Raman spectroscopy for rapid and inexpensive diagnosis of echinococcosis using the adaptive iteratively reweighted penalized least squares-Kennard–stone-back propagation neural network

Xiangxiang Zheng¹,⁵, Guodong Lü²,⁵, Guoli Du², Xiaxia Yue³, Xiaoyi Lü¹,⁴, Jun Tang³ and Jiaqing Mo¹

¹ College of Information Science and Engineering, Xinjiang University, Urumqi 830046, People’s Republic of China
² The First Affiliated Hospital of Xinjiang Medical University, Urumqi 830000, People’s Republic of China
³ Physics and Chemistry Detecting Center, Xinjiang University, Urumqi 830046, People’s Republic of China

E-mail: xiaoz813@163.com

Received 11 April 2018, revised 1 May 2018
Accepted for publication 2 May 2018
Published 6 June 2018

Abstract
A rapid and inexpensive method of screening and diagnosis for echinococcosis is proposed for Raman spectroscopy, together with improved neural networks. We use the adaptive iteratively reweighted penalized least squares (airPLS) algorithm to deduct the fluorescence background from the Raman spectra of healthy people and echinococcosis patients. The processed data was compressed into the principal component by the PLS method, and the Kennard–stone (KS) algorithm was used to divide it into a training set and a testing set. Finally, the data was put into the back propagation (BP) neural network for modeling and prediction. The results show that the true positive rate of echinococcosis diagnosis is (94.2857 ± 4.0721)%; the true negative rate is (95.2381 ± 0)% and the overall accuracy rate is (94.6939 ± 2.3269)%.

The Raman spectroscopy combined with the airPLS-KS-BP algorithm can achieve fast and accurate diagnosis of echinococcosis.

¹ Author to whom any correspondence should be addressed.
⁵ These authors contributed equally to this work and should be considered co-first authors.

Original content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Keywords: Raman spectroscopy, back propagation algorithms, data processing, diseases, fluorescence

(Some figures may appear in colour only in the online journal)

1. Introduction

Echinococcosis is distributed widely in many developing countries, and Xinjiang of China is one of the high incidence areas [1]. Studies have shown that in the absence of treatment, about 95% of echinococcosis patients will die within a decade [2]. Early diagnosis and intervention of echinococcosis is an effective approach to reduce its morbidity and mortality. At present, conventional examination methods include clinical symptom diagnosis, medical imaging technology diagnosis and immunologic diagnosis [3–5]. However, they have some problems; such as expensive instruments, cumbersome processes, technical person operation requirements and so on [6]. Therefore, it is very important to develop a fast, accurate and low-cost detection method for diagnosis of echinococcosis.

Raman spectroscopy is suitable for the analysis of biological samples [7, 8], and it is gaining more applications in medical diagnosis and research [9–11]. At present, one of the main challenges for Raman spectroscopy in clinical applications is that the auto-fluorescence intensities of organism tissues, which are excited by the laser light source, are superimposed in the Raman bands, so that the Raman signal intensity is only about 10−3 times the original excitation intensity [12]; which means the original spectrum cannot reflect the essential information of the cells. Therefore, it is very necessary to reduce and deduct the fluorescence background.

Currently, the fluorescent background problem is solved mainly through the surface-enhanced Raman Scattering (SERS) or data processing technology. SERS technology has a high sensitivity and specificity, so it has shown good application prospects in the detection of clinical medicine [13]. But some SERS materials depend on the specific sample and are not universal [14]; the commonly used SERS active substrates in the medical field, gold and silver nano-particles, have instability and uneven enhancement effects [15], and the SERS chip is very expensive. In addition, the post-processing of the Raman spectrum is also an effective method for deducting fluorescence backgrounds. At present, polynomial fitting, wavelet transforms and derivatives are three major popular algorithms of Raman spectral fluorescence background reduction [16–18]. However, the accuracy of manual polynomial fitting depends on the user’s experience, and automatic polynomial fitting performs poorly in a low signal-to-noise ratio environment. The wavelet transform may lose some useful spectral information when reconstructing the waveform, causing some spectral distortion. The derivative algorithm will change the shape of the original peak. Due to the problems of the above three methods, Zhang et al proposed the algorithm of adaptive iteratively reweighted penalized least squares (air-PLS) [19], which has achieved good results in deducting the fluorescence background, and has already been widely applied in various research fields [20, 21].

