Biomedical Optics EXPRESS

Classification for breast cancer diagnosis with Raman spectroscopy

Qingbo Li, Qishuo Gao, and Guangjun Zhang[*]

School of Instrumentation Science and Opto-Electronics Engineering, Beihang University, Xueyuan Road No.37, Haidian District, Beijing, 100191, China
[*]qbleebuaa@buaa.edu.cn

Abstract: In order to promote the development of the portable, low-cost and in vivo cancer diagnosis instrument, a miniature laser Raman spectrometer was employed to acquire the conventional Raman spectra for breast cancer detection in this paper. But it is difficult to achieve high discrimination accuracy. Then a novel method of adaptive weight k-local hyperplane (AWKH) is proposed to increase the classification accuracy. AWKH is an extension and improvement of K-local hyperplane distance nearest-neighbor (HKNN). It considers the features weights of the training data in the nearest neighbor selection and local hyperplane construction stage, which resolve the basic shortcoming of HKNN works well only for small values of the nearest-neighbor. Experimental results on Raman spectra of breast tissues in vitro show the proposed method can realize high classification accuracy.

OCIS codes: (170.5660) Raman spectroscopy; (300.6450) Spectroscopy, Raman; (170.6510) Spectroscopy, tissue diagnostics.

References and links

1. D. Carter, "New Global Survey Shows an Increasing Cancer Burden," Am. J. Nurs. **114**(3), 17–18 (2014).
2. R. Siegel, J. Ma, Z. Zou, and J. Ahmedin, "Cancer statistics, 2014," CA. 64(1), 9–29 (2014).
3. Q. B. Li, X. J. Sun, Y. Z. Xu, L. M. Yang, Y. F. Zhang, S. F. Weng, J. S. Shi, and J. G. Wu, "Diagnosis of gastric inflammation and malignancy in endoscopic biopsies based on Fourier transform infrared spectroscopy," Clin. Chem. **51**(2), 346–350 (2005).
4. R. R. Alfano, G. C. Tang, A. Pradhan, W. Lam, D. Choy, and E. Opher, "Fluorescence spectra from cancerous and normal human breast and lung tissues," IEEE J. Quantum Electron. **23**(10), 1806–1811 (1987).
5. R. R. Alfano, C. H. Liu, W. L. Sha, H. R. Zhu, D. L. Akins, J. Cleary, R. Prudente, and E. Clemer, "Human breast tissues studied by IR Fourier transform Raman spectroscopy," Lasers Life Sci. **4**(1), 23–28 (1991).
6. Y. Pu, W. B. Wang, Y. L. Yang, and R. R. Alfano, "Native fluorescence spectra of human cancerous and normal breast tissues analyzed with non-negative constraint methods," Appl. Opt. **52**(6), 1293–1301 (2013).
7. S. K. Teh, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, and Z. Huang, "Near-infrared Raman spectroscopy for early diagnosis and typing of adenocarcinoma in the stomach," Br. J. Surg. **97**(4), 550–557 (2010).
8. C. H. Liu, Y. Zhou, Y. Sun, J. Y. Li, L. X. Zhou, S. Boydston-White, V. Masilamani, K. Zhu, Y. Pu, and R. R. Alfano, "Resonance Raman and Raman Spectroscopy for Breast Cancer Detection," Technol. Cancer Res. Treat. **12**(4), 371–382 (2013).
9. A. F. García-Flores, L. Raniero, R. A. Canevari, K. J. Jalkanen, R. A. Bitar, H. S. Martinho, and A. A. Martin, "High-wavenumber FT-Raman spectroscopy for in vivo and ex vivo measurements of breast cancer," Theor. Chem. Acc. **130**(4–6), 1231–1238 (2011).
10. C. Yu, E. Gestl, K. Eckert, D. Allara, and J. Irudayaraj, "Characterization of human breast epithelial cells by confocal Raman microspectroscopy," Cancer Detect. Prev. **30**(6), 515–522 (2006).
11. A. Zoladek, F. C. Pascut, P. Patel, and I. Notingher, "Non-invasive time-course imaging of apoptotic cells by confocal Raman micro-spectroscopy," J. Raman Spectrosc. **42**(3), 251–258 (2011).
12. A. S. Haka, Z. Volynskaya, J. A. Gardecki, J. Nazemi, J. Lyons, D. Hicks, M. Fitzmaurice, R. R. Dasari, J. P. Crowe, and M. S. Feld, "In vivo Margin Assessment during Partial Mastectomy Breast Surgery Using Raman Spectroscopy," Cancer Res. **66**(6), 3317–3322 (2006).
13. A. S. Haka, Z. Volynskaya, J. A. Gardecki, J. Nazemi, R. Shenk, N. Wang, R. R. Dasari, M. Fitzmaurice, and M. S. Feld, "Diagnosing breast cancer using Raman spectroscopy: prospective analysis," J. Biomed. Opt. **14**(5), 054023 (2009).
14. M. V. P. Chowdary, K. Kalyan Kumar, S. Mathew, L. Rao, C. M. Krishna, and J. Kurien, "Biochemical

