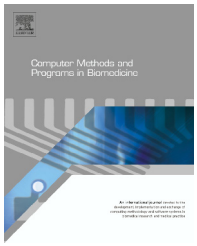
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 117 (2014) 387–403

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



Complex extreme learning machine applications in terahertz pulsed signals feature sets

X.-X. Yin[a,*], S. Hadjiloucas[b], Y. Zhang[a]

[a] Center for Applied Informatics and College of Enginering and Science, Victoria University, Footscray, VIC3011, Australia
[b] School of Systems Engineering, University of Reading, Reading RG6 6AY, UK

ARTICLE INFO

Article history:
Received 6 February 2014
Received in revised form
16 May 2014
Accepted 5 June 2014

Keywords:
THz
Complex extreme learning machine
Quaternary classification
Lagrangian
Multiclass classification

ABSTRACT

This paper presents a novel approach to the automatic classification of very large data sets composed of terahertz pulse transient signals, highlighting their potential use in biochemical, biomedical, pharmaceutical and security applications. Two different types of THz spectra are considered in the classification process. Firstly a binary classification study of poly-A and poly-C ribonucleic acid samples is performed. This is then contrasted with a difficult multi-class classification problem of spectra from six different powder samples that although have fairly indistinguishable features in the optical spectrum, they also possess a few discernable spectral features in the terahertz part of the spectrum. Classification is performed using a complex-valued extreme learning machine algorithm that takes into account features in both the amplitude as well as the phase of the recorded spectra. Classification speed and accuracy are contrasted with that achieved using a support vector machine classifier. The study systematically compares the classifier performance achieved after adopting different Gaussian kernels when separating amplitude and phase signatures. The two signatures are presented as feature vectors for both training and testing purposes. The study confirms the utility of complex-valued extreme learning machine algorithms for classification of the very large data sets generated with current terahertz imaging spectrometers. The classifier can take into consideration heterogeneous layers within an object as would be required within a tomographic setting and is sufficiently robust to detect patterns hidden inside noisy terahertz data sets. The proposed study opens up the opportunity for the establishment of complex-valued extreme learning machine algorithms as new chemometric tools that will assist the wider proliferation of terahertz sensing technology for chemical sensing, quality control, security screening and clinic diagnosis. Furthermore, the proposed algorithm should also be very useful in other applications requiring the classification of very large datasets.

* *Corresponding author*. + 61 0466 986 989.
E-mail address: xiaoxia.yin@vu.edu.au (X.-X. Yin).
http://dx.doi.org/10.1016/j.cmpb.2014.06.002
0169-2607/

1. Introduction

Terahertz (THz or T-ray) spectrometry and spectro-radiometry have become increasingly popular sensing modalities over the past two decades due in part to recent advances in continuous wave terahertz sources and detectors, but mostly due to the wide proliferation of THz time domain spectrometers (TDS). The later, utilize ultra-short laser pulse sources to perform time-resolved studies of molecular dynamics as well as explore spectroscopic imaging applications at millimetre and sub-millimetre frequencies (also known as the far-infrared part of the spectrum shown in Fig. 1). The terahertz part of the spectrum lying between the millimetre wave and infrared (100 GHz-10 THz) is particularly rich in terms of spectral features because at these frequencies we observe molecular rotation in gases, van der Waals bond or hydrogen-bond stretches and torsional bond deformations in liquids, as well as low frequency bond vibrations and phonon vibrations in crystals. Furthermore, this is a frequency range where current state-of-the-art electron-spin-resonance systems are operating [1,2], thus paving the way for better bio-molecule sensitivity on the basis of minute deviations in a sample's electron-spin resonance according to the physico-chemical state of the solvent. The higher frequencies associated with the THz spectrum correspond to the region where overtone and combination band spectroscopy can be performed; this is particularly interesting to environmental pollutants monitoring as well as in molecular astronomy. Infrared spectroscopy is unable to access lower frequency vibrational modes, making THz spectroscopy the only possible measurement modality in the above settings. A further attractive feature of THz spectroscopy as opposed to infrared spectroscopy is that samples have lower Rayleigh scattering at this part of the spectrum thus making non-invasive classification tests of samples while *in situ* (e.g. anthrax spores within an envelope) more reliable.

