2010 Symposium on Security Detection and Information Processing

Terahertz Spectra Applications in Identification of Illicit Drugs Using Support Vector Machines

Rui Pan\textsuperscript{a}, Shusen Zhao\textsuperscript{a}, Jingling Shen\textsuperscript{a,}\textsuperscript{*}

\textsuperscript{a}Beijing Key Laboratory for Terahertz Spectroscopy and Imaging, Key Laboratory of Terahertz Optoelectronics, Ministry of Education, Department of Physics, Capital Normal University, Beijing, 100048 China

Abstract

Support vector machine (SVM) was employed to classify terahertz absorption spectra for the purpose of illicit drugs identification. We successfully identified seven pure illicit drugs and found that it is a convenient and efficient method for drug identification. Then, we tried to identify drug mixtures based on the same training data, and found that the main content in a mixture can be recognized. We also tried to determine main content’s proportion of mixtures, using spectra data of various proportions, which were made for training data, based on Beer’s law. The results confirmed support vector machine to be an effective method for illicit drugs identification and content analysis.

© 2010 Published by Elsevier Ltd.

Key words: spectroscopy; illicit drugs identification; terahertz absorption spectra; support vector machines (SVM)

1. Introduction

Since illicit drugs endanger both human healthy and public security, it is important for security departments to detect and identify illicit drugs. Among the existing inspection techniques [1, 2], some of them may destroy samples as chemistry analysis, some need detect in low temperature as normal infrared spectroscopic and some just can get the shape of drugs instead of the type, and it even may be harmful to inspectors as x-ray. Therefore, a safe and non-destructive method is needed for the detection and identification of illicit drugs. Terahertz (THz) wave, as an electromagnetic wave which locates between microwave and infrared bands, has been shown to be a candidate to detect drugs [3-6]. Because of the inherent rotation vibration states of many organic molecules in gas and the low frequency vibration of crystal lattice in solids [7,8], it is possible that THz spectroscopy can be used to detect illicit drugs, such as methamphetamine (MA), papaverine, and cocaine.

The identification of THz absorption spectra using self-organization feature map (SOM) artificial neural network has been demonstrated that SOM can identify pure drugs from their THz absorption spectra [9]. As it is an
unsupervised neural network, SOM neural network requires substantial time for training, usually several hours, which is very inconvenient in practical applications.

With the THz absorption spectra, we use support vector machines (SVM) to identify drugs. SVM algorithms are based on statistical learning method presented by Vapnik et al [10]. Due to its excellent learning capability, SVM has been applied in many fields successfully, and became a hot topic in these years. Comparing with neural networks, such as SOM artificial neural networks and BP (Back-Propagation) neural networks [11], SVM is applicable in many areas with less training samples, less training time and quicker identifying, while employing it to classify THz spectral data is a new attempt.

In this paper, we demonstrate the powerful ability of SVM in three ways: identifying pure drugs, impure drugs and determining drug mixture contents. Seven illicit drugs’ and flour’s spectra are used as SVM training data. The data of five illicit drugs and three mixtures, which were extracted at other times, were classified by SVM. SVM was also trained by different proportion of MA-flour mixtures, and then it is used to identify the content of mixtures.

2. THz absorption spectrum

![Figure 1: Schematic setup of a THz-TDS system](image)

Seven illicit drugs (ketamine, methylephedrine, cocaine, ephedrine, pseudoephedrine, papaverine, and methamphetamine) were legally provided by the First Research Institute of Ministry of Public Security of China. All the samples were powders with a purity of over 99%. They were dried before being pressed into 0.6-1.0mm thick and 10.0mm diameter slices with a pressure of 5 tons. They were put into our THz time domain spectroscopy system (THz-TDSS), shown in figure 1, to get their THz fingerprint spectra within 0.2-2.5 THz. BS is a beam splitter, HWP is a half-wave polarizer, QWP is a quarter-wave polarizer, M1-M11 are reflecting mirrors, PM1-PM4 are four parabolic mirrors, L1-L3 are focusing lenses, P is a polarizer and PBS is wollaston prism. The laser source is a mode-locked Ti: sapphire laser beam with 100fs pulse duration and an 80 MHz repetition rate at a central wavelength of 780 nm. The laser pulse was split into pump and probe beams. The pump beam was focused onto the photoconductive antenna to generate THz pulses. The generated THz pulses transmitted through four off-axis PMs and were focused on a ZnTe crystal. The probe beam propagated after a series of mirrors and was focused on the same spot as the THz pulse on the ZnTe crystal. The ZnTe crystal was used to detect THz pulses by applying an electro-optic sampling technique. The sample was placed in the focal point between PM2 and PM3. The parts within the dashed lines in figure 1 were placed in a chamber full of dry air in order to eliminate the influence of vapour. The relative humidity in the chamber was less than 4% and the temperature was 296 K in the experiments. Thus, the hydration of the samples is assumed to be low and constant. The data acquisition time is 3 minutes for each spectrum. The number of time samples is 301, which was accordingly truncated at 157th frequency point corresponding to frequency of 2.5 THz, and resolution is 50 GHz.
Figure 2: THz absorption spectra: (a1) the training data of papaverine (a2) the identifying data of papaverine (b1) the training data of ephedrine (b2) the identifying data of ephedrine (c1) the training data of ketamine (c2) the identifying data of ketamine