#209903 - $15.00 USD Received 9 Apr 2014; revised 21 May 2014; accepted 23 May 2014; published 27 Jun 2014

correlation of Raman spectra of normal, benign and malignant breast tissues: a spectral deconvolution study," Biopolymers **91**(7), 539–546 (2009).

15. C. H. Liu, W. B. Wang, A. Alimova, V. Sriramoju, V. Kartazayev, and R. R. Alfano, "Monitoring changes of proteins and lipids in laser welded aorta tissue using Raman spectroscopy and basis biochemical component analyses," Proc. SPIE **7175**, 717504 (2009).
16. O. G. Okun, "K-local hyperplane distance nearest-neighbor algorithm and protein fold recognition," Pattern Recognit. Image Anal. **16**(1), 19–22 (2006).
17. P. Vincent, Y. Bengio, "K-local hyperplane and convex distance nearest neighbor algorithms," NIPS, 985–992 (2001).
18. M. S. Feld, R. Manoharan, J. Salenius, J. Orenstein-Carndona, T. J. Roemer, J. F. Brennan III, and Y. Wang, "Detection and characterization of human tissue lesions with near-infrared Raman spectroscopy," Proc. SPIE **2388**, 99–104 (1995).
19. G. Li, "Removing Background of Raman Spectrum Based on Wavelet Transform," *International Conference on Future Computer and Communication* (IEEE, 2009), pp. 198–200.
20. K. E. Shafer-Peltier, A. S. Haka, M. Fitzmaurice, J. Crowe, R. R. Dasari, and M. S. Feld, "Raman microspectroscopic model of human breast tissue: implications for breast cancer diagnosis in vivo," J. Raman Spectroscopy **33**(7), 552–563 (2002).
21. A. S. Haka, K. E. Shafer-Peltier, M. Fitzmaurice, J. Crowe, R. R. Dasari, and M. S. Feld, "Diagnosising breast cancer by using Raman spectroscopy," Proc. Nat. Acad. Sci. USA **102**(35), 12371–12376 (2005).
22. M. V. P. Chowdary, K. K. Kumar, J. Kurien, S. Mathew, and C. M. Krishna, "Discrimination of normal, benign, and malignant breast tissues by Raman spectroscopy," Biopolymers **83**(5), 556–569 (2006).

1. Introduction

Breast cancer is one of the major causes of female death. Data show 20% global increase in breast cancer from 2008 to 2012 [1]. In 2014, about 62570 cases of breast carcinoma in situ will be newly diagnosed in the United States. Breast cancer accounts for 15% of all female cancer deaths, which is second only to lung cancer in the United States [2]. In China, the incidence has also increased significantly in recent years, and ranked first in the female malignant tumors in some large cities, such as Beijing, Shanghai and Tianjin [3].

Since the early diagnosis is the key factor to increase the rate of survival time for the cancer patients, it is important to develop fast, less invasive, objective methods for the diagnosis of breast cancers. Raman spectroscopy, as a molecular spectroscopy, could detect the changes of molecular structure and composition. During the tumor formation, significant changes occurred in the structure and concentration of the main bimolecular, which constitute the cell and tissue, such as carbohydrates, lipids, proteins and nucleic acids. Because these changes occur earlier than the clinical symptoms appearance and tumor medical imaging detection, molecular spectroscopy has the potential to early diagnosis of the tumor [3–7]. Due to the characters such as sharp peaks, freeing from the interference of water, fewer samples required and without sample chemical treatment etc, Raman spectroscopy is promising to realize real-time and noninvasive detection at the molecular level.

