Heliyon 10 (2024) e27743

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Interval-based sparse ensemble multi-class classification algorithm for terahertz data

Chengyong Zheng [a], Xiaowen Zha [a], Shengjie Cai [b], Jing Cui [c,*], Qian Li [d], Zhijing Ye [e,*]

[a] School of Mathematics and Computational Science, Wuyi University, Jiangmen, 529000, China
[b] Shenzhen Kangguan Technology Co., LTD, Shenzhen, 518129, China
[c] Guangdong Jiangmen Chinese Traditional Medicine College, Jiangmen, 529020, China
[d] Terahertz Technology Application (Guangdong) Co., Ltd, Guangzhou, 510700, China
[e] Faculty of Innovation Engineering, Macau University of Science and Technology, Taipa, Macau

ARTICLE INFO

Keywords:
Terahertz spectrum
Classification
Sparse ensemble
Interval
Cross entropy

ABSTRACT

Terahertz time-domain spectroscopy (THz-TDS) has been widely used for food and drug identification. The classification information of a THz spectrum usually does not exist in the whole spectral band but exists only in one or several small intervals. Therefore, feature selection is indispensable in THz-based substance identification. However, most THz-based identification methods empirically intercept the low-frequency band of the THz absorption coefficients for analysis. In order to adaptively find out important intervals of the THz spectra, an interval-based sparse ensemble multi-class classifier (ISEMCC) for THz spectral data classification is proposed. In ISEMCC, the THz spectra are first divided into several small intervals through window sliding. Then the data of training samples in each interval are extracted to train some base classifiers. Finally, a final robust classifier is obtained through a nonnegative sparse combination of these trained base classifiers. With l_1 -norm, two objective functions that based on Mean Square Error (MSE) and Cross Entropy (CE) are established. For these two objective functions, two iterative algorithms based on the Alternating Direction Method of Multipliers (ADMM) and Gradient Descent (GD) are built respectively. ISEMCC transforms the problem of interval feature selection and decision-level fusion into a nonnegative sparse optimization problem. The sparse constraint ensures only a few important spectral segments are selected. In order to verify the performance of the proposed algorithm, comparative experiments on identifying the origin of Bupleurum and the harvesting year of Tangerine peel are carried out. The base classifiers used by ISEMCC are Support Vector Machine (SVM) and Decision Tree (DT). The experimental results demonstrate that the proposed algorithm outperforms six typical classifiers, including Random Forest (RF), AdaBoost, RUSBoost, ExtraTree, and the two base classifiers, in terms of classification accuracy.

* Corresponding authors.
E-mail addresses: 12810626@qq.com (J. Cui), xkinghust@163.com (Z. Ye).

https://doi.org/10.1016/j.heliyon.2024.e27743
Received 25 August 2023; Received in revised form 24 February 2024; Accepted 6 March 2024

1. Introduction

Electromagnetic wave with a frequency range of 0.1-10 terahertz (THz) is referred to as terahertz wave [1]. Due to its characteristics of penetration, fingerprint, and non-ionization, THz spectroscopy has demonstrated great potential in non-destructive detection, especially in the fields of food, medicine, environmental protection, and some industrial fields [2–5]. With THz spectra, Friska et al. [6] utilized fast independent component analysis (ICA) and random forest (RF) algorithms to identify adulterated rice. Pan et al. [7] applied the support vector machine (SVM) with an enhanced cuckoo search algorithm to classify ginseng of different growth ages. Zhu et al. [8] claimed that employing a noise reduction technique and a reconstruction approach can successfully tackle the problem of low spectral signal-to-noise ratio produced by different components of the biological mixture, including water. Liu et al. [9] proposed to use principal component analysis (PCA), local preservation protection (LPP), and Isomap methods to reduce the dimensionality of the THz spectral data, and then use the probabilistic neural network (PNN) and SVM to identify liver tumors. Huang et al. [10] used nonnegative matrix factorization (NMF) to decompose THz spectral data into the product of the weight matrix and characteristic matrix, and processed the weight matrix by K-means clustering to classify biological macromolecules. Liu and Zhang et al. [11] proposed a novel method combining THz time-domain spectroscopy (THz-TDS) with chemometrics for the classification of sand samples. They employed Savitzky-Golay (SG) smoothing and orthogonal signal correction (OSC) to pretreat spectra, and applied PCA, partial least squares (PLS) discriminant analysis and SVM to establish classification models for distinguishing sand samples from various deserts and grain sizes. Zhu and Wang et al. [12] used logistic regression (LR), SVM, and RF to classify the oxidation degree of coal, and provided a coal spontaneous combustion monitoring technology combining THz permittivity spectrum and machine learning algorithm. Zhang and Li et al. [13] used PCA to reduce the dimensionality of original THz spectral information, and then employed SVM, decision tree (DT), and RF to discriminate herbal medicines. Sarja et al. [14] proposed a classification method for plastic inorganic pigments based on THz spectroscopy and convolutional neural networks (CNN). Huang and Cao et al. [15] used THz spectroscopy to inspect mouse liver injury. They utilized the maximal information coefficient to select crucial features from the absorption coefficients and the refractive index spectra, and applied RF and AdaBoost to recognize different levels of liver injury.

