Near-infrared spectroscopic applications for diagnosis of endometrial carcinoma

Yuhong Xiang
Ke Xu
Zhuoyong Zhang
Capital Normal University
Department of Chemistry
Beijing, China 100048

Yinmei Dai
Capital University of Medical Science
Beijing Obstetrics and Gynecology Hospital
Beijing, China 100006

Peter de B. Harrington
Ohio University
Department of Chemistry and Biochemistry
Center for Intelligent Chemical Instrumentation
Athens, Ohio 45701-2979

Abstract. NIR spectra of 77 endometrium sections (malignant, hyperplasia, and normal samples) are collected. Partial least squares discriminant analysis (PLS-DA) and fuzzy rule-building expert systems (FuRES) are used for classification based on the NIR spectral data. The classification ability of two classifiers is evaluated by using ten bootstraps and five Latin partitions. The results indicate that the classification ability of FuRES is better than that of PLS-DA. The sensitivity, specificity, and accuracy obtained from FuRES for malignant endometrium diagnosis are 90.0 ± 0.7, 95.0 ± 0.8, and 93.1 ± 0.8%, respectively. The results demonstrate that NIR spectroscopy combined with the FuRES technique is promising for the classification of endometrial specimens and for practical diagnostic applications. © 2010 Society of Photo-Optical Instrumentation Engineers. [DOI: 10.1117/1.3512183]

Keywords: endometrial carcinoma; near-infrared spectroscopy; fuzzy rule-building expert system; bootstrapped Latin partitions.

1 Introduction

Endometrial carcinoma is the third major cause of gynecologic cancer death (behind ovarian and cervical cancer). Currently, it is the most common malignancy of the female genital tract in industrialized countries and is increasingly prevalent throughout the world year by year.1

Traditional diagnostic tests for endometrial carcinoma involve fractional curettage, which has some disadvantages and limitations, including: 1. blind curettage can cause the identification of small lesions to be missed; 2. biased evaluations of tumor location, grade, and penetration depth in the myometrium can lead to high false-positive rates;2 and 3. it is sometimes difficult to distinguish differentiated tissues between endometrial carcinoma and highly atypical hyperplasia in pathology. Consequently, there is an urgent demand for developing an accurate, rapid, and inexpensive approach to diagnose endometrial carcinoma.

Near-infrared spectroscopy (NIRS) has been successfully applied in noninvasive pathology studies. The absorption of near-infrared radiation (14,700 to 4000 cm−1) provides qualitative and quantitative information about the chemical composition of physical tissues. Any alteration in the composition of the tissues can be detected and used for diagnostic purposes. Most of the NIRS investigations dealing with human tissues were on breast.3,4 Other NIRS studies including cervix,5 brain,6 skin,7 prostate,8 lung,9 head and neck,10 pancreas,11 and colorectal tissues have also been reported.11 The goals of these studies were to disclose absorption peaks of one or more chromophores, characterizing the differences between malignant and normal tissues. Chemometrics provides a powerful tool for researchers to extract qualitative and quantitative information effectively from overlapped NIR peaks. Many pattern recognition techniques are widely used for classification such as partial least squares discriminant analysis (PLS-DA),12 cluster analysis (CA),13 artificial neural network (ANN),14 SIMCA,15 etc.

The aim of our work is to develop a new chemometric method to diagnose endometrial carcinoma based on the NIR spectra of endometrial tissues. The NIR spectra of endometrial specimens were collected and preprocessed. PLS-DA and fuzzy rule-building expert system (FuRES) models were used for discriminating various classes of samples. The better results of the FuRES classifier were obtained with sensitivity of 90.0 ± 0.7%, specificity of 95.0 ± 0.8%, and prediction accuracy was 93.1 ± 0.8% for detecting patients with endometrial carcinoma.

2 Theory

FuRES is a multivariate classifier based on information theory for modeling data in the form of an easy-to-interpret classification tree.16 Using the iterative dichotomiser 3 (ID3) algorithm, FuRES provides local modeling and implements conjugate gradient optimization for the global minima of fuzzy classification entropy. By this algorithm, a decision tree was constructed, taking the whole set of samples as a root node. Then it partitions the samples at each branch until all the samples belong to one class at each leaf (i.e., terminal node). A classification tree is constructed with fuzzy logic applied at each. Fuzzy classifiers are robust, so that outlying data objects would not cause the model to degrade as in other least squares classifiers.16

PLS is a multivariate statistical technique for modeling a relationship between dependent variables (Y) and independent variables (X). PLS-DA aims to find the variables and directions in multivariate space that discriminate the known classes in the
calibration set. In our work, a set of binary codes, [1 0 0], [0 1 0], and [0 0 1], were used to represent three types of pathological sections of endometrium tissue.