Terahertz transient spectrometers and imaging systems differ from their optical or infrared counterparts in that the signal-to-noise ratios in the acquired spectra is low due to a very inefficient process in the generation of the THz transients (lower by a factor of 10^5 compared to infrared time-domain spectroscopy systems centred at 800 nm). This introduces significant problems in the analysis and interpretation of spectra as well as the classification of samples. A further important difference of THz spectroscopy compared to its optical or infrared counterpart is that the longer wavelengths used enable the reliable recording of the phase delay across each frequency when the radiation is transmitted or reflected through the sample. In the case of time-domain spectrometry with a TDS system, one directly obtains reliable measurements of attenuation phase delay or dispersion at each spectral bin. Complementary information to traditional spectroscopic measurements may be therefore obtained. The current consensus in the bio-medical community is that advanced classification algorithms still need to be developed to assist screening, expert diagnosis, and subsequent treatment in an automated fashion.

It must be stressed that because of the significant cost associated with the installation and operation of THz transient spectrometers many of these systems are found mostly in national labs or most well-funded Physics, Chemistry or Biology research labs worldwide. The usual mistake by managers in these facilities is that optical experts are usually employed to run these systems. This decision may be partly justified on the basis that such systems are rather complex to operate requiring good understanding of optoelectronics as well as frequent alignment of the optical components in the system before performing an experiment. This practice, however, does not address the issue that a major bottle-neck resides in the analysis of the recorded spectra. It is not uncommon for users to have a relative lack of experience in the science of Chemometrics or the management of the associated very large data sets generated by these spectrometers. As a consequence THz-transient Chemometrics and sample classification is still at its infancy. The current work addresses this shortcoming by proposing novel classification modalities as chemometric tools specifically for these systems. The challenge for any automated THz pattern recognition systems is to explore the available spectral features in the input layer of the designed classifier as these were generated directly on the basis of the sample's THz response. Most molecules show rather complex THz absorption spectra with a multitude of absorption lines. In liquid or gaseous samples, those absorption lines are subject to thermal or pressure broadening at room temperature and within an imaging setting there may also be the result of several electromagnetic interactions [3] or pseudo-coherence errors [4] that would be associated with a thickness variation of the sample across its aperture when placed in the imaging system.

In previous works [5], we have shown that using signal processing techniques, it is possible to apodize [6] and denoise the corresponding time-domain signatures [7] or spectra or alternatively model them using a range of modelling techniques adopted from the systems identification literature [8,9]. We have further demonstrated that such analysis may be directly performed in the time, frequency or even the wavelet domains [10–13]. The current study is concerned with the classification of T-ray measurements on the basis of extracted features from their spectral signatures only. In this respect, the work follows directly to that performed by [14] where from the spectrum, a set of values were extracted as features, to be used as inputs to the classifier. In contrast, in the current work, we use directly the complex values associated with the Fourier transform of the time domain signatures, after taking into consideration separately the real and imaginary parts of the transform. Our goal is to explore the use of a complex value Extreme Learning Machine (ELM) [15,16] to classify the complex-valued THz datasets using complex valued labels.

The procedure is in many respects analogous to quaternary classification, where complex coupled hyper-planes are defined to accommodate the output of the classifier. The formulation uses a complex input space for the spectral signatures as well as optimisation variables that are all complex valued. In contrast to classic Support Vector machine (SVM) algorithms, complex-valued ELMs address the complex valued hyper-planes through the calculation of the smallest norm of output weights with the smallest training error [15]. In this respect, the operation of this algorithm is similar to its real-valued ELM counterpart [15]. The algorithm discards the normal threshold of b found in SVMs, without calculating