**Terahertz Spectrum Recognition of Pathogens Based on PCA-Siamese Neural Network**

ZENG Wan-dan, WU Cheng-wei, SHI Ru-jin  
Shanghai Institute of Technology  
SIT  
Shanghai,China  
E-mail: zengwd@sit.edu.cn, 344078971@qq.com  

LI Qian-xue, XIA Zhi-ping  
Academy of Military Medical Sciences  
Changchun, China  

Abstract: In the terahertz timedomain spectroscopy technique, 16 common pathogens were experimentally studied and their characteristic absorption spectra were obtained. The terahertz absorption spectra of 16 common pathogens were trained and identified by Siamese neural network method. First, the terahertz absorption spectra of the 16 pathogens were reduced by PCA to construct training data. Then, the constructed Siamese neural network model was trained by back propagation. Finally, the pathogens measured at different times were used as the test set to evaluate the model, after comparing with the training data, the matching absorption spectrum was obtained, and the recognition rate reached 97.34%. The recognition results fully indicate that the identification of different kinds of pathogens can be recognized by Siamese neural network, which provides an effective method of detection and identification of pathogens by terahertz spectroscopy.

Keywords—THz spectroscopy; machine learning; Siamese neural network; similarity learning

I. INTRODUCTION

The rapid development of terahertz technology in recent years has provided the basis for terahertz technology-related applications[1]. Terahertz has attracted a lot of attention in many fields due to its high frequency and high penetrability and low photon energy. Among them, biological and chemical substances can form a unique "fingerprint spectrum" in the terahertz band [2]. The quality of terahertz technology in the field of quality control, non-destructive testing, biomedicine, etc.[3,4] has broad application prospects. The terahertz spectrum analysis based on machine learning has also achieved good results[5,6,7]. Aiming at the classification of terahertz absorption spectra of pathogens, this paper proposes a classification method based on Siamese neural network. The method can integrate multiple distance features to achieve accurate classification of pathogens. Experiments show that the method has certain robustness and can assist decision-making in related works.

II. Data acquisition and preprocessing

The experiment uses a reflective terahertz generator (Fig 1). The laser used is Spectra Physics's self-mode-lockable tunable titanium sapphire laser with a laser center wavelength of 810 nm, a pulse width of 100 fs, and a repetition rate of 82 MHz. The laser output power is 980 mW. The femtosecond laser pulse generated by the laser is divided into two beams by a half-wave plate (HWP) and then by a beam splitter (BS): a beam passing through the BS is pumped light, passed through a chopper and a retarder (by reflection). After the mirrors M2 and M3 are formed, they are reflected and collimated and then converge on the emission crystal InAs<100> through the convex lens L1, thereby exciting the terahertz electromagnetic waves. The terahertz waves are collimated by the off-axis parabolic mirrors PM1 to PM4. On the electro-optical detection crystal; a beam reflected by the BS is used as the probe light, and after passing through a series of mirrors RM6 to RM11 and the convex lens L2, it is struck on the high-resistance silicon wafer through the polarizing plate P, and is reflected by the electro-optical detection crystal (ZnTe). On the electro-optic crystal, it meets the terahertz wave carrying the sample information, and then passes through the quarter-wave plate (QWP). The Wollaston prism (PBS) is divided into two beams of light whose polarization directions are perpendicular to each other. The differential detector demodulates the terahertz signal by measuring the difference between the two polarization components, and performs data acquisition by a computer to obtain sample information. We collected more than 50 of each bacterium in the terahertz spectrum of 16 pathogens bacteria (such as Enterobacter sakazakii, Acinetobacter baumannii, Salmonella enteritidis).

![Fig. 1 Schematic setup of the THz-TDS](image)

The first thing is smoothing. The collected spectral data have been partially de-noised, but there are still some noise and fluorescence background interference. Savitzky-Golay filter is used to smoothen the background interference of spectral data.

Second, to normalize the spectral data. The normalization process has two advantages:
1. Improve the convergence speed of the model
2. Improve the accuracy of the model, the effect is more obvious in the algorithm involving distance calculation.

