International Conference on Image and Signal Processing (ICISP10), Trois-Rivieres, QC, Canada, June 30-July 2, 2010

Classification of High-Resolution NMR Spectra Based on Complex Wavelet Domain Feature Selection and Kernel-Induced Random Forest

Guangzhe Fan[1], Zhou Wang[2], Seoung Bum Kim[3], and Chivalai Temiyasathit[4]

[1]Dept. of Statistics & Acturial Science, University of Waterloo, Waterloo, ON, Canada
[2]Dept. of Electrical & Computer Engineering, University of Waterloo, Waterloo, ON, Canada
[3]Dept. Industrial Systems & Information Engineering, Korea University, Seoul, Korea
[4]International College, King Mongkut's Inst. of Technology Ladkrabang, Bangkok, Thailand

Abstract. High-resolution nuclear magnetic resonance (NMR) spectra contain important biomarkers that have potentials for early diagnosis of disease and subsequent monitoring of its progression. Traditional features extraction and analysis methods have been carried out in the original frequency spectrum domain. In this study, we conduct feature selection based on a complex wavelet transform by making use of its energy shift-insensitive property in a multi-resolution signal decomposition. A false discovery rate based multiple testing procedure is employed to identify important metabolite features. Furthermore, a novel kernel-induced random forest algorithm is used for the classification of NMR spectra based on the selected features. Our experiments with real NMR spectra showed that the proposed method leads to significant reduction in misclassification rate.

Keywords: High-resolution NMR spectrum; Metabolomics; Classification tree; Random forest; Complex wavelet transforms; False discovery rate; Kernel

1 Introduction

High-resolution nuclear magnetic resonance (NMR) spectroscopy has been found to be useful in characterizing metabolic variations in response to disease states, genetic medication, and nutritional intake. Given the thousands of feature points in each NMR spectrum, the first step is to identify the features that are mostly related to the problems being studied. Many of the feature points are either redundant or irrelevant. Removing them may largely reduce the computational cost while improving the stability (e.g., noise robustness) of the subsequent analysis and classification processes. Such dimensionality reduction procedures are mostly carried out directly in the original frequency domain. The widely used methods for identifying important metabolite features in spectral data include principal component analysis (PCA) and partial least squares (PLS) [1,2], but the principle components from PCA or PLS do not provide clear interpretations with respect to the original features. In [3], a genetic programming based method was proposed to select a subset of original metabolite features in NMR spectra for the classification of genetically modified barley, though the method may not be reliable for high-dimensional and noisy data.

The wavelet transform provides a powerful and flexible framework for localized analysis of signals in both time and scale, but its applications to NMR spectra has not been fully exploited. In [4,5], wavelet transform was used for the detection of small chemical species in NMR spectra by suppressing the water signal. In [6], decimated discrete wavelet transform was employed to analyze mass spectrometry data (similar to NMR spectra) with class information. In [7], NMR spectra were analyzed using complex wavelet transforms, which have the important property of energy shift-insensitive. In particular, the false discovery rate based multiple test procedure leads to more reliable feature selection results when intensity and position shifts exist between multiple NMR spectra being compared (which is always the case in the data acquired in practice). NMR spectrum based classification is not only a practically useful application. It also provides a direct test of the quality of the feature extraction procedure. In [7], a simple classification tree algorithm was used. In our present study, we used Gabor wavelet transform which achieved the best result in [7]. We employed a new cross-validated testing scheme and identify different feature sets for our usage. To test the results of our feature selection scheme, we also compare three different classification approaches: classification tree, random forest and kernel-induced random forest, which is a novel algorithm for the classification of high-resolution NMR spectra. In [8], a classification tree algorithm was described in detail. Later [9] proposed the ensemble of classification tree which was called random forest. Random forest is a powerful classification tool which uses many randomly generated large classification trees and combines them to vote for a decision. The instability of a single classification tree was greatly reduced by the ensemble. A kernel-induced random forest method was proposed in [10]. A kernel function is computed for every two observations based on all the features or a reduced feature space. Then the observations are used to classify other observations via a recursive partitioning procedure and its ensemble model. The classification accuracy is improved with the kernel-induced feature space.