In conclusion, the feasibility of THz technology has been demonstrated in several areas. However, most studies have focused on dimension reduction or feature extraction, which are widely used in machine learning. Some studies have even explored cutting low-frequency segments directly. Due to the highly sensitive and noisy nature of THz data, classification information is often concealed within one or several characteristic peaks that account for only a small portion of the entire spectral data. The remaining data typically contains limited useful information. Using the whole spectrum data for dimension reduction (e.g., PCA) and feature extraction (e.g., NMF) will inevitably affect the classification performance.

In our previous work [16], a collaborative classification algorithm with multiple THz spectra (MVTHzCC) is put forward. Its feature selection is based on an optimal interval search followed by complementary feature seeking. A decision-level weighted fusion was utilized to make full use of the information provided by various THz spectra. MVTHzCC is good for ensuring that valid classification features in the THz spectra are found. However, its feature search process is time-consuming.

In order to efficiently search out the valid classification features in the THz spectral data, and consider the curve characteristics of THz spectra, an interval-based sparse ensemble multi-class classifier (ISEMCC) for THz spectral data classification is proposed. ISEMCC converts the problem of characteristic peak search and selection in the THz spectral data classification into the optimal sparse combination problem of the trained classifiers by THz interval data. ISEMCC has the following advantages:

- Through sparse optimization, the selection of optimal interval features and optimal decision level fusion is realized at the same time.
- The selection of the base classifier is flexible and has wide applicability.
- Compared with the base classifier, the proposed algorithm can significantly improve the classification accuracy.

The rest of this paper is organized as follows. Section 2 presents the proposed method. Section 3 reports experimental results. Section 4 concludes this paper.

2. Interval-based sparse ensemble multi-class classifier

In THz-based substance identification, the identification information of substances is usually hidden in one or more small intervals of THz spectral data. How to find these small intervals is the key to THz-based classification and recognition. Hence, we provide an interval-based sparse ensemble multi-class classifier (ISEMCC). First, ISEMCC divides the THz spectral data into many small intervals. Then it uses the data of training samples in each interval to train some base classifiers. Finally, the final strong classifier is achieved by a nonnegative sparse combination of the trained base classifiers. Thus, the optimal interval searching and the optimal decision level fusion of interval classifiers are acquired at the same time.

2.1. Symbols

Let $X_{trn} = \{(x_1, y_1), \cdots, (x_{n_{trn}}, y_{n_{trn}})\}$ be the training THz spectral data, where, n_{trn} is the number of training data, and $y_i \in R^C$ is the one-hot coding vector of the label of the i-th training data, i.e., all the entries are zero except the c-th if x_i belongs to the c-th class, C is the total number of classes. Let f be a base classifier. In order to gather small intervals from THz spectral data, a sliding window, and its sliding step should be given. Assuming a total of n_b intervals have been obtained by sliding a small window with

width w and sliding step h, a trained base classifier f_j $(j = 1, 2, \cdots, n_b)$ can be obtained by use of the training data that spectra belong to the j-th interval. Let $f_j(x_i) = p_j^i \in R^C$ be the output of f_j on x_i, $Y_i = [p_1^i, p_2^i, ..., p_{n_b}^i]$ be the output of f on all n_b intervals of x_i. It should be noted here that $p_j^i \in R^C$ can be either a one-hot encoded vector or a probability prediction vector, depending on whether the base classifier makes a probability prediction. We'll demonstrate in subsection 3.2 that probability prediction is not necessary, a general class prediction is sufficient.

2.2. Two objective functions

How to find a few intervals from all the n_b intervals so that the decision-level fusion of the classifiers trained on these interval segments is optimal? This problem is equivalent to finding sparse non-negative vectors $\alpha \in R^{n_b}$ such that $Y_i\alpha$ and y_i $(i = 1, \cdots, n_{Trn})$ are as consistent as possible.

In order to measure the consistency of $Y_i\alpha$ and y_i, the Mean squared error (MSE) method and Cross entropy (CE) method are discussed in this paper.

For MES method, $\|Y_i\alpha - y_i\|$ is adopted to measure the consistency of $Y_i\alpha$ and y_i, and the optimization objective function is defined as follows

$$L_{MSE}(\alpha) = \frac{1}{2n_{trn}} \sum_{i=1}^{n_{trn}} \|Y_i\alpha - y_i\|^2 + \lambda_{mse}\|\alpha\|_1, \alpha \geq 0, \tag{1}$$

where $\lambda_{mse} > 0$ is a regular parameter.

For the CE method, the CE loss function widely used in neural network training is adopted [17]. In fact, $Y_i\alpha$ can be regarded as the output of the last layer of a neural network, and the SoftMax function can be used to obtain the probability output of $Y_i\alpha$. Given a vector $x \in R^n$, the SoftMax function is defined as follows

$$\mathrm{softmax}(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}.$$

Let $z_i = \mathrm{softmax}(Y_i\alpha)$, the optimization objective function of the CE method is defined as follows

$$L_{CE}(\alpha) = -\sum_{i=1}^{n_{trn}} \sum_{k=1}^{C} y_i(k) \log[z_i(k)] + \lambda_{ce}\|\alpha\|_1, \alpha \geq 0, \tag{2}$$

where the $\lambda_{ce} > 0$ is a regular parameter, and $z_i(k)$ are the k-th entry of vector z_i.

