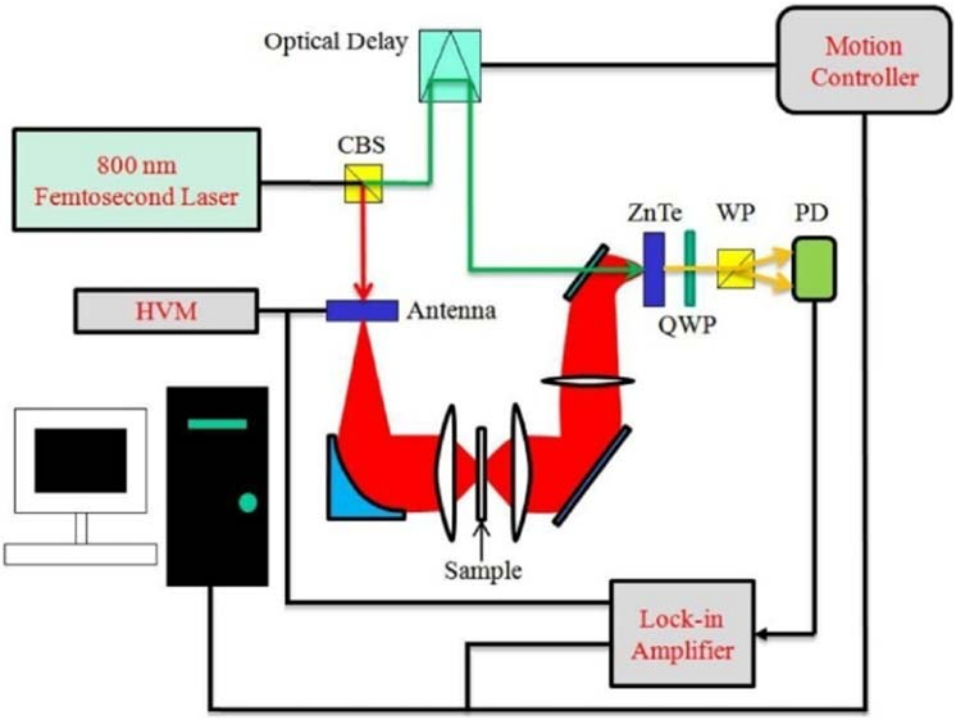

Fig. 1. The schematic diagram of THz-TDS system.