2 Method

2.1 Complex wavelet transform

We consider complex wavelets as dilated/contracted and translated versions of a complex-valued "mother wavelet" $w(x) = g(x)e^{j\omega_c x}$, where ω_c is the center frequency of the modulating band-pass filter, and $g(x)$ is a slowly varying and symmetric real-valued function. The family of wavelets derived from the mother wavelet can be expressed as:

$$w_{s,p}(x) = \frac{1}{\sqrt{s}} w\left(\frac{x-p}{s}\right) = \frac{1}{\sqrt{s}} g\left(\frac{x-p}{s}\right) e^{j\omega_c (x-p)/s}, \quad (1)$$

where $s \in R^+$ and $p \in R$ are the scale and translation factors, respectively. Considering the fact that $g(-x) = g(x)$, the wavelet transform of a given real signal $f(x)$ can be written as:

$$F(s,p)=\int_{-\infty}^{\infty} f(x)\,w^{*}_{s,p}(x)\,dx=\left[\, f(x)*g\left(\frac{x}{s}\right)e^{j\omega_c x/s}\,\right]_{x=p}. \qquad (2)$$

In other words, we can use this to compute the wavelet coefficient $F(s,p)$ at any given scale s and location p. Using the convolution theorem and the shifting and scaling properties of the Fourier transform, it is not difficult to derive that

$$F(s,p)=\frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\sqrt{s}\,G(s\omega-\omega_c)\,e^{j\omega p}\,d\omega, \qquad (3)$$

where $F(\omega)$ and $G(\omega)$ are the Fourier transforms of $f(x)$ and $g(x)$, respectively. Now suppose that the function $f(x)$ has been shifted by a small amount Δx, i.e., $f'(x)=f(x+\Delta x)$. This corresponds to a linear shift in the Fourier domain: $F'(\omega)=F(\omega)e^{j\omega_c\Delta x}$. Substitute this into Eq. (3) and assume that Δx is small relative to the envelop function $g(x)$, then we can derive

$$F'(s,p)=F(s,p)\ . \qquad (4)$$

This implies that the magnitude (or energy) of the complex wavelet coefficient does not change significantly with a small translation. Such an energy shift-insensitive property is very important in the analysis of NMR spectra because a small misalignment between multiple NMR spectra is unavoidable (even after preprocessing), and the misalignment may interfere with direct comparisons between NMR spectra.

Among the various complex wavelets available, we choose the Gabor wavelets mainly for two reasons: First, according to the Gabor uncertainty principle, the time-frequency resolution of a signal is fundamentally limited by a lower bound on the product of its bandwidth and duration, and the Gabor filters are the only family of filters that achieve this lower bound [11]. In other words, the Gabor filters provide the best compromise between simultaneous time and frequency signal representations. Second, the Gabor wavelets are easily and continuously tunable for both the center frequencies and for bandwidths.

2.2 Feature selection based on a multiple testing procedure

The most straightforward approach for feature selection in the wavelet transform domain is thresholding. However, this method may result in ignorance of small magnitude coefficients that are indeed important for classification. In this study, we identify complex wavelet coefficient features to maximize the separation of classes. More specifically, a multiple testing procedure that controls the false discovery rate (FDR) is employed to identify significant Gabor coefficients that discriminate between the spectra under different conditions. The FDR is the error rate in multiple hypothesis tests and is defined as the expected proportion of false positives among all the hypotheses rejected [12]. In our problem, the rejected hypothesis is interpreted as the significant coefficients necessary for classification.

The FDR-based procedure is explained with our experimental data. Let δ_{jk} be the magnitude of the Gabor coefficient at the *k*-th position of the *j*-th class. Our

experimental data comprise 136 NMR spectra in which half of the spectra were taken from the zero-SAA phase and the other half were taken from the supplemented-SAA phase. The goal is to identify a set of δ_k that maximizes the separability between the two SAA phases. For each wavelet coefficient, a null hypothesis states that the average magnitudes of Gabor coefficients are equal between the two SAA phases, and the alternative hypothesis is that they differ. The two-sample *t* statistic is

$$t_k = \frac{\bar{\delta}_{1k} - \bar{\delta}_{2k}}{\sqrt{\frac{\hat{\sigma}_{1k}^2}{n_1} + \frac{\hat{\sigma}_{2k}^2}{n_2}}} \tag{5}$$