The normalization method chosen is Min-Max Normalization:

---

978-1-7281-3380-5/19/$31.00 ©2019 IEEE
The terahertz offset of terahertz spectrum ranges from 0.1 to 2.2 THz. For the collected spectral data, it is characterized by discrete sampling of terahertz frequency. Because of the high sampling frequency and the high coupling degree between the feature points, there is redundancy in the feature points, PCA algorithm is adopted to reduce the feature dimension. Finally, three principal component features were retained to replace the original 140 features.

Feature reduction:

The terahertz offset of terahertz spectrum ranges from 0.1 to 2.2 THz. For the collected spectral data, it is characterized by discrete sampling of terahertz frequency. Because of the high sampling frequency and the high coupling degree between the feature points, there is redundancy in the feature points, PCA algorithm is adopted to reduce the feature dimension. Finally, three principal component features were retained to replace the original 140 features.

Where the $x_{\text{max}}$ is the maximum value of the feature, $x_{\text{min}}$ is the minimum value of the feature, $x^*$ is the normalized feature.

Feature reduction:

The terahertz offset of terahertz spectrum ranges from 0.1 to 2.2 THz. For the collected spectral data, it is characterized by discrete sampling of terahertz frequency. Because of the high sampling frequency and the high coupling degree between the feature points, there is redundancy in the feature points, PCA algorithm is adopted to reduce the feature dimension. Finally, three principal component features were retained to replace the original 140 features.

Finally, three principal component features were retained to replace the original 140 features.

The Siamese neural network consists of two main parts: Branch model: Used to extract spectral features, using the same torso network for both spectra in a spectral pair. Head model: Used to compare the eigenvectors of the output in the branch model to determine whether the spectra in the pair are matched.

The branch model uses the ordinary neural network model, in which the number of network layers is set to 4 layers, the number of neurons is 12, 8, 4, 2 respectively, and the activation function is ReLU. To prevent overfitting, L2 regularization was added with a regularization coefficient of 10-6. The network outputs a two-dimensional feature to represent the spectral properties of the pathogen.

A single-layer neural network used by the head model has an activation function of sigmoid, which can control the similarity in the interval $[0,1]$. Before the decision is made in the head model, the feature vector output from the branch model needs to be converted into a distance feature. The head model compares the similarities and differences between two feature vectors, and the distance metric is involved here. For each pair of features, calculate the sum, product, absolute difference, and squared difference.

The head model relies on these four distance characteristics to produce similarity. Through the learning of the head model, the network can trade off between the matched zero and non-zero values. A neural network layer with the same weight is used for each feature.

The training of the Siamese neural network is end-to-end, which can achieve better classification results than training the branch model and the head model separately. In training, the similarity of the network output is scored, which means that the output of the network is actually the dissimilarity between the bacteria. The higher the similarity between the two bacteria is equivalent to the lower the dissimilarity, the setting will be even more conducive to network optimization, it is easier to search for very similar but different samples, then the optimization problem is converted to linear assignment problem.

The similarity matrix is randomly initialized, and the dissimilarity of the same pathogen pair on the diagonal and the positive samples of the same kind of pathogen is set to infinity, in order to avoid the selection of these positive samples by the Hungarian algorithm when searching for negative samples. For each generation of similar matrices, the Hungarian algorithm is used to search for the most difficult to distinguish spectral pairs. After each round of training, the dissimilarity of the difficult matching spectral pairs selected in the similarity matrix is set to infinity, and the optimized attention is focused
on these spectral pairs. The Hungarian algorithm[10] selects the combination with the least degree of dissimilarity. The spectral pair in the combination is the negative sample of the target in the Siamese neural network. The significance of setting the dissimilarity to infinity is that it enlarges the distance between the negative sample with the positive ones in the feature space.

