Classification of Proteomic Data with Logistic Kernel Partial Least Squares Algorithm

Zhenqiu Liu
Department of Statistics
The Ohio State University
1958 Neil Avenue
Columbus, OH 43210, USA

Dechang Chen
Department of Preventive Medicine and Biometrics
Uniformed Services University of the Health Sciences
4301 Jones Bridge Road
Bethesda, MD 20814, USA

Jianjun Tian
Mathematical Biosciences Institute
The Ohio State University
231 W. 18th Avenue
Columbus, OH 43210, USA

Abstract

In this paper we introduce the logistic kernel partial least squares (LKPLS) algorithm for classification of health vs. cancer using mass spectrometry (MS). Wavelet decomposition is proposed for feature selection and data preprocessing. LKPLS combines the logistic regression with the kernel partial least squares algorithm. The method is applied to real life cancer samples. Experimental comparisons show that LKPLS outperforms other methods in the analysis of MS data.

1 Introduction

Proteins carry out and modulate the vast majority of chemical reactions which together constitute 'life'. Proteomics is an integral part of the process of understanding biological systems and uncovering disease mechanisms. Because of their high level of variability and complexity, it is an extremely challenging endeavor to conduct massive analysis of thousands of proteins. In the last decade, mass spectrometry has increasingly become the method of choice for analysis of complex protein samples. Mass spectrometry measures two properties of ion mixtures in the gas phase under the vacuum environment: the mass-to-charge ratio (m/z) of ionized proteins in the mixture and the number of ions present at different m/z values. The output is a mass spectrum or chart with a series of spike peaks, each representing the ions of a specific m/z value in the sample. The heights of peaks and the m/z values of the peaks are a fingerprint of the sample. Mass spectrometry has not only been used intensively to identify proteins via peptide mass fingerprints, but also found promising applications in cancer classification [1, 3, 5, 8]. An important goal of cancer classification is to predict cancer on the basis of peptide/protein intensities.

While MS is increasingly used for protein profiles, significant challenges have arisen with regard to analyzing the data sets. The critical pre-precessing steps include baseline correction, peak identification and alignment, data normalization and visualization, and feature selection. The final and most important step is the classification of disease status with the selected features. Recent publications on cancer classification with MS data have mainly focused on how to identify features for classification and which classification method is more accurate than others. Particularly, T statistics and principal component analysis (PCA) have been used to select features. Classification methods such as linear discrimination analysis, k-nearest neighbor classification, decision trees [1], and support vector machines have been used to distinguish between cancer and normal samples [3, 8]. However, features provided by PCA do not necessarily yield good classification results. Experiments can be used to show that features derived from the partial least squares method (PLS) usually give more accurate predictions. Besides, some feature selection methods have been applied to the peaks instead of the original MS data, which can affect the performance of classification because of the peak finding algorithms employed.

In this paper we propose a novel analysis procedure LKPLS for classification of MS data. LKPLS combines the

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 IEEE

kernel partial least squares (KPLS) with logistic regression in a natural way. KPLS is a generalization and nonlinear version of PLS. LKPLS involves three steps: feature space transformation, dimension reduction, and classification. The proposed algorithm can not only predict the class label but also provide the probability of each sample falling into a specific class. Feature selection based on wavelet decomposition of original data has also been proposed in this paper. We assess the performance of the proposed algorithm and feature select method using two real life MS data sets.

This paper in organized as follows. In Section 2, we discuss the DWT feature selection method. LKPLS algorithm is given in Section 3. Computational results are described in Section 4. Conclusions and remarks are provided in Section 5.

2 Feature Selection Methods

A MS data set with n samples is a $p \times (n+1)$ matrix $(mz, X) = [mz, \mathbf{x}_1, \ldots, \mathbf{x}_n]$ where p is the number of m/z ratios, mz is a column vector for the measured m/z ratios, and $\mathbf{x}_j$ are the corresponding intensities of the jth sample. Let $\mathbf{y}' = [y_1, \ldots, y_n]$ denote the cancer status of the samples. Our goal is to predict the label y_j based on the intensity profile $\mathbf{x}_j$. As usual, such prediction often requires the task of feature selection. In the following, we propose one feature selection method based on the wavelet decomposition.