Raman spectroscopic diagnosis technology of breast cancers has been developed recently. There are many investigations focus on Fourier Transform Raman spectroscopy (FTRS), Confocal Raman microspectroscopy (CRS), Resonance Raman spectroscopy (RRS) and Surface-enhanced Raman spectroscopy (SERS) for breast cancer diagnosis [8–15]. Using them, the Raman spectra could be acquired with lower fluorescence, higher spatial resolution, but these technologies generally use a large-sized Raman spectrometer or a large desktop microscope, which is also expensive and difficult to achieve clinical portable diagnosis. For Conventional Raman spectroscopy (RS), the Raman spectrometers tend to be small-sized, portable and low cost. Combined with the optical fiber probe, RS has promise for in vivo and in situ cancer detection. While due to the strong fluorescence background interference and low spectral signal-to-noise ratio, it is difficult to achieve high discrimination accuracy by using the miniature Raman spectrometer. Therefore, it is significant to investigate the discrimination analysis method for high classification accuracy. A few studies in [12–14] use the miniature Raman spectrometer to collect the RS spectra to diagnose breast cancers.

In this paper, a novel algorithm of adaptive weight K-local hyperplane (AWKH) is investigated for classification of the acquired Raman spectra from cancerous and normal

#209903 - $15.00 USD Received 9 Apr 2014; revised 21 May 2014; accepted 23 May 2014; published 27 Jun 2014

human breast tissues. It is an extension and improvement of K-local hyperplane distance nearest-neighbor (HKNN) [16]. HKNN performs well only for small values of the number of nearest-neighbor (K) because it assumes that every single feature of the training data is equally relevant for the nearest neighbors selection [17]. The feature weights measure the importance of each single feature in classification. For AWKH, the feature weight is estimated by using the ratio of the between-group to with-group sums of squares for the data assigned to the given classes. Then the higher weight corresponds to a feature with better class separation capability. In the paper, AWKH realized higher accuracy for the discrimination of the acquired Raman spectra compared to the classifiers support vector machine (SVM) and HKNN.

2. Materials and methods

2.1 Tissue specimens

A total of sixteen breast tissue samples were obtained from female patients in Peking University Third Hospital, including four normal tissues and twelve cancerous tissues. The mean age was 56 years with the oldest 88 years and the youngest 33 years. After the spectra were acquired, the samples were stored in liquid nitrogen and sent for the frozen section pathological diagnosis as the reference in the spectral analysis. The experimental procedures were approved by the Medical Ethics Committee of Peking University Third Hospital and the patients.

2.2 Raman spectral measurements

In order to promote the development of a clinical portable, low-cost and in vivo cancer diagnosis instrument, an Ocean Optics QE65Pro miniature fiber optic Raman spectrometer at a 785nm excitation wavelength was employed to acquire the conventional Raman spectra.

Specimens without any chemical treatment were frozen using liquid nitrogen and maintained until thawed at room temperature. They were placed in the glass slide for Raman spectral measurement. The integration time is 30s. All the spectra were acquired in the wavelength range of interest, from 700 to 2000 cm^{-1}. In the Spectral acquisition process, every sample was measured at different pathology locations, and for every same pathology location three spectra were measured and averaged in order to reduce the noise level. Each Raman spectrum was labeled according to the pathological diagnosis. In order to reflect the experiment results objectively, the sample spectra were collected on the same environmental conditions and the experiments were conducted two days. 75 Raman spectra (16 normal and 59 cancerous) obtained in the first day and 58 Raman spectra (18 normal and 40 cancerous) obtained in the second day.

2.3 Software

All the examined preprocessing and classification algorithms were implemented and tested in Matlab 2009a. In addition, the SVM toolbox was used.

2.4 Preprocessing algorithm

The spectra collected using Ocean Optics QE65Pro Raman spectrometer yielded noise and fluorescence background. The noise was removed by wavelet transform and the fluorescence background was removed by fitting the smoothed spectra to a third-order polynomial function.