where $\bar{\delta}_{1k}, \hat{\sigma}_{1k}^2$, and n_1 are the sample mean, variance, and the number of samples from the first condition, respectively. Similarly, $\bar{\delta}_{2k}, \hat{\sigma}_{2k}^2$, and n_2 are obtained from the second condition. By asymptotic theory, t_k approximately follows a *t*-distribution on the assumption that the null hypothesis is true. Using this, the *p*-values for δ_k can be obtained. In multiple testing problems, it is well known that applying a single testing procedure leads to an exponential increase of false positives. To overcome this, the methods that control family-wise error rates have been proposed. The most widely used one is the Bonferroni method that uses a more stringent threshold [13]. However, the Bonferroni method is too conservative, and it often fails to detect the "true" significant features. A more recent multiple testing procedure that controls FDR was proposed by Benjamini and Hochberg [12]. The advantage of the FDR-based procedure is that it identifies as many significant hypotheses as possible while keeping a relatively small number of positives [14,15].

2.3 Kernel-induced Classification Tree and Random Forest

A classification model was used to examine the advantage of using the complex wavelet transform and FDR-based feature selection in NMR spectra. We used a classification tree, one of the widely used classification methods. Classification trees partition the input (feature) space into disjoint hyper-rectangular regions according to performance measures such as misclassification errors, Geni index, and cross-entropy and then fit a constant model in each disjoint region [16]. The number of disjoint regions (equivalent to the number of terminal nodes in a tree) should be determined appropriately because a very large tree overfits the training set, while a small tree cannot capture important information in the data. In general, there are two approaches to determining the tree size.

The first approach is the direct stopping methods that attempt to stop tree growth before the model overfits the training set. The second approach is tree pruning that removes the leaves and branches of a full-grown tree to find the right size of the tree. In the present study the Geni index was used as a performance measure. To determine tree size, we stop the growth of a tree when the number of data points in the terminal node reaches five.

In order to estimate the true misclassification rate of classification tree models, we used a cross-validation technique. Specifically, we used a four-fold cross validation in

which the experimental data were split into four groups corresponding to four subjects. Three subjects were used for training the models, and the one remaining subject was used for testing. This process was repeated three more times. The final classification results from the four different testing samples were then averaged to obtain the misclassification rates (or cross-validated error rates) of the classification tree models.

A *kernel* is a function *K*, such that for all $\vec{x}_i$ and $\vec{x}_j \in X^p$, $i, j = 1, 2, \ldots, n$

$$K(\vec{x}_i, \vec{x}_j) = <\boldsymbol{\varphi}(\vec{x}_i), \boldsymbol{\varphi}(\vec{x}_j)> \tag{6}$$

where $\boldsymbol{\varphi}$ is a (non-linear) mapping from the input space to an (inner product) feature space. If the observation *i* is fixed in the training sample, and observation *j* is a new input, then the kernel function above can be treated as a new feature defined by observation *i*, denoted as $K(\vec{x}_i, \cdot)$. Some popular kernels are inner product kernel, polynomial kernel and Gaussian (radial basis) kernel.

A classification tree model is a recursive partitioning procedure in the feature space. Starting from the root node, at each step, a greedy exhaustive search is implemented to find the best splitting rule such as "$X_i < c$" for numerical features. If the answer is yes, then the observation will move to the left child node and move to the right child node otherwise. The procedure is implemented recursively until a very large binary tree is constructed. A large tree usually overfits the training sample. Then cross-validation is used to prune the tree back to its proper size. A single classification tree described above is highly interpretable but quite instable and weak in prediction. An example is shown in Fig. 1. A random forest algorithm is simply a replication of the classification tree procedure while introducing a random vector in the construction space, such as limiting the number of features to be searched at every step growing the tree and/or bootstrapping the data set. The trees in a random forest are usually very large and need no pruning. Due to the instability of classification trees, they are quite different and diversified when the random vector is introduced in the process. Each classification tree generally is a low-bias but high-variance model. When they are combined to vote for a decision, the variance is reduced and the classification power is very strong. Another nice result is that including more trees in the random forest will not overfit the training sample. However, the random forest described above can only deal with numeric data. For non-numeric data such as images, they cannot be directly used.

