Contents lists available at SciVerse ScienceDirect

Cancer Epidemiology

The International Journal of Cancer Epidemiology, Detection, and Prevention

journal homepage: www.cancerepidemiology.net

Near infrared spectroscopy combined with least squares support vector machines and fuzzy rule-building expert system applied to diagnosis of endometrial carcinoma

Fan Yang [a], Jing Tian [a], Yuhong Xiang [a], Zhuoyong Zhang [a,*], Peter de B. Harrington [b]

[a] Department of Chemistry, Capital Normal University, Beijing 100048, China
[b] Center for Intelligent Chemical Instrumentation, Department of Chemistry and Biochemistry, Clippinger Laboratories, Ohio University, Athens, OH 45701-2979, USA

ARTICLE INFO

Article history:
Received 29 July 2011
Received in revised form 18 October 2011
Accepted 20 October 2011
Available online 17 November 2011

Keywords:
Endometrial carcinoma
Near infrared spectroscopy
Least squares support vector machines
Fuzzy rule-building expert system
Cancer diagnosis

ABSTRACT

Objective: The feasibility of early diagnosis of endometrial carcinoma was studied by least squares support vector machines (LS-SVM) and fuzzy rule-building expert system (FuRES) that classified near infrared (NIR) spectra of tissues. *Methods:* NIR spectra of 77 specimens of endometrium were collected. The spectra were pretreated by principal component orthogonal signal correction (PC-OSC) and direct orthogonal signal correction (DOSC) methods to improve the signal-to-noise ratio (SNR) and remove the influences of background and baseline. The effects of modeling parameters were investigated using bootstrapped Latin-partition methods. *Results:* The optimal LS-SVM model of the PC-OSC pretreatment method successfully classified the samples with prediction accuracies of 96.8 ± 1.4%. *Conclusions:* The proposed procedure proved to be rapid and convenient, which is suitable to be developed as a non-invasive diagnosis method for cancer tissue.

1. Introduction

Endometrial carcinoma is one of the most common gynecologic cancers. About 46,470 cases of endometrial cancers were estimated to occur and 8120 deaths in the United States in 2011 [1]. Endometrial adenocarcinoma occurs during the reproductive and menopausal years. The median age of women with this malignancy is early in the seventh decade of life, although most patients are aged 50–59 years. Approximately 5% of women younger than 40 years have adenocarcinoma, and 20–25% of women are diagnosed before menopause. In recent years, there has been an increasing of endometrial cancer occurring in young patients [2]. At present, a total abdominal hysterectomy (surgical removal of the uterus) with bilateral salpingo-oophorectomy is the most common therapeutic approach [3]. To improve the quality of patient life, early and accurate detection is urgently needed.

Current diagnostic tests for endometrial cancer include ultrasonographic measurements (monitoring of endometrial thickness). Biopsy and MRI [4] have been used in routine preoperative examination, but they are either invasive or expensive and not suitable for fast screening endometrial carcinoma in large scale. A method for cheap and fast screening endometrial carcinoma is needed. Although biopsy sampling is currently the most accurate and widely used screening technique, the diagnosis of biopsy sample has some limitations. For example, it is difficult to distinguish between the endometrial cancer and atypical endometrial hyperplasia [5]. Thus, there is need to develop an accurate, fast, convenient, and inexpensive method to diagnose the endometrial cancer at the early stage.

In recent years, near infrared (NIR) spectroscopy has attracted attention as a simple and inexpensive method for pathology studies. Typical applications were biodiesel analysis [6], characterization of human breast epithelial cells [7], dairy products analysis [8], resins bulk concentration evaluation during adsorption process [9], skin lesions [10]. NIR spectra can be assigned mostly to overtones and combinations of the molecular vibrations of C–H, N–H, and O–H groups, which are common to biological molecules [8–13]. Therefore, the absorptions of near-infrared radiation 769–2500 nm (13,000–4000 cm^{-1}) lead to a complex spectrum, which provides qualitative and quantitative information about the chemical composition of the tissue [14,15]. The concept of diagnosis based on NIRS is simple: cancer tissues differ from their healthy counterparts in their morphology and compositions (lipids, proteins, carbohydrates, and water). SVM was used for establishing the relationship between the endometrial cancer tissues and NIR spectra. The classes of tissue samples have been differentiated based on histopathology [16–19]. SVM was used for establishing the relationship between the endometrial cancer

* Corresponding author. Tel.: +86 10 68902490x803; fax: +86 10 68903047.
E-mail address: gusto2008@vip.sina.com (Z. Zhang).

1877-7821/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.canep.2011.10.009

tissues and NIR spectra. The classes of tissue samples have been differentiated based on histopathology. Any alteration in the chemical composition of the tissues can thus be detected by NIR spectroscopy and evaluated with the help of chemometrics methods which can extract latent information form the spectrum, therefore, chemometrics provides NIRS broad applications in various aspects [12,13,20–23].

Because cancer tissues differ from normal tissue in composition and histology, several research groups have been trying to establish diagnosis models based on the correlation between the tissue histology and NIR spectra. The studies of spectral differences between human cancer and normal tissues have been reported for breast [24], cervix [25], lung [26], colorectal [15], prostate [27,28] and other tissues. These studies were mostly based on the assignment of position and intensity of peaks. At shorter near-IR wavelengths, the heme proteins (oxyhemoglobin, deoxyhemoglobin, and myoglobin) and cytochromes dominate the spectra, and their absorptions are indicative of regional blood flow and oxygen consumption [19].