### IV. EXPERIMENTAL ANALYSIS

On this data set, the accuracy of Siamese neural network is slightly better than that of traditional machine learning algorithm. This is because the noise retained without smoothing is amplified between network layers, so smoothing needs to be done before normalization. The reason for selecting Min-Max normalization is that the mean and variance of samples are influenced by the nearest interpolation, and the maximum value does not change before and after the interpolation. Hungarian algorithm regularizes negative samples after each round of training. Reducing the error rate does not affect the prediction of positive samples.

#### Table 1 Result of spectroscopy classification

| Algorithm         | Accuracy |
|-------------------|----------|
| SVM               | 94.41%   |
| KNN               | 92.89%   |
| Logistic Regression | 89.60%  |
| Random Forest     | 91.47%   |
| Decision Tree     | 90.71%   |
| Siamese           | 96.34%   |

In terms of accuracy, Siamese network is slightly better than traditional machine learning algorithm; under the MAP5 standard, Siamese network has obvious advantages. Siamese network can predict the five most possible results according to the size of the predicted similarity.

#### Table 2 Ablation experiments

| Smoothing | Normalization | Hungarian | Accuracy |
|-----------|---------------|-----------|----------|
| ✓         | ✓             |           | 91.85%   |
| ✓         |               | ✓         | 94.20%   |
| ✓         |               |           | 94.71%   |
| ✓         | ✓             | ✓         | 96.34%   |

Smoothing treatment and normalization treatment increased by 2.47 and 1.24 percentage points, respectively, while smoothing treatment before normalization treatment increased by 4.08 percentage points, which was 0.37 more than that of single smoothing and normalization treatment. This is because the noise retained without smoothing is amplified between network layers, so smoothing needs to be done before normalization. The reason for selecting Min-Max normalization is that the mean and variance of samples are influenced by the nearest interpolation, and the maximum

**CONCLUSION**

According to the properties of terahertz spectrum, normalization of terahertz spectrum contributes greatly to the accuracy of model prediction. Hungarian algorithm is very important for training Siamese network to optimize negative samples.

Because of the poor interpretability of the neural network, the similarity of Siamese neural network data output has some limitations. The network optimization reduces the similarity of mismatched samples based on the distance between training data, and only has reference for the similarity below threshold. In order to further improve the accuracy of the model, more sample data can be added and a network design with stronger feature extraction ability can be designed to build a more complete spectral library.

**REFERENCES**

[1] Wu Shuai, Qu Hao, Tu Hao. Progress towards the application of terahertz technologies[J]. Application of Electronic Technique, 2019, 43(07): 3-7+18.
[2] Liu Sheng-gang, Zhong Ren-bin. Journal of Universit y of Electronic Science and Technology of China, 2009, 38 (5): 481.
[3] Tonouchi M. Nature Photonics, 2007, 1(2): 97.
[4] Bax ter J B, Guglietta G W. Analytical Chemistry, 2011, 83(12): 4342.
[5] Jia Yan, Chen Si-jia, Li Ning. Identification of Terahertz Absorption Spectra of Illicit Drugs Using Back Propagation Neural Networks[J]. Chinese Journal of Lasers, 2007(05): 719-722.
[6] Ma Shuai, Shen Tao, Wang Rui-qi. Terahertz Spectroscopic Identification with Deep Belief Network[J]. Spectroscopy and Spectral Analysis, 2015, 35(12): 3325-3329.
[7] Liu Junxu, Du Bin, Deng Yuqiang. Deng. Terahertz-Spectral Identification of Organic Compounds Based on Differential PCA-SVM Method Chinese Journal of Lasers[J]. Chinese Journal of Lasers, 2019, 46(06): 344-351.
[8] J. Bromley, I. Guyon, Y. Lecun, et al. Signature Verification Using a Siamese Time Delay Neural Network[C] Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993]. Morgan Kaufmann Publishers Inc. 1993.
[9] S. Chopra , R. Hadsell , Y. LeCun. Learning a Similarity Metric Discriminatively, with Application to Face Verification [C] IEEE Computer Society Conference on Computer Vision & Pattern Recognition, 2005.
[10] H W. Kuhn , The Hungarian method for the assignment problem[J]. Naval Research Logistics, 1955, 2(1-2): 83-97.