Instead of using the original features in the data space to construct splitting rules, kernel functions based on observations are used. Since the definition of kernel is very flexible to handle various types of data, the potential of random forest is greatly extended and enhanced. Not only the feature space is enlarged, but also some complicated and non-linear patterns between observations can be directly learned by random forest. Please see the second tree in the figure for an example of a kernel-induced classification tree. A kernel-induced random forest is simply a replication of many such trees with a random vector introduced. By default, the kernel we use is Gaussian.

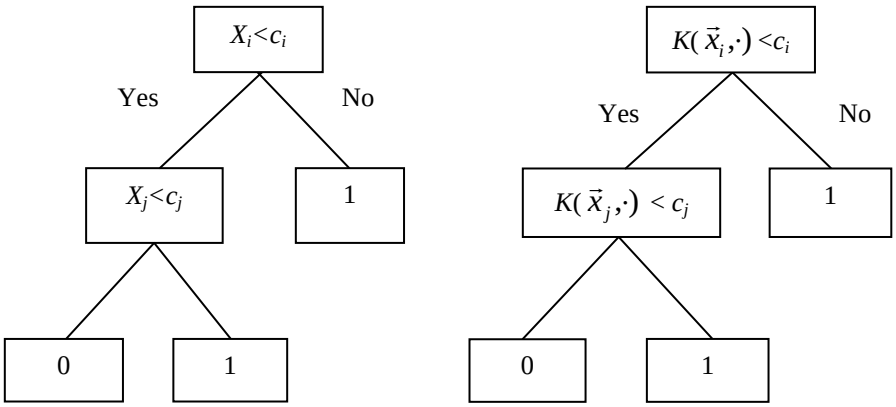

Fig. 1. Left: an example of a classification tree with 3 terminal nodes. Note that X_i and X_j are features in the data space. Right: an example of a kernel-induced classification tree with 3 terminal nodes. Note that X_i and X_j are observation vectors in the data space and the $K(\ ,\ )$'s are kernel functions defined on these observations.

3 Result

3.1. Experimental data

We used plasma samples obtained from four healthy subjects under controlled metabolic conditions in the Emory General Clinical Research Center (GCRC). The subjects signed an informed consent approved by the Emory Institutional Review Board. During the 12-day GCRC admission, the subjects consumed defined diets at standardized intervals. For the first two days (equilibration), the subjects consumed balanced meals from a plan in which foods were selected to ensure adequate energy, protein and sulfur amino acid (SAA) intake (SAA at 19 mg/kg/day). After this phase, subjects were placed on constant semipurified diets designed to alter SAA intake. The diets provided adequate energy and amino acid nitrogen to meet the estimated maintenance needs of individual subjects. The L-amino acid component of the diet was altered to provide zero SAA during the initial five days and 117 mg/kg per day during the latter five days of the GCRC stay. Blood was drawn serially 34 times from four subjects over ten days, and ^{1}H-NMR spectra were obtained by a Varian INOVA 600 MHz instrument. During the first 17 time points, blood was collected from each subject consuming zero SAA (zero-SAA phase) and 117 mg/kg per day SAA during the latter 17 time points (supplemented-SAA phase). Thus, the total number of spectra used in this study is 136 (4 subjects×34 spectra).

Raw NMR spectra require preprocessing, which includes phase/baseline correction, elimination of uninformative spectral regions containing no significant metabolite signals, alignment, and normalization relative to the internal standard. The NUTS software (Acron NMR Inc., Livermore, CA) was used for phase/baseline

correction. To adjust for the variable suppression of the large water signal in NMR spectra and enhance the detection of metabolites, water signal and other uninformative spectral regions were eliminated. We used MATLAB (Mathwork Inc., Natick, MA) with the beam search algorithm [17] for initial spectral alignment. Finally, normalization of NMR spectra was achieved by scaling to the integral of the internal standard.