Figure 2 shows some of THz absorption spectra of the samples, the absorption spectra of papaverine, ephedrine and ketamine. Two spectra of each drug were measured at different days which were used for training and identifying, respectively. The relatively absorption coefficient we obtained is from equation (1):
\[ \alpha(\omega) = \frac{2}{L} \ln\left(\frac{4n_z(\omega)}{\rho(\omega)[n_z(\omega) + 1]^2}\right) \]  

where \( \rho(\omega) \) is the mode of complex transmission coefficient, \( n_z(\omega) \) is refractive index of sample, \( L \) is the thickness of sample.

3. Support Vector Machines [12-16]

Figure 3: Schematic optimal category plane: Solid points and hollow points are two types of samples, while \( H \) represents optimal category plane, and the distance from \( H_1 \) to \( H_2 \) is called margin. The optimal category plane will not only be able to classify two types of samples, but also get the largest margin.

SVM, which was originally designed for binary classification, can be extended for multi-class classification. The originally SVM transforms input space to a high-dimensional feature space by linear or nonlinear transformation which is defined by inner product function, and gets an optimal category plane, shown in figure 3. The category plane separates the different classes as a boundary. Different inner product kernel functions produce different SVM algorithms. There are four widely used kernel functions which are linear inner product function, polynomial inner product function, radial basis function and Sigmoid inner product function. The advantage of SVM is that even if the space dimensions increase, the complexity of the calculation does not change much. In the general supervision learning methods, there are two sets of data; one is used to make classifier, and the other is used to test classifier’s performance. We input pretreated experimental data to SVM, set parameters to train SVM, and then obtain a model which is used to identify test data set. In this paper, we used LIBSVM [17] with RBF kernel to classify samples.

4. Identification Results

4.1 Identification of pure drugs

The training data of SVM are THz absorption spectra of pure ketamine, methylephedrine, cocaine, ephedrine, pseudoephedrine, papaverine, MA and flour, ten groups of data for each, and no polyethylene within. The samples data of methylephedrine, ketamine, ephedrine, pseudoephedrine, papaverine, measured on other days, were used to be classified by SVM. The data (THz absorption spectra) of papaverine, ephedrine and ketamine, for example, for training and identifying are shown in figure 2. The classification results are satisfied. In fact, we have five different THz spectra of each sample to be identified and in each time the SVM gave a correct output.

Besides the strong ability of identifying pure illicit drugs, SVM greatly reduces training and identifying time. In practical use the training data are inevitably to be adjusted, such as increasing numbers or types. In this case the trained identification model must to be retrained. To finish the process it takes several hours to retrain SOM with the
same data, while just tens of seconds to retrain SVM. That is, we can finish the whole process including training and identifying by SVM within 1 minute. Therefore, it is possible in practical applications by SVM to adjusting training data at any time and come to identification results quickly and accurately.

4.2 Identification of drug mixtures

In practical use, the samples to be identified are usually drug mixtures. Therefore it is useful to get drug samples adulterated with other materials, flour for example, identified by a trained SVM, and the training data of SVM are the same ones for identifying pure samples. We chose ketamine mixed with polyethylene (50% to 50%), captured ketamine, MA mixed with flour (no polyethylene) and some pure drugs to be identified by SVM. The results are shown in Table 1.