2.3. Solutions to two objective functions

This subsection provides solutions to the two optimization problems $\min_\alpha L_{MSE}(\alpha)$ and $\min_\alpha L_{CE}(\alpha)$.

2.3.1. MSE method

The close-form solution of $\min_\alpha L_{MSE}(\alpha)$ cannot be obtained directly. In order to solve $\min_\alpha L_{MSE}(\alpha)$, the alternating direction method of multipliers (ADMM) [18,19] is applied.

First, a free variable υ is introduced, and the optimization problem of (1) is transformed to

$$min_\alpha \frac{1}{2n_{trn}} \sum_{i=1}^{n_{trn}} \|Y_i\alpha - y_i\|^2 + \lambda_{mse}||\upsilon||_1, \text{ s.t. } \upsilon = \alpha, \upsilon \geq 0.$$

The Lagrangian function is defined as follows:

$$L(\alpha, u, \upsilon) = \frac{1}{2n_{trn}} \sum_{i=1}^{n_{trn}} \|Y_i\alpha - y_i\|^2 + \lambda_{mse}||\upsilon||_1 + \frac{\mu}{2}||\upsilon - \alpha + u||^2 + l^+(\upsilon),$$

where u is the Lagrange multiplier, μ is the non-negative penalty parameter, and $l^+(\upsilon)$ is the indicative function: when $\upsilon \geq 0$, $l(\upsilon) = 0$, otherwise $l(\upsilon) = +\infty$.

By setting

$$\frac{\partial L}{\partial \alpha} = \frac{1}{n_{trn}} \sum_{i=1}^{n_{trn}} Y_i{}^T(Y_i\alpha - y_i) + \mu(\alpha - \upsilon - u) = 0,$$

we get

$$(\sum_{i=1}^{n_{trn}} Y_i{}^T Y_i + n_{trn}\mu I)\alpha = \sum_{i=1}^{n_{trn}} Y_i{}^T y_i + n_{trn}\mu(\upsilon + u),$$

where I is the identity matrix. The above formula gives

$$\alpha = (\sum_{i=1}^{n_{trn}} Y_i^T Y_i + n_{trn}\mu I)^{-1}[\sum_{i=1}^{n_{trn}} Y_i^T y_i + n_{trn}\mu(\upsilon + u)].$$

Similarly, by setting the first-order partial derivative of $L(\alpha, u, \upsilon)$ with respect to υ to be 0, we get

$$\upsilon = S^+_{\frac{\lambda_{mse}}{\mu}}(\alpha - u) = \begin{cases} \alpha - u - \frac{\lambda_{mse}}{\mu}, & \alpha - u > \frac{\lambda_{mse}}{\mu}, \\ 0, & else. \end{cases}$$

According to the ADMM algorithm [18,19], Algorithm 1 can be obtained.

Algorithm 1 ADMM for $\min_\alpha L_{MSE}(\alpha)$.

Input: Error bound of iteration ε, maximum number of iterations n_{iter}.
1: Initialize $\upsilon_0 = \mathbf{0}$, $u_0 = \mathbf{0}$, $k = 0$;
2: **repeat**
3: $\alpha_{k+1} = (\sum_{i=1}^{n_{trn}} Y_i^T Y_i + n_{trn}\mu I)^{-1}[\sum_{i=1}^{n_{trn}} Y_i^T y_i + n_{trn}\mu(\upsilon_k + u_k)]$;
4: $\upsilon_{k+1} = S^+_{\frac{\lambda_{mse}}{\mu}}(\alpha_{k+1} - u_k) = \begin{cases} \alpha_{k+1} - u_k - \frac{\lambda_{mse}}{\mu} & \alpha_{k+1} - u_k > \frac{\lambda_{mse}}{\mu} \\ 0 & else \end{cases}$;
5: $u_{k+1} = u_k + \upsilon_{k+1} - \alpha_{k+1}$;
6: $k = k + 1$;
7: **until** $\|\upsilon_{k+1} - \upsilon_k\| < \varepsilon$ or $k > n_{iter}$
Output: $\alpha = \upsilon_k$.

2.3.2. CE method

The close-form solution of $\min_\alpha L_{CE}(\alpha)$ also cannot be directly obtained. So gradient descent algorithm is adopted.

It can be inferred that the gradient of $L_{CE}(\alpha)$ is [20]

$$\nabla L_{CE} = \sum_{i=1}^{n_{trn}} Y_i^T(\mathrm{softmax}(Y_i\alpha) - y_i) + \lambda_{ce}\mathrm{sign}(\alpha),$$

where sign(·) is the sign function.

Initializing the weight α and the learning rate τ, according to the gradient descent algorithm, an iterative algorithm for solving $\min_\alpha L_{CE}(\alpha)$ can be described as Algorithm 2.

Algorithm 2 Gradient descent algorithm for $\min_\alpha L_{CE}(\alpha)$.

Input: Learning rate τ, error bound of iteration ε, maximum number of iterations n_{iter}. Initialize: $t = 0$, $\alpha_0(i) = \frac{1}{n_b}$ for $i = 1, 2, \cdots, n_{tm}$;
1: **repeat**
2: $\alpha_{t+1} = \alpha_t - \tau[\sum_{i=1}^{n_{trn}} Y_i'(\mathrm{softmax}(Y_i\alpha_t) - y_i) + \lambda_{ce} * \mathrm{sign}(\alpha_t)]$;
3: $\alpha_t = \max(\alpha_t, 0)$;
4: $t = t + 1$;
5: **until** $\|\alpha_{t+1} - \alpha_t\| < \varepsilon$ or $t > n_{iter}$.
Output: $\alpha = \alpha_t$.