3.2 Classification results

To evaluate the adequacy of the metabolite features obtained from Sections 3.1 and 3.2, we used the Gabor wavelet transformed data to show our classification performances for different approaches. Originally, there were 8444 features in the domain. After Gabor wavelet transformation, there were 8181 features. Then with the FDA-based procedure at the significance level of 0.01, there were 20 features selected. To provide an honest estimate of the classification performance of the different approaches under the FDA procedure, we use the following strategy: we divide the 136 spectra into four folds, i.e., each subject with 34 spectra is one fold. We use this subject-based cross validation because we try to show our potential to use current information on some subjects to classify other subjects. If the whole data set is randomly divided into a number of folds without the consideration of subjects, the result may not be informative to evaluate the true performance of our approach for the future. Then we separate each fold (subject) out as test set and apply the FDA procedure on the training data combined from the rest of the three folds. Note that when the procedure of cross-validation is used, these selected features may not be the same for different folds. Classification error rates obtained from cross validation are shown to evaluate the efficacy of the three proposed classification methods with our feature selection scheme (Table 1). Our experiences suggest that feature selection is important for all three classification approaches. In the selected feature space, the kernel-induced random forest performs the best among the three methods while random forest ranks the second.

Table 1. Cross-validated Misclassification Rate for the three classification approaches with FDA feature selection.

Method	Classification Tree	Random Forest	Kernel-induced Random Forest
Error rate	33.1%	29.5%	26.5%

4 Conclusion

We have used a complex wavelet transform combined with the FDR-based feature selection method to improve feature selection and classification of high-resolution

NMR spectra. We also compared three different classification methods and introduced the novel approach of kernel-induced random forest. The ability of wavelet transforms to break down the original spectrum into different resolution levels allows us to investigate the metabolite feature with different scales. The energy shift-insensitive property in the complex wavelet transform can efficiently handle misalignment and enables direct comparison among multiple NMR spectra. The FDR-based feature selection procedure treats all the wavelet coefficients simultaneously and systematically identifies important features in NMR spectra. The selected features greatly improve the classification accuracy when the kernel-induced random forest is used since the kernels contain more efficient information in the selected feature space.

References

1. Goodacre, R., York, E.V., Heald, J.K., Scott, I.M.: Phytochemistry 62, 859–863 (2003)
2. Tapp, H.S. , Defernez, M., Kemsley, E.K.: Journal of Agricultural And Food Chemistry 51 6110–6115 (2003)
3. Davis, R.A., Charlton, A.J., Oehlschlager, S., Wilson, J.C.: Chemometrics and Intelligent Laboratory Systems 81 50–59 (2006)
4. Barache, D., Antoine, J., Dereppe, J.: Journal of Magnetic Resonance 128 1–11(1997)
5. Gunther, U.L., Ludwig, C., Ruterjans, H.: Journal of Magnetic Resonance 156 19–25(2002)
6. Qu, Y., Adam, B.-L., Thornquist, M, Potter, J.D., Thompson, M.L., Yasui, Y., Davis, J., Schellhammer, P.F., Cazares, L., Clements, M., Write, G.L., Feng, Z.: Biometrics 59 143–151(2003)
7. Kim, S.B., Wang, Z., Oraintara, S., Temiyasathit, C., Wongsawat, Y.: Chemometrics and Intelligent Laboratory Systems 90 161–168(2008)
8. Breiman, L.: “Random forests,” Machine Learning. 45, 5-32(2001)
9. Breiman, L., Friedman, J., Olshen, R., and Stone, C.: Classification and Regression Trees, Belmont, CA: Wadsworth, (1984)
10. Fan, G.: "Kernel-Induced Classification Tree and Random Forest," Technical Report, Dept. of Statistics and Actuarial Science, University of Waterloo (2009)
11. Gabor, D.: Journal of Institution Electrical Engineering 429–457(1946)
12. Benjamini, Y., Hochberg, Y.: Journal of The Royal Statistical Society Series B. Methodological 57 289–300(1995)
13. Shaffer, J.P.: Annual Review of Psychology 46 561–584(1995)
14. Kim, S.B., Tsui, K-L., Borodovsky, M.: International Journal of Bioinformatics Research and Applications 2 193–217(2006)
15. Storey, J.D.: Annals of Statistics 31 2013–2035(2003)
16. Hastie, T., Tibshirani, R., Friedman, J.: The Element of Statistical Learning, Springer, New York (2001)
17. Lee, G.C., Woodruff, D.L.: Analytica Chimica Acta 513 413--416 (2004)