The bootstrapped Latin partitions (BLPs) method can provide an unbiased evaluation of pattern recognition models.\(^{17}\) The method was developed based on traditional cross-validation and resampling methods. With the Latin-partition method, block cross-validation is implemented. The dataset is randomly partitioned into \(n_{\text{part}}\) equally sized blocks. One group is left out for validation and the others are used for model building. Each group is used once for prediction and \((n_{\text{part}} - 1)\) times for training. The objects are each used once and only once for prediction, so that the results for the dataset can be pooled. Bootstrapping is a resampling method that uses sampling with replacement.\(^{18}\)

The prediction error is averaged among the bootstraps, and the standard deviation of the errors allows confidence intervals to be calculated. Therefore, the use of bootstrapped Latin-partition validation can help characterize the inherent variations in the data that result from different partitions between the training and prediction datasets.\(^{19}\)

### 3 Materials and Method

A Nicolet 6700 (Thermo Scientific, Waltham, Massachusetts) extended Fourier transform near-infrared (FT-NIR) spectrometer (Thermo Electron) equipped with an InGaAs detector was used for the NIR measurement. The spectrometer was controlled by OMNIC service software, version 7.3. Data analysis was done using MATLAB software (The MathWorks Incorporated, South Natick, Massachusetts).

A total of 77 pathological sections of endometrium tissue were provided by Beijing Obstetrics and Gynecology Hospital affiliated with Capital Medical University, Beijing. The histopathological diagnoses indicated that the 77 cases were divided into three groups, including 29 cases of endometrial carcinoma with the age of patients between 35 and 71, 30 cases of endometrial hyperplasia within ages 29 to 63, and 18 cases of normal endometrium tissue within ages 19 to 35. Tissue specimens were prepared as routine paraffin sections.

The NIR spectra of pathological sections were collected by an optical integrating sphere system at room temperature with diffuse reflectance mode. All sections were scanned from 4000 cm\(^{-1}\) to 10,000 cm\(^{-1}\), and the spectrum of each sample was obtained as a mean of 64 scans, with a resolution of 4 cm\(^{-1}\). Each section was measured at five different locations, and the average spectra of tissue sections were used as the spectra of cases in the following analysis.

### 4 Results and Discussions

The spectra were first smoothed using the Savitzky-Golay algorithm with a five-point cubic polynomial to eliminate high-frequency noise and baseline drift, followed by multiplicative scatter correction (MSC) to correct the baseline effects caused by physical conditions, such as optical path length, thickness, and medium of sections. The spectra were normalized to unit length and the average of all the normalized spectra was subtracted from each spectrum.

Principal component analysis (PCA) was applied to visualize differences among the samples. The principal component scores show that the three classes of samples are not resolved.

The average values of calibration and prediction errors with respect to latent variable number are given in Fig. 1, with 95% confidence intervals obtained from ten bootstraps and five Latin partitions. The latent variable number of 3 was chosen based on this evaluation.

A confusion matrix obtained from the average prediction results of ten bootstraps is given in Table 1. In this matrix, the rows represent the true class, and columns represent the predicted class. The average values of the ten bootstraps are presented with 95% confidence intervals. In this test, the sensitivity was 77.9 ± 2.1%, the specificity was 65.6 ± 2.1%, and prediction accuracy was 70.3 ± 2.1% for detecting malignant samples.

FuRES constructs a classification tree that allows the visualization of the inductive structure of the rules. Membership functions in FuRES allow values between 0 and 1 and provide a measure of the degree of similarity of elements in the total population.\(^{20}\)

#### Table 1 The confusion matrices obtained from the validations of PLS-DA and FuRES classifiers.

<table>
<thead>
<tr>
<th></th>
<th>Malignant</th>
<th>Hyperplasia</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PLS-DA model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malignant</td>
<td>22.6 ± 0.6</td>
<td>5.4 ± 0.6</td>
<td>1.0 ± 0</td>
</tr>
<tr>
<td>Hyperplasia</td>
<td>15.3 ± 0.9</td>
<td>13.5 ± 1.0</td>
<td>1.2 ± 0.2</td>
</tr>
<tr>
<td>Normal</td>
<td>0.0 ± 0</td>
<td>0.0 ± 0</td>
<td>18.0 ± 0</td>
</tr>
<tr>
<td><strong>FuRES model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malignant</td>
<td>26.1 ± 0.2</td>
<td>1.9 ± 0.2</td>
<td>1.0 ± 0</td>
</tr>
<tr>
<td>Hyperplasia</td>
<td>0.5 ± 0.3</td>
<td>28.0 ± 0.5</td>
<td>1.5 ± 0.5</td>
</tr>
<tr>
<td>Normal</td>
<td>0.4 ± 0.3</td>
<td>0.0 ± 0</td>
<td>17.6 ± 0.3</td>
</tr>
</tbody>
</table>
provide a powerful tool for the detection of early endometrial carcinoma. This method could be developed into a noninvasive method for clinical diagnosis for other cancers.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant numbers 20875065 and 30772322), Beijing Municipal Natural Science Foundation (2102010), and Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (PHR20100718).

References