Table 1 Result of drug mixtures identifying

<table>
<thead>
<tr>
<th>Samples</th>
<th>Cocaine</th>
<th>Ephedrine</th>
<th>Papaverine</th>
<th>Ketamine+Polyethylene</th>
<th>Captured Ketamine</th>
<th>MA70%+Flour30%</th>
<th>MA20%+Flour80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results</td>
<td>Cocaine</td>
<td>Ephedrine</td>
<td>Papaverine</td>
<td>Ketamine</td>
<td>Ketamine</td>
<td>MA</td>
<td>Flour</td>
</tr>
</tbody>
</table>

Table 1 tells that SVM classifies ketamine with polyethylene as ketamine, captured ketamine as ketamine, 70% MA and 30% flour as MA, 20% MA and 80% flour as flour. The result is reasonable. It was known from our former experiments that in the frequency range of 0.2 THz to 2.5 THz there are no absorption peaks in THz spectra of polyethylene, flour and most diluter in captured ketamine. That is, the spectra are nearly linear for these samples. So they have little effect on the appearance of mixture’s spectra when their portion is not very high, in other words, the spectral characteristics of the drug will not be obliterated. As an example the absorption spectra of pure MA and MA70%+flour30% mixture are shown in figure 4 (a) and (b). While for a mixture with smaller portion of drug the characteristic of its spectra will be obliterated in the linear spectra as shown in figure 4 (c) and (d), which are MA20%+flour80% and pure flour, respectively. In this case, SVM cannot identify drug from the mixture.

![Absorption spectra](a)

![Absorption spectra](b)

![Absorption spectra](c)

![Absorption spectra](d)
It has confirmed from the above identification results that the training model based on pure samples is also suitable for the mixtures, and that if the drug’s portion is over 70% in a mixture it still can be identified (Seventy percent is not the resolution of SVM.) Due to the limitation of pure drugs, we have no many samples to mix for getting other proportions. We can not determine what the minimum proportion of a real drug which can be identified by SVM is. Any way, the fact that SVM can identify some mixtures, especially it classifies the captured ketamine as ketamine, is already close to practical applications.

4.3 Determination of proportion

After successfully identified the pure and mixture samples, we explored the SVM’s ability in telling substance contents in a sample. In doing these, lots of known contents spectra data getting from experiments are needed for training process. However, the limitation of pure drugs makes it impossible to get so many samples with different proportions. Fortunately, based on Beer’s law which was also supported by our THz spectral experiments, we had made a series of different proportions THz spectra by calculating pure MA and flour spectral data, which were used in the training process. The MA-flour mixtures in proportions of 0% to 100% MA with step of 10% are shown in figure 5. The identification data of MA-flour mixture were obtained by THz-TDS experiment. Four sets of data for each training sample and three sets of data for each identified sample were applied in the process of SVM operation.

From the results we have seen that the mixtures with 20% MA were identified as MA content of 20%, and the mixtures with 70% MA were identified as MA content of 70%. It indicates that samples can be accurately classified by SVM that were trained by data with various proportions. It is important to point out that all the data we used are based on absolute absorption coefficient, so the absorption spectra are independent from sample’s thickness and mass. A mixture with certain contents has a unique spectrum.

In this way, drug content in samples can be identified. It provides us a fast way to determine a substance proportion in a mixture from its THz absorption spectrum. Regarding the advantage of the SVM identification, the operation speed is superior to SOM. Compare with SOM neural networks, which needs several hours to be trained, SVM, with better classification accuracy, only needs less than 1 minute with the same training data. This contents determination method is also non-destructive which is necessary in many applications especially in detection of illicit drug. It can be a new way for content analysis.

5. Conclusions

SVM was applied to identify THz absorption spectra of pure and impure illicit drugs. By THz absorption spectra of pure illicit drugs, the SVM accurately identified pure illicit drugs and successfully classified mixtures containing
illicit drugs with component over 70%. We also found that SVM can determine the content percentage of the main component in a mixture. Since the classification performance is excellent, SVM as an efficient classification system is feasible and worthy of consideration. Future studies will be about data processing techniques for absorption spectra, acquiring more data of spectra to train SVM, and developing the application in identifying medicine.

Acknowledgment

We thank Ting He and Wei Xiong for the fruitful discussions. This work was funded by the Science and Technology Program of Beijing Educational Committee under Grant No.KM200910028005. It is also supported by the National Keystone Basic Research Program (973 Program) under Grant Nos.2007CB310408 and 2006CB302901.

References