Optik 125 (2014) 6914–6919

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

Classification and recognition of transgenic product by terahertz spectroscopy and DSVM

Jianjun Liu [a], Zhi Li [a,b,*], Fangrong Hu [c], Tao Chen [c], Aijun Zhu [c]

[a] School of Mechano-Electronic Engineering, Xidian University, Xi'an, Shanxi 710071, PR China
[b] Guilin University of Aerospace Technology, Guilin, Guangxi 541004, PR China
[c] School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin, Guangxi 541004, PR China

ARTICLE INFO

Article history:
Received 23 November 2013
Accepted 20 June 2014

Keywords:
DSVM
PSO
Terahertz
Recognition
Transgenic

ABSTRACT

The purpose of this paper is to construct a classification model that can identify class accurately and control imbalance. A novel adaptive decision support vector machine (DSVM) is proposed for the recognition of transgenic cotton seed based on terahertz spectroscopy (THz), which make the traditional support vector machine is ability of adaptive decision, and select optimal parameters by using particle swarm optimization (PSO). For the classification and recognition of the transgenic cotton seeds, firstly, the factor analysis (FA) is applied to reduce the dimension and extract the feature spectrum of original spectral information. Secondly, the feature spectrum is selected and fed into the model of DSVM to recognize the different transgenic cotton seeds. The experimental results show that the proposed method can effectively classify the different transgenic cotton seeds, and its recognition rate surpasses the comparative method evidently.

1. Introduction

Terahertz usually refers to electromagnetic wave with the frequency in 0.1 Tz–10 THz (wavelength 30 μm–3 mm), and the band between microwave and infrared, which belongs to far-infrared band. Theoretical studies show that the vibration and rotational energy levels of most biological molecules (DNA, protein) are in THz band. Thus, it is possible to using THz time-domain spectroscopy (THz-TDS) detection and distinguishes biological [1–4].

With the popularization of transgenic technology and transgenic products, safety inspection and assessment of transgenic food attracts more and more attention. Up to now, most research focuses on the security check of explosives [5,6]. In food safety, especially, the transgenic food detection has not yet formed a complete system in THz field. Because of the transgenic product safety and environmental protection etc., caused a consumer boycott, the governments of all countries have introduced new regulations and some related tests are in continuous improvement [7–9]. At present, there are several commonly used transgenic testing protocols mainly including PCR and ELISA detection methods [10–15]. But those methods are impotent for detection of the other changes which caused by gene implant. Hence, it is very important to develop a new method to detection transgenic food as a supplement.

The goal of this study is to use THz spectroscopy, combined with spectral pretreatment technique and pattern recognition method to classify the transgenic cotton seed. It is significant and valuable for detection transgenic food, and lays the foundation for the gene product testing.

2. Theory and algorithm

2.1. Particle swarm optimization

Particle swarm optimization (PSO), which models the social behavior of groups of animals, is one of the modern heuristic algorithms for optimization. The algorithm, originally proposed by Kennedy and Eberhart in 1995 [16], is designed to avoid local minima by having a group of search directions that follow the optimal direction [17].

2.2. Support vector machine

A support vector machine is used here to build classifiers for three transgenic cotton seeds. SVM, a learning machine based on statistical learning theory, is first proposed by Vapnik in 1995 [18]. An SVM based on the Vapnik-Chervonenkis (VC) theory and

* Corresponding author at: School of Mechano-Electronic Engineering, Xidian University, Xi'an, Shanxi 710071, China. Tel.: +86 773 2191029; fax: +86 773 2191029.
E-mail address: liujianjunworkemail@gmail.com (Z. Li).

http://dx.doi.org/10.1016/j.ijleo.2014.08.054
0030-4026/

structural risk minimization (SRM) principle balances minimizing the generalization error and maximizing the geometric margin, aiming to achieve the best generalization ability [2,3,19–21].

Consider the optimization problem which the decision optimal is obtained when the classification margin is maximization, that is:

$$\begin{cases} \min \dfrac{1}{2} \parallel \omega \parallel^2 + \lambda U(m_j, C_j, \theta) \\ s.t \quad y_i^j[(\omega \cdot x_i^j) + b] \geq 1 - \xi_i^j, \quad i = 1, 2, \ldots, l, \quad j = 1, 2 \\ \xi_i^j \geq 0, \quad i = 1, 2, \ldots, l, \quad j = 1, 2 \end{cases} \tag{1}$$

for which λ is a parameter that can balance the decision loss minimization with classification margin maximization, U express the decision function that make the learning process consistent with risk preference, m_j reflect the imbalance of sample and make the model is ability of adaptive, C_j is describe decision of model.

2.3. Adaptive decision support vector machine

Because of the SVM is integrated into a goal of classification error in classification rule and the classification error minimization is NP completely [22]. Therefore, the traditional SVM is not to find the optimal solution by two or convex programming method. The model (1) is able to determine whether the classification is error or not by using 0–1 step function of the distance between point and classification boundary. The 0–1 step function is not differentiable, but can be approximated to the smoothing by the exponential function at any precision. So this NP complete problem is transformed into a general optimization problem by gradient descent method solution. An approximate smooth method for non negative variables is adopted in this paper [23]: $t(x, p) = 1 - \exp(-px), p > 0, x \geq 0$.