2.4. Strong classifier construction

Using the α calculated by Algorithm 1 or Algorithm 2, a strong classifier F can be constructed as follows

$$F = \sum\nolimits_{j=1}^{n_b} \alpha_j f_j.$$

Given a test sample x, we get $F(x) = \sum_{j=1}^{n_b} \alpha_j f_j(x)$, and the final output of F is $\arg\max F(x)$.

3. Experiments and analysis

In order to verify the effectiveness of the proposed algorithm, in this section, some experiments are carried out to identify the origin of Bupleurum and the year of Tangerine peel based on THz spectral data. SVM (linear kernel) and DT are used as the base classifieres for ISEMCC. With the MSE method, the proposed algorithms using SVM and DT as the base classifiers are denoted by ISEMCC(MSE+SVM) and ISEMCC(MSE+DT), respectively. Similarly, ISEMCC(CE+SVM) and ISEMCC(CE+DT) denote the proposed algorithms using SVM and DT as base classifiers but applying the CE method, respectively.

3.1. Data preparation and experimental setup

The acquisition of experimental data is mainly divided into Three steps. First, some Bupleurum samples with different origins and some Tangerine peel samples with different harvesting years are gathered. Second, for Bupleurum and Tangerine peel, 10 and 6

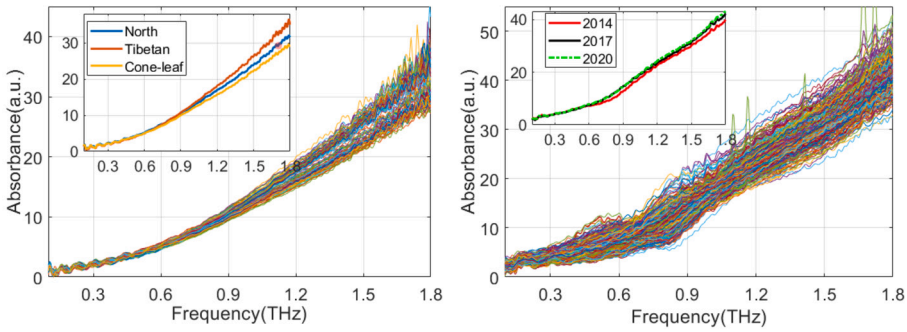

Fig. 1. THz Spectral Curves of Bupleurum (left) and Tangerine Peel (right), each contain a subfigure of the average THz spectral curves.

Table 1
Classes, sample numbers, and batch numbers of two THz spectral datasets.

Data Name	Bupleurum			Tangerine peel		
Label	North	Tibetan	Cone-leaf	2014	2017	2020
Batch Number	3	5	5	12	18	18
Sample Number	30	50	50	72	108	108

parallel circular thin flakes per batch of samples are prepared, respectively. In order to produce a circular thin flake, an appropriate amount of samples are taken firstly for crushing, grinding, and 100 mesh sieving to obtain some sieved powder, and about 0.1 g sieved powder is then weighed and put into a tableting mold to be pressed with a pressure of 24 MPa for 30 s to produce a circular thin flake with a diameter of about 13 mm and a thickness of about 1 mm. Lastly, a transmission mode THz-TDS commercial device (LZ9000 Terahertz Technology Application (Guangdong) Co., Ltd, Guangzhou, China) is used to gather the THz spectral data from these circular thin flakes. In order to collect THz spectral data, a circular thin flake is first placed in a mold, and then the mold is placed in the sample chamber of THz-TDS. Nitrogen is continuously blown in the sample chamber to ensure the dryness of the acquisition environment.

Fig. 1 shows the spectral curves of Bupleurum and Tangerine peel. The subgraphs in the left and right figures show the average THz spectral curves of Bupleurum and Tangerine Peel, respectively.

Table 1 shows the number of batches and samples of each category of the collected THz spectral data of Bupleurum and Tangerine peel. The absorption coefficients in the frequency range of 0.1-1.8 THz are used in the experiments.

To evaluate the performance of a model, the cross-validation method named leave-one-batch-out (LOBO) [16] is used. LOBO uses the following approach:

1. Splits a dataset into a training set and a testing set, using all but one batch of observations as part of the training set.
2. Build a model using only data from the training set.
3. Use the model to predict the response values of the one batch of observations left out of the model, and calculate the accuracy.
4. Repeat this process until each batch of observations have been tested once.
5. Calculate the overall accuracy (OA) using all the predicted response values of all observations.

The reason why the conventional random sample division method is not used is that the conventional random sample division method has the phenomenon of sample leakage, which will lead to falsely high experimental accuracies [16]. This is because the THz spectral data collected from multiple slices of the same batch (belonging to a class) are very similar. Conventional random sample division will divide part of these very similar samples into a training sample set and the rest into a test sample set, causing the problem of sample leakage. In view of the fact that in the actual classification process, it is impossible to have a batch of samples, some of which belong to the training sample set and some of which belong to the test sample set. Therefore, the batch leave-one-out method is more tally with the actual.