The selection of disease prediction and analysis methods will also affect the efficiency and accuracy of the prediction results. At present, the main prediction methods are regression analysis and artificial neural networks (ANNs), where the back propagation neural network (BPNN) has the characteristics of high-precision, non-linearity, self-organization, self-learning and self-adaptation, and it has a good application in attributing the recognition of cells. In our previous study, Raman spectroscopy combined with the PCA-BPNN model was used to diagnose echinococcosis patients and it achieved good results [22]. The over-fitting of the BPNN was successfully solved by using the principal component analysis (PCA) method. However, the PCA method does not consider output variables when selecting variables. Focusing on these issues, Jia et al [23] used the PLS algorithm to compress the principal component. Compared with the PCA algorithm, the relationship between the independent variables extracted by the PLS algorithm and the dependent variable is larger, and when dealing with multi-collinearity, high redundancy and multi-noise data, the PLS algorithm has higher accuracy and stability.

This article uses the airPLS algorithm to deduct the fluorescence background, and uses the PLS algorithm to compress the principal component. On the basis of this, for the training set, which requires representative data [24], we use the KS-BP algorithm to select the training set and classify the Raman spectral data of echinococcosis patients.

2. Materials and methods

2.1. Instrument

The Raman spectra experiments were performed by means of a laser Raman spectrometer (LabRAM HR Evolution RAMAN SPECTROMETER, HORIBA Scientific Ltd.) with a 10× objective at an ambient temperature. To guarantee a good signal-to-noise ratio and present fluorescence emission, the samples were excited by the 532 nm green light from a Spectra Physics Ar ion laser.

2.2. Serum sampling

The blood serum samples from the echinococcosis key laboratory of Xinjiang Medical University First Affiliated Hospital were randomly selected, with clear diagnosis and complete data from 55 cases of healthy people and 68 cases of echinococcosis patients.

2.3. Pre-treatment of data

The noises of Raman spectra are mainly in two categories: one is the thermal noise of the instrument, the other is the environmental interference noise. The background interference is...
mainly the baseline drift caused by the fluorescence background of the Raman spectrum [25]. In this study, the self-normalization algorithm is used to eliminate the noise of the data after removing the fluorescence background. And in order to reduce the influence of harmful factors—such as electronic noise, light scattering and instability of the output laser [26]—each sample was scanned twice and the the average spectrum was taken as its final spectrum.

2.4. Model evaluation

The true positive rate, the true negative rate and the overall accuracy rate were used as the evaluation indexes of the ANN model. The three indicators are defined as follows:

\[
\text{true positive rate} = \frac{A}{A + C} \times 100\% \quad (1)
\]

\[
\text{true negative rate} = \frac{D}{B + D} \times 100\% \quad (2)
\]

\[
\text{overall accuracy rate} = \frac{A + D}{A + B + C + D} \times 100\% \quad (3)
\]

where A, B, C and D represent true-positive, false-positive, false-negative and true-negative samples, respectively; ‘positive’ on behalf of someone suffering from echinococcosis while ‘negative’ on behalf of a normal case.

3. Result and analyses

3.1. Raman spectroscopic analysis

Figure 1 shows representative Raman spectra from healthy people and echinococcosis patients. Figure 2 shows Raman spectra of different samples with echinococcosis. It can be seen from figure 1 that the shapes of two spectral curves are similar, but the characteristic peaks are different, which is the basis of qualitative judgment. It can be seen from figure 2 that the Raman spectra of different patients suffering from echinococcosis are affected by the background of fluorescence, resulting in a baseline drift in the Raman spectrum and the change in range of the fluorescence background becomes very large, which has brought some adverse effects on qualitative judgment. Therefore, it is necessary to deduct the background of the spectrum.

3.2. Fluorescence background deduction

The airPLS algorithm is a method proposed in recent years to correct the spectral baseline drift, which can effectively deduct the fluorescence background in the Raman spectrum, and its baseline estimation is fast and flexible [27]. Therefore, this paper adopts the airPLS algorithm to eliminate the Raman fluorescence background.

Figure 3 shows the original and calibration spectra of a representative sample. It can be seen that the algorithm can effectively deduct the Raman fluorescence background, while maintaining the original spectral peak type.

3.3. Data compression

If the full spectrum data as variables are put directly into the neural network for modeling, the computational amount

Figure 1. Representative Raman spectra from healthy people and echinococcosis patients.
Figure 2. Raman spectra of different samples with echinococcosis.
Figure 3. Raman spectra of different samples with echinococcosis.
will be too large, and not all data will be useful for modeling. Therefore, the spectral data which are weaker and not related to the sample should be eliminated. In this paper, the PLS algorithm is used to compress the principal component, and the optimal principal components are determined by the 10-fold cross validation method.

The spectral samples with and without background deduction are processed by the PLS algorithm. The first, second and third principal components of the score matrix are extracted for drawing and analysis. As can be clearly seen from figure 4(a), without background deduction, healthy people and echinococcosis patients cannot be separated from the samples. After background deduction, the samples of healthy people and echinococcosis patients can be well separated, and the aggregation of each sample has been significantly improved.