After approximation smoothing, the optimization problem (1) is equivalent to the follow formula (2):

$$\min \frac{1}{2} \parallel \omega \parallel 2 + \lambda \sum_{j=1}^{2} \sum_{i=1}^{l} \frac{Cj}{mj} (1 - \exp(-p\xi_i^j)) \tag{2}$$

The Lagrange function of formula (2) is described as:

$$L(\omega, b, \xi_i^j) = \frac{1}{2} \parallel \omega \parallel 2 + \lambda \sum_{j=1}^{2} \sum_{i=1}^{l} \frac{Cj}{mj} (1 - \exp(-p\xi_i^j)) - \sum_{j=1}^{2} \sum_{i=1}^{l} \alpha_i^j [y_i^j ((\omega, x_i^j) + b) - 1 + \xi_i^j] - \sum_{j=1}^{2} \gamma_i^j \xi_i^j \tag{3}$$

where γ_i^j and ξ_i^j are the Lagrange multiplier, and the minimization of (3) must meet equation that $\partial L/\partial\omega = \partial L/\partial b = \partial L/\partial\xi_i^j = 0$. Hence, we can get the following formula:

$$\begin{cases} \omega = \sum_{j=1}^{2} \sum_{i=1}^{l} y_i^j \alpha_i^j x_i^j \\ \sum_{j=1}^{2} \sum_{i=1}^{l} y_i^j \alpha_i^j = 0 \\ \dfrac{\lambda C_j p}{m_j} \exp(-p\xi_i^j) = \alpha_i^j + \gamma_i^j, \quad j = 1, 2 \end{cases} \tag{4}$$

for which p is a large enough positive which make the function $1 - \exp(-p\xi_i^j)$ approximate to the step functions $\theta(\xi_i^j)$ in arbitrary precision.

Let $\delta_i^j = \alpha_i^j + \gamma_i^j$ then $\exp(-p\xi_i^j) = \frac{m_j}{\lambda C_j p}\delta_i^j$, $j = 1, 2$, and the dual form of formula (4) are defined as:

$$\begin{cases} \max W = \sum_{j=1}^{2} \sum_{i=1}^{l} \alpha_i^j - \sum_{j=1}^{2} \sum_{i=1}^{l} \sum_{k=1}^{l} \alpha_i^j \alpha_k^j y_i^j y_k^j (x_i^j, x_k^j) - \dfrac{1}{p} \sum_{j=1}^{2} \sum_{i=1}^{l} (1 + \ln \dfrac{\lambda C_j p}{m_j} - \ln \delta_i^j) \delta_i^j \\ s.t. \quad \sum_{j=1}^{2} \sum_{i=1}^{l} y_i^j \alpha_i^j = 0 \\ \delta_i^j = \alpha_i^j + \gamma_i^j, \quad i = 1, 2, \ldots, l, \quad j = 1, 2 \\ \alpha_i^j \geq 0, \quad i = 1, 2, \ldots, l, \quad j = 1, 2 \\ \gamma_i^j \geq 0, \quad i = 1, 2, \ldots, l, \quad j = 1, 2 \end{cases} \tag{5}$$

In the nonlinear case, the data is mapped to a feature space by a linear transformation, and the inner product of feature space is replaced by kernel transformation $K(x_i^j, x_k^j) = \varphi(x_i^j)\varphi(x_k^j)$. The dual form of optimization problems in feature space is described as:

$$maxW = \sum_{j=1}^{2} \sum_{i=1}^{l} \alpha_i^j - \sum_{j=1}^{2} \sum_{i=1}^{l} \sum_{k=1}^{l} \alpha_i^j \alpha_k^j y_i^j y_k^j K(x_i^j, x_k^j) - \frac{1}{p} \sum_{j=1}^{2} \sum_{i=1}^{l} (1 + \ln \frac{\lambda C_j p}{m_j} - \ln \delta_i^j) \delta_i^j \tag{6}$$

Solve the optimization problem (5) or (6) is that you can find the optimal separating hyper plane and obtain the solution of problems.

2.4. Algorithm process

The procedure of the proposed approach is based on the DSVM, which the parameters are optimized by PSO is presented as follows:

(1) Particle initialization and set PSO parameters;
(2) Data preparation: normalize the feature values of the data set to the range [−1, +1], generate the train set *Tr* and test set *Te* by cross validation rule, scan *Tr* and *Te*, identify those feature data m_1, m_2, T_1, T_2, input the decision parameters C_1 and C_2;
(3) Obtain the *I*-th generation of particle swarm S_I and let $i = 1$;
(4) Train the DSVM model: read the particle *i*, calculate the fitness rate f_i of particle *i* and output the rate of validation error and support vector by using cross validation rule;
(5) Judgment: if $f_i \geq f^*$ then go to step (8). Otherwise, let $i = i + 1$ and go to step (6);
(6) Judgment: if $i \leq N$ then return to step (5). Otherwise, go to step (7);