All experimental code is written in MATLAB. For the base classifier SVM (linear kernel), the box constraint is set to 100, and one-versus-one coding design is adopted to deal multiclass problem. All parameters of DT are set by default values.

The hyperparameters λ, μ, τ, n_{iter} and ε of ISEMCC are set to 0.01, 10, 0.1, 200 and 10^{-4}, respectively. The window width and sliding step are selected from {31, 41, 51, 61, 71, 81, 91, 101, 111, 121, 131, 141, 151, 161, 171, 181, 191, 201, 211} and {20, 50, 80}, respectively.

3.2. Experimental results

In this subsection, we will first provide the classification comparison results of ISEMCC when the base classifier outputs class prediction and probabilistic prediction, and then demonstrate the relationship between ISEMCC's training accuracy and the $alpha$

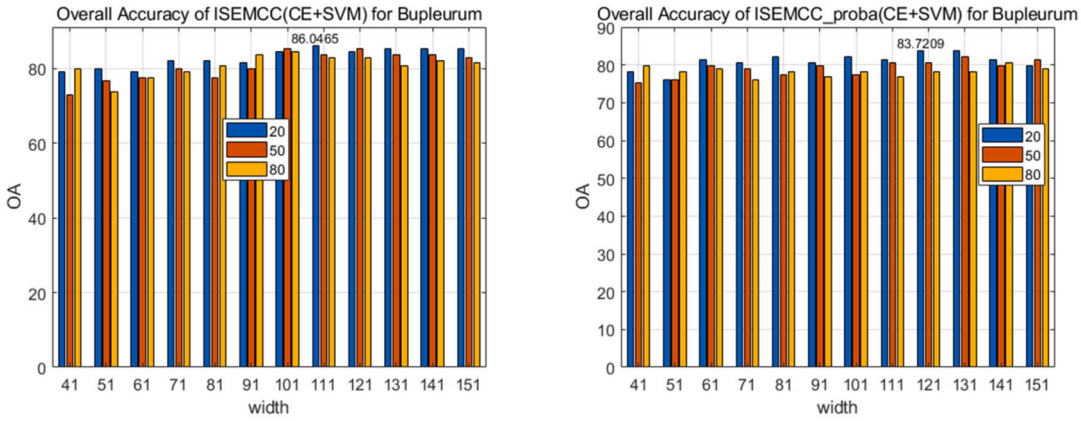

Fig. 2. Results of comparative experiment between "one-hot" scheme (left) and "probability" scheme (right) on Bupleurum dataset.

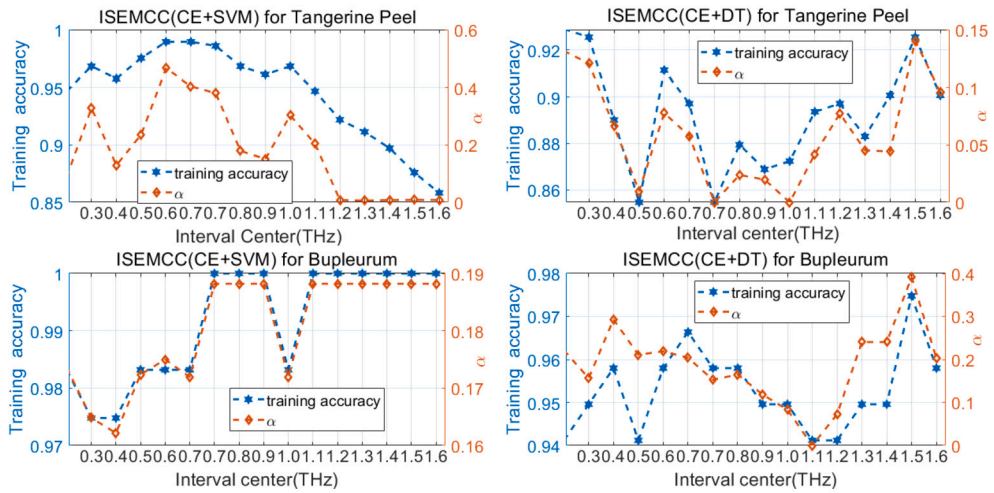

Fig. 3. Training accuracies and values of α in each interval obtained by ISEMCC(CE+SVM) and ISEMCC(CE+DT) on dataset Bupleurum and Tangerine peel.

value on each interval, followed by some test results for window width and sliding step size. Finally, we will show some comparative experimental results.

3.2.1. Comparison between class prediction and probabilistic prediction

In order to implement ISEMCC, the first question we need to decide is whether to choose class prediction ("one-hot" scheme) or probability prediction ("probability" scheme) as the output of base classifier. Therefore, taking ISEMCC (CE+SVM) as an example, we compared the "one-hot" scheme and "probability" scheme on Bupleurum dataset. Fig. 2 shows the results, from which we can see that neither scheme shows significant advantages, and the highest accuracy of the "one-hot" scheme is higher than that of the "probability" scheme. Based on this observation, "one-hot" scheme is adopted in the subsequent experiments unless otherwise stated.

3.2.2. Relationship between ISEMCC's training accuracy and the α value on each interval

Vectors α in formula (1) and (2) determines the weight of each base classifier trained by data belong to each interval. Fig. 3 presents the training accuracies and values of α obtained by ISEMCC(CE+SVM) and ISEMCC(CE+DT) on dataset Bupleurum and Tangerine peel. The window width and sliding step are set by 71 and 20, respectively.