The wavelet transform [18, 19] was introduced as follows:

The discrete wavelet transform is defined as:

#209903 - $15.00 USD Received 9 Apr 2014; revised 21 May 2014; accepted 23 May 2014; published 27 Jun 2014

$$f(t)=\sum_{k\in z}c_{J,k}\psi_{J,k}(t)+\sum_{j=1}^{J}\sum_{k\in z}d_{j,k}\psi_{j,k}(t) \tag{1}$$

where $\psi_{J,k}(t)$ is the wavelet basis function, $c_{J,k}$ is the J layer approximation coefficient of spectral signal that is also the low frequency coefficient, $d_{j,k}$ is the j layer detail coefficient of spectral signal as well as the high frequency coefficient .

The specific process is shown as follows:

Step 1: choose a wavelet function and a decomposition scale.

Step 2: deal with the high frequency coefficients of wavelet decomposition by threshold processing. A soft threshold function is used in this paper:

$$\hat{w}_{j,k}=\begin{cases}\operatorname{sgn}(w_{j,k})(|w_{j,k}|-\lambda) & |w_{j,k}|\geq\lambda; \\ 0 & |w_{j,k}|<\lambda;\end{cases} \qquad \lambda=\sigma*(2\log(N))^{1/2} \tag{2}$$

Step 3: reconstruct the spectrum signal, according to the J layer low frequency coefficient and the high frequency coefficient after threshold processing from 1th to the jth layer.

2.5 AWKH algorithm

Adaptive weight k-local hyperplane (AWKH) algorithm is an improvement and extension of K-local hyperplane distance nearest-neighbor (HKNN) algorithm. HKNN [16] performs well only for small values of K, it sufferers from bias for data with high dimensions, AWKH resolves the problem by considering the features weights when calculate the distance between the test set samples and hyperplane. The feature weight is estimated by using the ratio of the between-group to with-group sums of squares. Feature weights are computed such that higher weight corresponds to a feature with better class separation capability. And the bias when HKNN is used in high dimensions is settled by considering the shape of the neighborhood around the test sample. Raman spectra of breast tissues contain some specific peaks which are beneficial to classification but not common exist. Since AWKH only considers the relationship between samples, so when dealing with Raman spectroscopy, AWKH can obtain high accuracy.

The specific process of the AWKH algorithm can be summarized as follows:

Suppose the training set consists of L samples with J classes. Each training sample consists of d input features $x_i=(x_{i1},...,x_{id})^T$ with known class label $y_i=c\ \ (i=1,...,L;c=1,...,J)$ .The goal is to predict the class label of a query with input vector $q=(q_1,...,q_d)^T$.

Step 1: calculate the feature weight w of the training sample according to the formula as follows:

$$R_j=\frac{\sum_c(\bar{x}_{cj}-\bar{x}_j)^2}{\sum_i\sum_c I(y_i=c)(x_{ij}-\bar{x}_{cj})^2}$$
$$w_j=\frac{\exp(R_j)}{\sum_{j=1}^{d}\exp(R_j)} \qquad \forall j=1,...,d \tag{3}$$

where $\bar{x}_j$ denotes the jth component of the grand class centroid and $\bar{x}_{cj}$ denotes the jth component of class centroid of class c; $I(\bullet)$ denotes the indicator function, it equals1when $y_i=c$, otherwise, it equals 0; x_{ij} denotes the jth component of the ith training sample.

#209903 - $15.00 USD Received 9 Apr 2014; revised 21 May 2014; accepted 23 May 2014; published 27 Jun 2014

Step 2: calculate the weighted Euclidean distance metric D between x_i and q , the formula is as follows:

$$D(x_i, q) = (\sum_{j=1}^{d} w_j (x_{ij} - q_j)^2)^{1/2} \tag{4}$$

Step 3: according to the Euclidean distance D , select K nearest neighbors of class c $p_c = (p_{c1}, ..., p_{cK})$ for the given query q , then construct the local hyperplane of class c with p_c :

$$\begin{aligned} &LH_c(q) = \{s \mid s = \sum_{i=1}^{K} \alpha_i V_{\cdot i} + m_c\} \\ &m_c = \frac{1}{K} \sum_{i=1}^{K} p_{ci} \\ &V_{\cdot i} = p_{ci} - m_c \\ &\alpha = (\alpha_1, ..., \alpha_K)^T \end{aligned} \tag{5}$$