3.4. Sample division

According to the 10-fold cross validation method, the number of PLS principal components is determined to be three, so the first three principal components are taken as the new variables to construct the sample matrix, and the BPNN is used for classification. In order to enable the BPNN to correctly predict new samples, the changes of the samples to be analyzed must be included in the training set as much as possible, so valid algorithms must be used to select some representative samples which can cover the whole change range in bio-information components as much as possible. In this study, the KS algorithm is adopted to select 74 samples (60%) as the training set, and 49 samples (40%) as the testing set from 123 samples.

3.5. Modeling and results comparison

The input layer and output layer structures of the BPNN are related to the practical application, and key to the design is the structure of the hidden layer. In this paper, the three-layer BPNN structure is adopted, the input layer is the first three principal components, the number of output layer nodes is one, and the number of hidden nodes is calculated according to

\[ m = \sqrt{n + 1 + a} \]  (4)

where \( m \) is the number of hidden layer nodes, \( n \) is the number of output layer nodes, \( l \) is the number of input layer nodes and \( a \) is a constant between 1 and 10.

Combined with many rounds of experiments, the optimal number of the hidden layer nodes is taken as three, the transfer function of the hidden layer uses the Sigmoid function, the output layer uses the Purelin function, the training function uses the Trainlm function, the maximal iteration times are 2000, the display error once for every 500 times, the target error is 0.00000001, the learning rate is 0.05 and the remaining training parameters are the default values. The values of the output neurons for healthy persons and echinococcosis patients are set as one and two, respectively. When the output values are less than or equal to 1.5, we judge it as a healthy person, otherwise as an echinococcosis patient.

The initial weights of the BPNN were randomly initialized by the network [28], so in order to eliminate the random

| Table 1. Model diagnostic results. | Table 2. Simulation results of model. |
|-----------------------------------|--------------------------------------|
| Processing method | True positive rate | True negative rate | Overall prediction accuracy rate |
| PLS-KS-BPNN | 76.9231% | 86.9565% | 81.6327% |
| airPLS-PCA-KS-BPNN | 72.7273% | 74.0741% | 73.4694% |
| airPLS-PLS-KS-BPNN | 82.1429% | 80.9524% | 81.6327% |
| airPLS-PLS-KS-BPNN | 92.8571% | 95.2381% | 93.8776% |

Figure 4. (a) Result of raw spectra. (b) Result after background deduction.
interference, the model randomly selected five consecutive diagnostic results, which are shown in table 1. Among them, the highest diagnosis rate is as follows: in 49 testing sets of data, there are 28 cases of echinococcosis patients, all of which are confirmed, none are misdiagnosed and the true positive rate is 100%; there are 21 cases of healthy persons, of which 20 cases are confirmed, one case is misdiagnosed, the true negative rate is 95.2381% and the overall accuracy rate is 97.95992%. The lowest diagnosis rate is as follows: in 49 testing sets of data, there are 28 cases of echinococcosis patients, of which 25 cases are confirmed, three cases are misdiagnosed and the true positive rate is 89.2857%; there are 21 cases of healthy persons, of which 20 cases are confirmed, one case is misdiagnosed, the true negative rate is 95.2381% and the overall accuracy rate is 91.8367%.

Based on the random characteristic of the initial weights of the BPNN, the calculation of the echinococcosis diagnosis rate can consider five times of running results. Among 49 testing sets of data, the true positive rate is 94.2857% ± 4.0721%, the true negative rate is 95.2381% ± 0% and the overall accuracy rate is 94.6939% ± 2.3269%.

The random running results of the PLS-KS-BPNN model, the airPLS-PCA-KS-BPNN model and the airPLS-PLS-BPNN model for the original spectra are shown in table 2.

It can be seen that the true positive rate, the true negative rate and the overall accuracy rate after using the airPLS-PLS-KS-BPNN model are all enhanced, where using the airPLS algorithm to deduct the background can effectively eliminate the spectral drift and noise, using the PLS algorithm to compress the data can give the model higher precision and stability and using the KS algorithm to divide the training set can improve the representativity of the sample, which demonstrates that the airPLS-PLS-KS-BPNN model can achieve a rapid and accurate prediction of echinococcosis very well.

4. Conclusion

The diagnosis and prediction of healthy people and echinococcosis patients are performed by Raman spectroscopy combined with the PLS-KS-BPNN model. The experimental results show that the airPLS-PLS-KS-BPNN model can better diagnose echinococcosis. Compared with the traditional diagnostic methods, it has the advantages of low cost, simple operation and fast analysis; so, it is suitable for a rapid and accurate diagnosis of echinococcosis. In the next step, we will further develop more clinical trials and analyze their experimental data to improve the accuracy and reliability of diagnosis.

Funding

National Natural Science Foundation of China (NSFC) (61765014); Reserve Talents Project of National High-level Personnel of Special Support Program (QN2016XX0324); Urumqi Science and Technology Project (No. P161310002 and Y161010025); Reserve Talents Project of National High-level Personnel of Special Support Program (Xinjiang[2014]22).