Fig. 3 suggests that, THz data in different intervals have different separability, and in each interval, the value of α is roughly positively correlated with the training accuracy of the base classifier and exhibits a certain sparsity (some values of α are zero). In addition, in each interval, base classifier SVM and DT demonstrate different classification performance. For example, for Tangerine peel, the classification performance of SVM is good at 0.6 and 0.7 THz but decreases significantly after 1.2 THz, while DT demonstrates good performance at 0.2 and 1.5 THz and deteriorates at 0.5 and 0.7 THz. These make the proposed ISEMCC a classification method that can be interpreted.

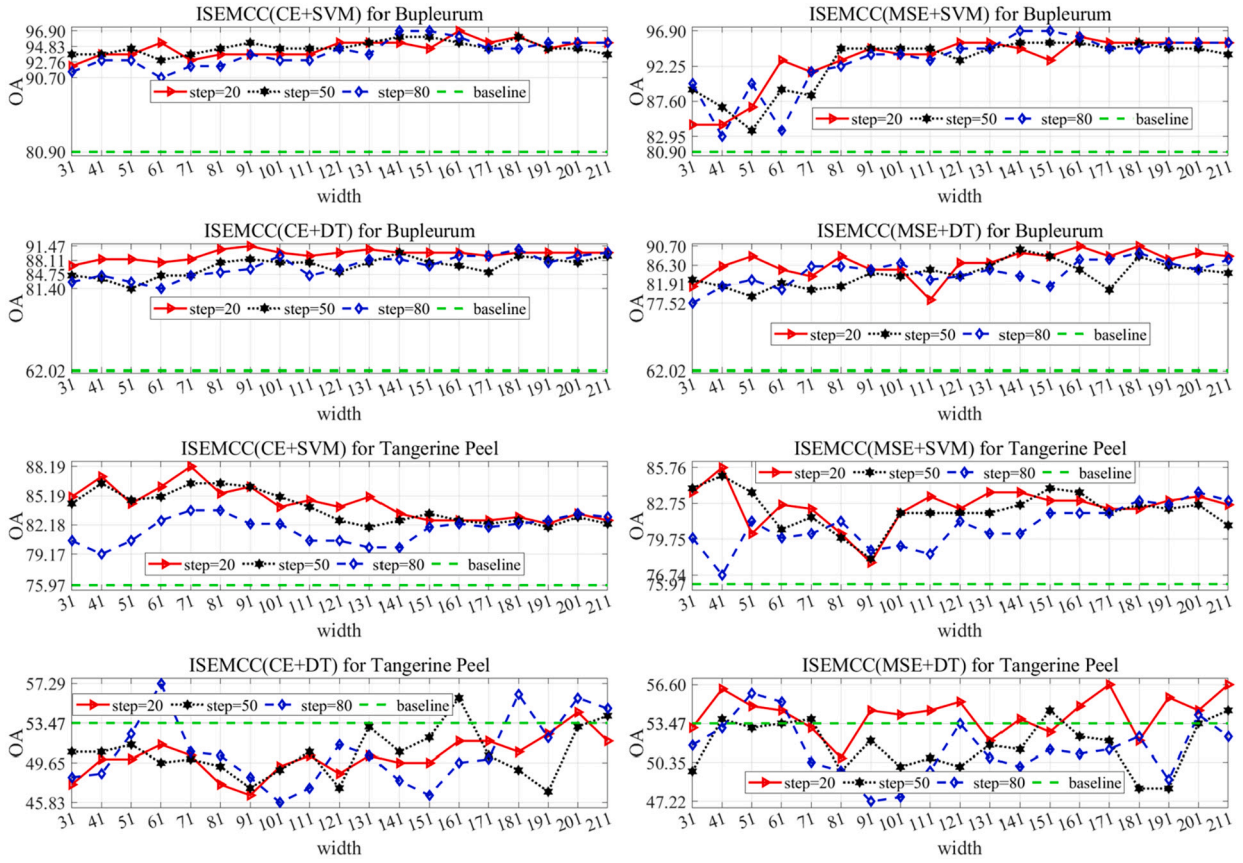

Fig. 4. Classification accuracies of ISEMCC with different window widths under three step sizes on two THz spectral data sets:the baseline in each subfigure shows the OA of the corresponding base classifer.

3.2.3. *Window width and sliding step testing*

The window width and sliding step are two important hyperparameters of the proposed algorithm. They are applied to partition the data into small interval segments, and control the input data dimension of each base classifier and the amount of sample information obtained by each base classifier. Furthermore, they also directly determine the number of base classifiers.

In order to test the sensitivity of the proposed algorithm to window width and sliding step, extensive testing on the data sets of Bupleurum and Tangerine peel has been conducted. Fig. 4 displays the changes in overall accuracy (OA) for ISEMCC(CE+SVM), ISEMCC(MSE+SVM), ISEMCC(CE+DT), and ISEMCC(MSE+DT) with varying window widths and three representative sliding steps. The baseline in each subfigure shows the OA of the corresponding base classifer.