Step 4: calculate the minimum distance between q and $LH_c(q)$:

$$\begin{aligned} J_c(q) &= \min_{\alpha} \sum_{j=1}^{d} w_j (V_{j\cdot} \alpha + m_{cj} - q_j)^2 + \lambda \alpha^T \alpha \\ &= \min_{\alpha} (s - q)^T W (s - q) + \lambda \alpha^T \alpha \\ W &= diag(w_1, ..., w_d) \end{aligned} \tag{6}$$

where λ is the regularization parameter, which avoids α being too large. Solve the equation $\frac{\partial J_c(q)}{\partial \alpha} = 0$, then achieve $\alpha = (\mathrm{U}^\mathrm{T}\mathrm{V}+\lambda I_{\mathrm{n}_\mathrm{c}})\backslash(\mathrm{U}^\mathrm{T}(\mathrm{q}\text{-}\mathrm{m}_\mathrm{c}))$, where $\mathrm{U}^\mathrm{T}=\mathrm{V}^\mathrm{T}\mathrm{W}$.

Step 5: the class label of q is assigned as: $\mathrm{label(q)=argmin}_\mathrm{c} J_c(q)$.

3. Results and discussion

133 spectra were obtained by Raman spectroscopic method with the scan region 700 cm^{-1} to 1800 cm^{-1}. Each Raman spectrum was labeled according to the pathological diagnosis.

3.1 Spectral preprocessing

Symmlet-5 wavelet filter and four-decomposition scale were adopted to reduce noise, and then a third-order polynomial was adopted to remove fluorescence background and baseline corrected. The mean Raman spectra of normal and cancerous tissues before preprocessing and after preprocessing are shown respectively (see Fig. 1, Fig. 2).

#209903 - $15.00 USD Received 9 Apr 2014; revised 21 May 2014; accepted 23 May 2014; published 27 Jun 2014

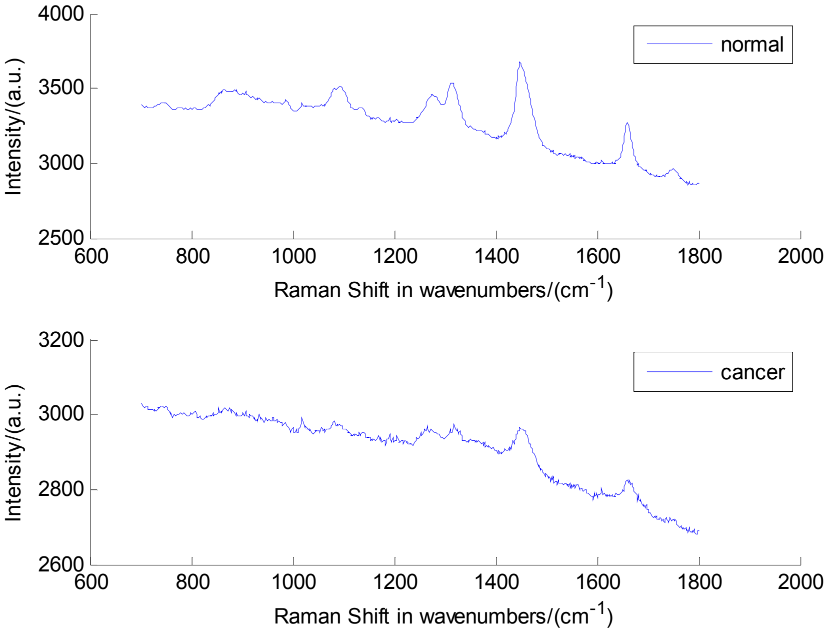

Fig. 1. Typical Raman spectra of breast tissues before preprocessing.