Owing to its nature, the quality of the NIR spectra is often worsened by the large number of overtone bands typically observed in near-infrared region, which are extensively overlapped with weak intensity of absorption, and sensitive to the external conditions (such as changes of temperature, pressure, apparatus and physical condition of samples, etc.) [23,29–31]. Therefore, it is difficult to assign NIR absorbance bands for biological samples. Chemometrics can enable NIR spectroscopy for many applications [6,8,10,13,21–23,29,31]. Previous studies on the early diagnosis of endometrial cancer in our laboratory have furnished good results [32]. Spectral preprocessing methods such as direct orthogonal signal correction (DOSC) [33], principal component orthogonal signal correction (PC-OSC) [34] are used to enhance NIR spectral classification models. The least squares support vector machines (LS-SVM) method, proposed by Suykens et al. [35], is a simplification of the traditional support vector machines (SVM). LS-SVM encompasses similar advantages by solving a set of linear equations (linear programming), which is much easier and computationally more efficient than solving the nonlinear equations (quadratic programming) employed by traditional SVM which includes support vector regression (SVR) and support vector classification (SVC) approaches depending on the problems. In this work LS-SVM was used for classification of tissue samples [36,37]. The fuzzy rule-building expert system (FuRES) is based on the concept of classification trees with a minimal neural network at each branch [38]. FuRES generates reproducible models and is a powerful classifier which had been used for several applications [39–41]. Various multivariate methods have been used to solve qualitative or quantitative problems. However, most methods used were based on linear algorithms, such as partial least squares (PLS) [42], linear discriminant analysis (LDA) [43], etc. In this work, we are trying to explore nonlinear classification methods for solving more complicated practical problems.

In this work, NIR spectra of a total of 18 normal, 30 hyperplasia, and 29 malignant tissue slices were collected. The spectra were converted to second derivatives by a Savitzky–Golay polynomial filter after multiplicative scatter correction (MSC). Then DOSC and PC-OSC were used to remove unwanted background variances from the spectra. The optimal LS-SVM based on NIR spectra of endometrial tissue can provide an efficient method for the early diagnosis of endometrial cancer.

2. Theoretical basis

2.1. LS-SVM

SVM is an algorithm from the machine learning community, developed by Vapnik and co-workers [44]. Due to remarkable generalization performance, SVM has attracted much attention and gained extensive application in pattern recognition and regression problems. SVM maps input data into a high dimensional feature space where objects may become linearly separable by a hyperplane. One reason that the SVM often performs better than other methods is that SVM was designed to minimize structural risk which has been shown to be superior to the traditional empirical risk minimization principle employed by conventional neural networks [12]. So the SVM is usually less vulnerable to overfitting the training data. Suykens and Vandewalle [35] proposed a modified version of the SVM called a least squares SVM (LS-SVM), which results in a set of linear equations instead of a quadratic programming problem and can extend the applications of the SVM. The LS-SVM model can be expressed as:

$$y(x) = \sum_{k=1}^{N} \alpha_k \varphi(x, x_k) + b \tag{1}$$

where x_k is the input vector, $y(x)$ is the corresponding target output, α_k is the Lagrange multiplier, φ is the kernel function, and b is the bias term.

The radial basis function (RBF) can model nonlinear relationships between spectra and target attributes, reduce the computational complexity of the training procedure, and give good performance. Thus, the RBF kernel was chosen as the kernel function of the LS-SVM in this paper. The LS-SVM model can be expressed as:

$$\phi(x_i, x_j) = e^{-(|x_i - x_j|^2 / 2\sigma^2)} \tag{2}$$

where σ^2 is the variance of the Gaussian kernel.

2.2. Fuzzy rule-building expert system (FuRES)

Using the iterative dichotomiser 3 (ID3) algorithm, FuRES provides local modeling and implements conjugate gradient optimization for the global minima of fuzzy classification entropy, $H(C|A)$ [38,45] to obtain a rule utilizing a linear discriminant. Fuzzy entropy $H(C|A)$ based on Shannon's information theory is minimized with the constraint that the first derivative of the entropy with respect to the computational temperature is maximized. Each weight vectors **w** is normalized so that a computational temperature t controls the fuzziness of the logistic function. Equations are given as

$$\chi_1(x_k) = \left(1 + e^{-(x_k w - b)/t}\right)^{-1} \tag{3}$$

$$\chi_2(x_k) = 1 - \chi_1(x_k) \tag{4}$$

$$p(c_i|a_j) = \frac{\sum_{k=1}^{n_i} \chi_j(x_k)}{\sum_{k=1}^{n} \chi_j(x_k)} \tag{5}$$

$$H(C|a_j) = -\sum_{i=1}^{n} p(c_i|a_j) \ln[p(c_i|a_j)] \tag{6}$$

$$H(C|A) = \sum_{j=1}^{2} p(a_j) H(C|a_j) \tag{7}$$

for which each rule in the tree is comprised of a weight vector **w**, a bias value b, and a computational temperature t. The multivariate rule resembles a sigmoidal neural network processing element with the exception that the weight vector is normalized to unit length. The kth object is multiplied by the normalized weight vector **w** and corrected for bias in Eqs. (3) and (4). The computational temperature controls the degree of fuzziness of the rule's logic given by $\chi_a(x_k)$ or membership function with large t