Fig. 4 suggests that the OA fluctuation of ISEMCC under three sliding steps is relatively consistent, indicating that it is not sensitive to sliding steps. However, it is a little sensitive to the change of window width: In general, as the window width increases, the OA begins to increase and then tends to be stable or decrease. The reason may be that when the window width is small, each base classifier can obtain too little sample information, making it difficult to correctly classify samples. As the window width increases, the classification ability of the base classifier shows some improvement. As the window width increases further, the model exhibits saturation or supersaturation. Fig. 4 also shows that except ISEMCC(CE+DT) and ISEMCC(MSE+DT) for Tangerine Peel, the OAs of the proposed ISEMCC with different window width and sliding step are significantly higher than that of the corresponding base classifer. However, the results of ISEMCC(CE+DT) and ISEMCC(MSE+DT) for Tangerine Peel are a bit abnormal, the reasons of which still need to be further analyzed.

According to the experimental results shown in Fig. 4, the minimum OA, maximum OA, optimal window widths and sliding step sizes of the proposed algorithm are shown in Table 2.

3.2.4. *Comparative experiments*

In order to verify the classification performance of the proposed algorithm, six algorithms including SVM (linear kernel and RBF kernel), DT, RF, AdaBoost, RUSBoost [21] and ExtraTree[1] are used for comparison. The THz spectral data are still Bupleurum and

[1] https://github.com/rtaormina/MATLAB_ExtraTrees.

Table 2
Minimum OA, maximum OA, Optimal window width and sliding step settings of ISEMCC.

Data set	Algorithm	Optimal (width, step)	Minimum OA	Maximum OA
Bupleurum	CE+SVM	(141, 80), (151, 80), (161, 20)	90.70	96.90
	MSE+SVM	(141, 80), (151, 80)	82.95	96.90
	CE+DT	(91, 20)	81.40	91.47
	MSE+DT	(161, 20), (181, 20)	77.52	90.70
Tangerine peel	CE+SVM	(71, 20)	79.17	88.19
	MSE+SVM	(41, 20)	76.74	85.76
	CE+DT	(61, 80)	45.83	57.29
	MSE+DT	(171, 20), (211,20)	47.22	56.60

Table 3
Experimental comparison between ISEMCC and six typical classification algorithms.

Algorithm	Bupleurum	Tangerine peel
SVM(linear)	83.72%	82.64%
SVM(RBF)	84.50%	83.68%
DT	76.74%	56.94%
RF	91.47%	58.68%
AdaBoost	92.25%	65.63%
RUSBoost	79.07%	65.28%
ExtraTree	81.40%	63.19%
ISEMCC(MSE+DT)	90.70%	56.60%
ISEMCC(CE+DT)	91.47%	57.29%
ISEMCC(MSE+SVM)	**96.90%**	**85.76%**
ISEMCC(CE+SVM)	**96.90%**	**88.19%**

Tangerine peel. The parameters of all the six comparison algorithms have been optimized using Bayesian optimization with 100 objective function evaluations. Table 3 shows the experimental results.

As can be seen from Table 3:

- Apart from ISEMCC with DT on the Tangerine peel data set, the proposed algorithm outperforms its base classifiers in terms of classification accuracy.
- SVM-based ISEMCC is better than DT-based ISEMCC.
- CE-based ISEMCC is slightly better than MSE-based ISEMCC.
- Overall, the classification accuracy of the proposed algorithm is significantly higher than that of all comparison algorithms.

4. Conclusion

In the classification of THz spectral data, it is difficult to obtain ideal results by using conventional classification models directly, and there are still few classification methods that fit the characteristics of THz spectral data. According to the characteristics that the identification information of THz spectral data usually exists in one or more small characteristic peaks, the proposed algorithm transforms the search and selection problem of THz characteristic peaks into a sparse optimization problem by dividing the THz spectral data into intervals, which greatly improves the classification accuracy and makes the proposed algorithm explainable. Admittedly, the way in which the intervals are divided has an obvious impact on the performance of the proposed algorithm, and there seems to be no other efficient method of searching for interval-dividing parameters at present, except for cross-validation methods.

Whether the proposed method is also applicable to the classification of hyperspectral, near-infrared spectroscopy and other data remains to be further studied. The proposed algorithm can also be used for regression problems after slight modification, which will be further investigated in detail.

CRediT authorship contribution statement

Chengyong Zheng: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Formal analysis, Conceptualization. **Xiaowen Zha:** Writing – original draft, Validation, Software. **Shengjie Cai:** Visualization, Validation, Software, Data curation. **Jing Cui:** Writing – review & editing, Validation, Investigation. **Qian Li:** Supervision, Resources, Funding acquisition, Data curation. **Zhijing Ye:** Writing – original draft, Validation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that support the findings of this study are available from Chengyong Zheng (zcy@wyu.edu.cn), upon reasonable request.

Acknowledgement

This work was supported by the Wuyi University Hong Kong-Macau joint Research and Development Fund under Grant 2022WGALH16, supported in part by the National Natural Science Foundation of China under Grant 62001337, the Faculty Research Grants of the Macau University of Science and Technology under Grant FRG-22-101-FIE, and the Special Foundation in Key Fields for Universities of Guangdong Province under Grant 2023ZDZX4060.