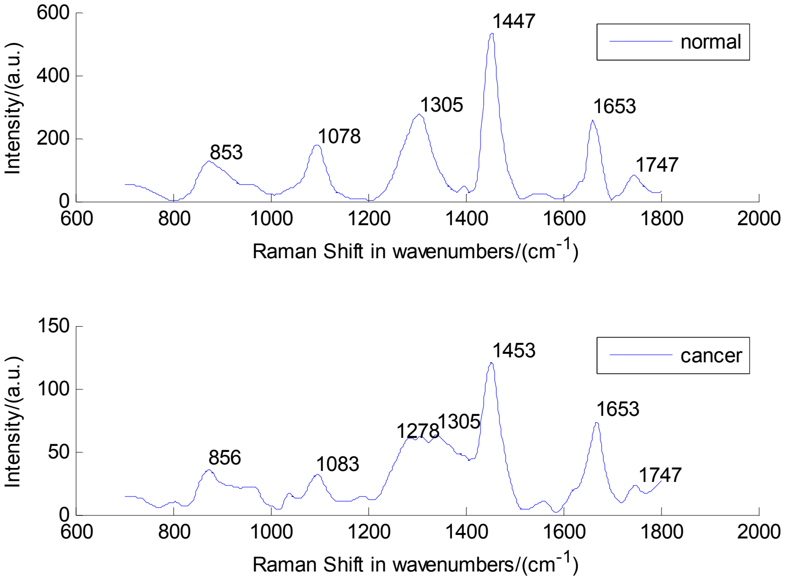

Fig. 2. Typical Raman spectra of breast tissues after preprocessing.

The raw spectra of normal tissues showed evident peaks (see Fig. 1), while, there are only small peaks in the raw spectra of cancerous tissues because of the effect of the noise and the fluorescent background.

#209903 - $15.00 USD   Received 9 Apr 2014; revised 21 May 2014; accepted 23 May 2014; published 27 Jun 2014

The quality of Raman spectra has improved greatly after data preprocessing (see Fig. 2). The Raman spectra are smoother, the Raman peaks of normal and cancerous tissues are distinguished, and especially the differences between Raman spectra of normal and cancerous tissues are more pronounced after preprocessing.

The essence of the wavelet transform is that project the spectrum signal in the wavelet basis function, decompose the spectrum signal in time domain and frequency domain, get the wavelet approximation coefficients and detail coefficients. Where, the detail signal reflects the local nuances, and most of them are noise in the high frequency region. So wavelet transform could be used to remove the noise of the Raman spectra, and optimize the quality of spectra.

The Raman peaks of normal tissues and cancerous tissues (see Fig. 2) are displayed. The spectral profile of normal tissues is indicative of higher levels of lipids. In comparison, the spectral profile of pathological tissues indicates the presence of more proteins and fewer lipids. The spectral features (1078, 1305, 1447, 1653 and 1747cm^{-1}) of normal tissues indicate a dominance of lipids. The spectral profiles of cancerous tissues (1083, 1278, 1453cm^{-1}) indicate the presence of proteins. The peak intensities of 1305, 1653, 1747 cm^{-1} in cancerous tissues decrease obviously compared to those in normal tissues. The peak position representing protein molecules appears at 1278 cm^{-1} in cancerous tissues, while almost disappears in normal tissues. These changes reflect that during tumor formation, the protein, lipid and nucleic acid molecular changed in the configuration, component and quantity, and the proportion of proteins significantly increased against to the greatly reduced lipids proportion. This observation corroborates earlier studies [20, 21]. As is well known, cancerous tissues contain more proteins relative to normal tissues and adipose–rich noncancerous, which is the basis of spectroscopic diagnosis.

Specific assignments of individual peaks could be found in Table 1.

Table 1. Peak positions and assignments of Breast Tissue[α]

Peak position(cm^{-1})	Major assignment
1078	C-C or C-O stretch (lipid)
1278	Amide III(C-N stretch) (protein)
1305/1308	Amide III, α-helix, C-C str&C-H (protein)
1447	Scissoring mode of methylene (CH2) (lipid)
1453	CH2 deformation (protein)
1653	lipid
1747/1750	C = O stretch (lipid)

[α] See [8, 18, 22].

#209903 - $15.00 USD Received 9 Apr 2014; revised 21 May 2014; accepted 23 May 2014; published 27 Jun 2014

3.2 Statistical analysis

The whole data set was split into a training set and test set, and each classifier was learned on the training set and applied on the test set.

The 75 Raman spectra (16 normal and 59 cancerous) obtained in the first day after preprocessing were selected as the training set, and the 58 Raman spectra (18 normal and 40 cancerous) obtained in the second day after preprocessing were selected as the test set.