References

[1] Han S et al 2016 Investigation of the use of serology and ultrasonography to detect hepatic cystic echinococcosis in Heilongjiang, China, using a Bayesian framework Acta Tropica 162 212
[2] Wu C et al 2017 Genetic variation of mitochondrial genes among Echinococcus multilocularis isolates collected in western China Parasites Vectors 10 265
[3] Laca L 2017 Importance of complex diagnostic approach in differential diagnosis of alveolar echinococcosis: helmint-hologia Helminthologia 52 298–302
[4] Unal E et al 2017 A rare retrovesical hydatid cyst and value of transrectal ultrasonography in diagnosis: a case report and review of the literature Med. Ultrasonogr. 19 111–3
[5] Tina P et al 2016 Hepatic echinococcal cysts: a review J. Clin. Transl. Hepatol. 4 39–46
[6] Tamarozzi F et al 2016 Comparison of the diagnostic accuracy of three rapid tests for the serodiagnosis of hepatic cystic echinococcosis in humans PLoS Negl. Trop. Dis. 10 e0004444
[7] Yeturu S et al 2016 Handheld Raman spectroscopy for the early detection of plant diseases: abutilon mosaic virus infecting Abutilon sp Anal. Methods 8 3450–7
[8] Conrad A O et al 2016 Application of infrared and Raman spectroscopy for the identification of disease resistant trees Front. Plant Sci. 6 1152
[9] Richardson W et al 2015 Ensemble multivariate analysis to improve identification of articular cartilage disease in noisy Raman spectra J. Biophotonics 8 555–66
[10] Stone N et al 2008 Advanced transmission Raman spectroscopy: a promising tool for breast disease diagnosis Cancer Res. 68 4424–30
[11] Matousek P et al 2016 Development of deep subsurface Raman spectroscopy for medical diagnosis and disease monitoring Chem. Soc. Rev. 45 1794–802
[12] Jarvis R M et al 2004 Surface-enhanced Raman spectroscopy for bacterial discrimination utilizing a scanning electron microscope with a Raman spectroscopy interface Anal. Chem. 76 5198–202
[13] Yu L et al 2017 Label and label-free based surface-enhanced Raman scattering for pathogen bacteria detection: a review Biosens. Bioelectron. 94 131–40
[14] Grady A O et al 2001 Quantitative Raman spectroscopy of highly fluorescent samples using pseudosecond derivatives and multivariate analysis Anal. Chem. 73 2058–65
[15] Yan X et al 2017 The optimal hotspots of dynamic surface-enhanced Raman spectroscopy (D-SERS) for drugs quantitative detection Anal. Chem. 89 4875–81
[16] Jirasek A et al 2004 Accuracy and precision of manual baseline determination Appl. Spectrosc. 58 1488–99
[17] Hu Y et al 2007 A background elimination method based on wavelet transform for Raman spectra Chemometr. Intell. Lab. Syst. 85 94–101
[18] Gan F et al 2006 Baseline correction by improved iterative polynomial fitting with automatic threshold Chemometr. Intell. Lab. Syst. 82 59–65
[19] Zhang Z M et al 2010 An intelligent background—correction algorithm for highly fluorescent samples in Raman spectroscopy J. Raman Spectrosc. 41 659–69
[20] Huang Y Y et al 2010 Micro-raman spectroscopy of algae: composition analysis and fluorescence background behavior Biotechnol. Bioeng. 105 889–98
[21] Mani-Varnosfaderani A et al 2016 Estimating complicated baselines in analytical signals using the iterative training of Bayesian regularized artificial neural networks Anal. Chim. Acta 940 56–64
[22] Cheng J Y et al 2017 Study on the echinococcosis blood serum detection based on Raman spectroscopy combined with neural network Optoelectron. Lett. 13 77–80
[23] Jia W et al 2015 An optimized classification algorithm by BP neural network based on PLS and HCA Appl. Intell. 43 176–91
[24] Chen W R et al 2016 Representative subset selection and outlier detection via isolation forest Anal. Methods 8 7225–31
[25] Liland K H et al 2016 Model-based pre-processing in Raman spectroscopy of biological samples J. Raman Spectrosc. 47 643–50
[26] Guan D et al 2015 Noninvasive liver diseases detection based on serum surface enhanced Raman spectroscopy and statistical analysis Opt. Express 23 18361–72
[27] Zhang Z M et al 2010 Baseline correction using adaptive iteratively reweighted penalized least squares Analyst 135 1138–46
[28] Mattfeldt T et al 1999 Prediction of prostatic cancer progression after radical prostatectomy using artificial neural networks: a feasibility study BJU Int. 84 316–23