References

[1] V. Fedotov, Phase control of terahertz waves moves on chip, Nat. Photonics 15 (10) (2021) 715–716, https://doi.org/10.1038/s41566-021-00887-8.
[2] Y. Peng, C. Shi, Y. Zhu, M. Gu, S. Zhuang, Terahertz spectroscopy in biomedical field: a review on signal-to-noise ratio improvement, PhotoniX 1 (1) (2020) 12, https://doi.org/10.1186/s43074-020-00011-z.
[3] M. Bauer, F. Friederich, Terahertz and millimeter wave sensing and applications, Sensors (Basel, Switzerland) 22 (24) (2022) 9693, https://doi.org/10.3390/s22249693.
[4] X. Fu, Y. Liu, Q. Chen, Y. Fu, T.J. Cui, Applications of terahertz spectroscopy in the detection and recognition of substances, Front. Phys. 10 (2022), https://doi.org/10.3389/fphy.2022.869537.
[5] S. Huang, H. Deng, X. Wei, J. Zhang, Progress in application of terahertz time-domain spectroscopy for pharmaceutical analyses, Front. Bioeng. Biotechnol. 11 (2023), https://doi.org/10.3389/fbioe.2023.1219042.
[6] J. Friska, M. Navaneetha Velammal, A. Rajeshwari, P. Hannah, Blessy, Random Forest (RF) based identification of rice powder mixture using terahertz spectroscopy, J. Phys. Conf. Ser. 1979 (1) (2021) 012056, https://doi.org/10.1088/1742-6596/1979/1/012056.
[7] S. Pan, H. Zhang, Z. Li, T. Chen, Classification of Ginseng with different growth ages based on terahertz spectroscopy and machine learning algorithm, Optik 236 (2021) 166322, https://doi.org/10.1016/j.ijleo.2021.166322.
[8] Y. Zhu, C. Shi, X. Wu, Y. Peng, Terahertz spectroscopy algorithms for biomedical detection, Acta Opt. Sin. 41 (1) (2020) 0130001, https://doi.org/10.3788/AOS202141.0130001.
[9] H. Liu, Z. Zhang, X. Zhang, Y. Yang, Z. Zhang, X. Liu, F. Wang, Y. Han, C. Zhang, Dimensionality reduction for identification of hepatic tumor samples based on terahertz time-domain spectroscopy, IEEE Trans. Terahertz Sci. Technol. 8 (3) (2018) 271–277, https://doi.org/10.1109/TTHZ.2018.2813085.
[10] J. Huang, J. Liu, K. Wang, Z. Yang, X. Liu, Classification and identification of molecules through factor analysis method based on terahertz spectroscopy, Spectrochim. Acta, Part A, Mol. Biomol. Spectrosc. 198 (2018) 198–203, https://doi.org/10.1016/j.saa.2018.03.017.
[11] P. Liu, X. Zhang, B. Pan, M. Wei, Z. Zhang, P.B. Harrington, Classification of sand grains by terahertz time-domain spectroscopy and chemometrics, Int. J. Environ. Res. 13 (1) (2019) 143–160, https://doi.org/10.1007/s41742-018-0159-y.
[12] H. Zhu, H. Wang, J. Liu, W. Wang, R. Gao, Y. Zhang, Application of terahertz dielectric constant spectroscopy for discrimination of oxidized coal and unoxidized coal by machine learning algorithms, Fuel 293 (2021) 120470, https://doi.org/10.1016/j.fuel.2021.120470.
[13] H. Zhang, Z. Li, T. Chen, J. Liu, Discrimination of traditional herbal medicines based on terahertz spectroscopy, Optik 138 (2017) 95–102, https://doi.org/10.1016/j.ijleo.2017.03.037.
[14] A. Sarjaš, B. Pongrac, D. Gleich, Automated inorganic pigment classification in plastic material using terahertz spectroscopy, Sensors 21 (14) (2021) 4709, https://doi.org/10.3390/s21144709.
[15] P. Huang, Y. Cao, J. Chen, W. Ge, D. Hou, G. Zhang, Analysis and inspection techniques for mouse liver injury based on terahertz spectroscopy, Opt. Express 27 (18) (2019) 26014, https://doi.org/10.1364/OE.27.026014.
[16] C. Zheng, S. Cai, Q. Li, C. Li, X. Li, A collaborative classification algorithm with multi-view terahertz spectra, Results Phys. 42 (2022) 106023, https://doi.org/10.1016/j.rinp.2022.106023.
[17] Y. Zhou, X. Wang, M. Zhang, J. Zhu, R. Zheng, Q. Wu, Mpce: a maximum probability based cross entropy loss function for neural network classification, IEEE Access 7 (2019) 146331–146341, https://doi.org/10.1109/ACCESS.2019.2946264.
[18] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. 3 (1) (2011) 1–122, https://doi.org/10.1561/2200000016.
[19] C.Y. Zheng, H. Li, Q. Wang, C. Philip Chen, Reweighted sparse regression for hyperspectral unmixing, IEEE Trans. Geosci. Remote Sens. 54 (1) (2016) 479–488, https://doi.org/10.1109/TGRS.2015.2459763.
[20] L. Li, M. Doroslovački, M.H. Loew, Approximating the gradient of cross-entropy loss function, IEEE Access 8 (2020) 111626–111635, https://doi.org/10.1109/ACCESS.2020.3001531.
[21] C. Seiffert, T.M. Khoshgoftaar, J. Van Hulse, A. Napolitano, Rusboost: a hybrid approach to alleviating class imbalance, IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum. 40 (1) (2010) 185–197, https://doi.org/10.1109/TSMCA.2009.2029559.