The training set and the test set are normalized to zero mean and unit variance first, then, classify the test set by AWKH、HKNN and SVM classifier respectively. The two parameters K and λ for AWKH were set as [1:20] and 10 respectively. The parameter K for HKNN in reference [16] was set as [1:20]. Then, select the result with highest testing accuracy as the optimized classification result.

The experimental results are summarized in Table 2 and Table 3. Here, the optimized parameters for AWKH are $K = 4$, $\lambda = 10$, and $K = 3$ for HKNN. Table 2 displays the classification results of test set with AWKH. Table 3 shows the results obtained with three different methods.

Table 2. Classification results of test set with AWKH

	The predicted cancerous number (T +)	The predicted normal number (T-)
The real cancerous number (D +)	39	1
The real normal number (D-)	1	17

Table 3. Comparison of the results for the AWKH, HKNN, SVM

Method	Sensitivity (%)	Specificity (%)	Positive* (%)	Negative** (%)	Accuracy (%)
AWKH	97.5	94.4	97.5	94.4	96.6
SVM	92.5	88.9	94.9	84.2	91.4
HKNN	92.5	77.8	90.2	82.4	87.9

* The positive predictive value, ** the negative predictive value

In Table 3, it can be seen that AWKH achieves the highest testing accuracy among three different classifiers. Especially, AWKH is much more accurate than SVM classifier.

The classification accuracy with different K value using AWKH and HKNN is shown respectively (see Figs. 3 and 4). In wake of the increase of K value, the accuracy with HKNN decreased (see Fig. 4). The accuracy of AWKH stays stable for K value between 4 and 20 (see Fig. 3).

#209903 - $15.00 USD Received 9 Apr 2014; revised 21 May 2014; accepted 23 May 2014; published 27 Jun 2014

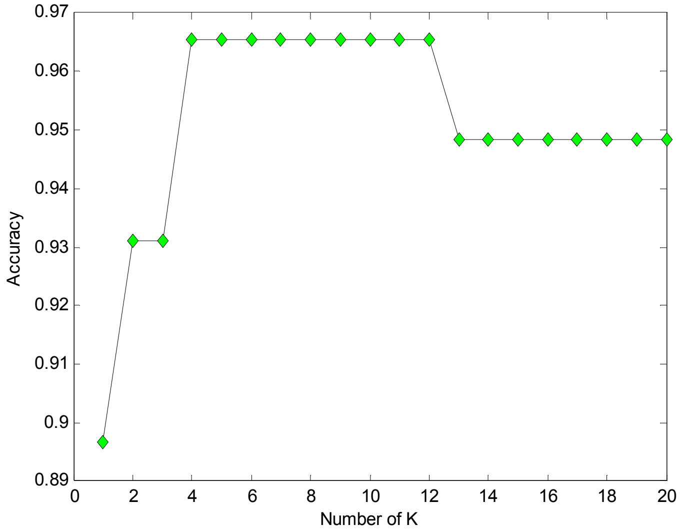

Fig. 3. The classification accuracy at different K value for AWKH.

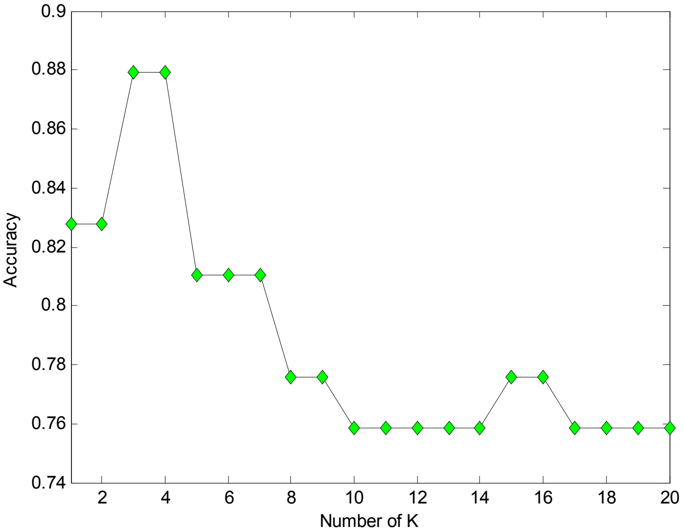

Fig. 4. The classification accuracy at different K value for HKNN.

The optimal value of the parameter λ depends on K . For small K , the model can achieve good results without λ . With larger K , the model tends to be various and more complex, so that the regularization can help to improve the performance. In contrast, HKNN does not have the advantage.

The feature weights measure the importance of every single feature of spectral data. HKNN performs well only for small values of K because it assumes that every single feature is equally relevant for classification which may yield unsatisfactory performance when data with high dimensions. AWKH computes the ratio of the between class to the within class squared distances to estimate the features weights. The nearest neighbors are selected by the weighted Euclidean distance between the test sample and training set. The resulting nearest

#209903 - $15.00 USD Received 9 Apr 2014; revised 21 May 2014; accepted 23 May 2014; published 27 Jun 2014

neighbors are then associated with the most discriminant feature space. The local hyperplance constructed based on these neighbors is more convincing which leading to the classification result directly. With small K , HKNN may be well formulated, but with large K , HKNN will suffer the unsatisfactory performance. Moreover, for the higher dimensionality of the extracted features, the more points from each class are needed to accurately estimate the localized model, hence K should be larger. AWKH considers the features weights make it fairly robust on the choice of K , which is generally a desirable characteristic of a K-local learning algorithm.

Then the data processing was conducted two more times. The 58 Raman spectra obtained in the second day were selected as the training set, and the other spectra obtained in the first day were selected as the test set. Table 4 shows the results obtained by three different methods with optimal parameters.

Table 4. Comparison of the results for the AWKH, HKNN, SVM

Method	Sensitivity (%)	Specificity (%)	Positive* (%)	Negative** (%)	Accuracy (%)
AWKH	94.9	93.8	98.2	83.3	94.7
SVM	93.2	93.8	98.2	78.9	93.3
HKNN	88.1	87.5	96.3	66.7	90.7

* The positive predictive value, ** the negative predictive value

Finally, the total 133 Raman spectra were split into two data sets randomly for ten times, Every time 80 Raman spectra after preprocessing were selected as the training set, and the other 53 Raman spectra after preprocessing were selected as the test set. Then the algorithms were examined. Table 5 shows the average accuracy of the ten experiments using three different methods with optimal parameters.

Table 5. Comparison of the results for the AWKH, HKNN, SVM

Method	Accuracy (%)
AWKH	95.8
SVM	92.4
HKNN	87.6

From the experimental results above, AWKH shows great advantage for the classification of Raman spectra of breast tissues.

Although the two algorithms have similar mechanisms for AWKH and HKNN, AWKH performed better in the experiment. The data sets with irrelevant or redundant features like Raman spectra data can be classified more accurate with AWKH because it considers the features weights. For SVM, kernel function needs to be used for every single sample and the choices of the parameters for the kernel is important, which are complex and unstable. But it is worth noting that, SVM can perform well when the parameters are optimal and it has advantage for the large-scale test set.

4. Conclusions

As evident from the studies conducted so far, it is quite feasible to classify normal and pathological breast tissues optically. The ultimate goal of optical spectroscopy methods is to develop clinical portal, low-cost and in vivo cancer diagnosis instrument. For such applications, a miniature laser Raman spectrometer with a 785nm excitation was employed to

#209903 - $15.00 USD Received 9 Apr 2014; revised 21 May 2014; accepted 23 May 2014; published 27 Jun 2014

acquire the conventional Raman spectra of breast tissues. Then the preprocessing procedures were investigated. At the end of the paper, a novel classification algorithm AWKH is proposed. This novel algorithm improves the HKNN method by stressing the feature weight.

The experimental results show that the proposed classification algorithm is an effective method. AWKH achieved high classification accuracy even when the strong fluorescence background interference and low spectral signal-to-noise ratio were obtained by the miniature laser Raman spectrometer. It is helpful to promote the development of clinical portable diagnosis technology and a desire to apply the technology in vivo breast cancer diagnosis using Raman spectroscopy in the later research.

Acknowledgments

This work is supported by Programs for Changjiang Scholars and Innovative Research Team (PCSIRT) in University of China (IRT0705).

#209903 - $15.00 USD Received 9 Apr 2014; revised 21 May 2014; accepted 23 May 2014; published 27 Jun 2014