Wavelet Decomposition for Feature Selection

MS data have several special characteristics: the dimension of the data is large, the data points are not necessarily independent, and the measurements are usually more or less noisy. These problems motivate the use of compression techniques to describe the sequential data with a few features that capture the basic shape of the sequence. A wavelet transform is a way to decompose a signal in a chosen number of its constituent part. Fourier analysis also has this property but wavelet analysis has some advantages when analyzing signals of non-stationary nature. Wavelet provides more irregular shapes and the wavelet decomposition is a local one, so that if the information relevant to our prediction problem is constrained in a particular part or parts of the curve, as typically it is, this information will be carried in a very small number of wavelet coefficients. These properties make wavelets ideal for analyzing signals with discontinuities and sharp changes while allowing temporal locating the features of signal. Wavelets are families of functions that can accurately describe other functions in a parsimonious way. The signal is projected into the time frequency plane. The basis functions are $\Psi_{j,k}(t) = 2^{\frac{j}{2}}\Psi(2^j t - k)$, where Ψ is the mother wavelet function. Any square integrable real function $f(t)$ can be represented in terms of bases as

$$f(t) = \sum_{j,k} c_{j,k}\Psi_{j,k}(t),$$

where $c_{j,k} = \langle \Psi_{j,k}(t), f(t) \rangle$ are the coefficients of the discrete wavelet transform (DWT). There is a fast algorithm to get the coefficients with $O(n)$ time. A simple and commonly used wavelet is the Haar wavelet with the mother function

$$\Psi_{Haar}(t) = \begin{cases} 1, \text{ if } 0 < t < 0.5 \\ -1, \text{ if } 0.5 \leq t < 1 \\ 0, \text{ otherwise} \end{cases}$$

In this paper, we use Haar wavelet ([2]) because of its simplicity. One can use any other orthonormal wavelet basis functions and achieve similar results as we present here. In wavelet decomposition the signal is separated successively into slow and fast components using a pair of finite impulse response filters. In the first stage the high pass and the low pass filters separate the signal into components above $f_s/4$ and components below $f_s/4$. The second stage receives the low frequency components as input and separates them into components above $f_s/8$ and components below $f_s/8$, and so on. The number of stages depends the slowest component that is desired.

Each MS input $\mathbf{x}_j$ can be treated as an original signal for MS data. We can apply DWT to the input. Then the obtained coefficients can be used as the features [6].

Let the vector $\mathbf{z}_j$ represent the wavelet coefficients of $\mathbf{x}_j$. Let l denote the smooth part and h denote a detailed part of one decomposition step. Then in

- step 1: $\mathbf{x}_j = l_{1j} + h_{1j}$
- step 2: $\mathbf{x}_j = l_{2j} + h_{2j} + h_{1j}$
- Step i: $\mathbf{x}_j = l_{ij} + h_{ij} + \ldots + h_{1j}$

In this way each x_{ij} has a corresponding wavelet coefficient z_{ij}, $i = 1, \ldots, p$, i.e., $\mathbf{x}_j \rightarrow \mathbf{z}_j$.

Note the size of feature matrix selected with DWT method is still around $p \times n$, but many entries are zeros or near zeros. Traditionally, some heuristic rules can be applied to reduce the feature dimension. For instance, we can either keep the first few coefficients or the largest coefficients of DWT as the features. However, both methods are just based on the signal itself and do not consider the associated class information. Our proposed LKPLS algorithm combines the feature selection and classification together and choose features not only based on the data but also the class labels.

Features can also be selected from the original MS data with the two way T test statistic. By assuming each m/z

value is independent, T test selects the features by the following formula:

$$c_j = \frac{|\mu_j^+ - \mu_j^-|}{\sqrt{\frac{(\sigma_j^+)^2}{n^+} + \frac{(\sigma_j^-)^2}{n^-}}},$$

where μ_j^+ and μ_j^- are the mean values of the jth feature over positive and negative samples, respectively, σ_j^+ and σ_j^- are the corresponding standard deviations, n^+ and n^- are the number of positive and negative training samples, respectively. We can then choose the features with significant p values.

3 LKPLS Algorithm

Given a training dataset $\{\mathbf{z}_i\}_{i=1}^n$ with class labels $\{y_i\}_{i=1}^n$ and a test feature dataset $\{\mathbf{z}_t\}_{t=1}^{n_t}$ with labels $\{y_t\}_{t=1}^{n_t}$, the algorithm LKPLS is stated as follows:

1. Compute the kernel matrix, for the training data, $K = [K_{ij}]_{n\times n}$, where $K_{ij} = K(\mathbf{z}_i, \mathbf{z}_j)$. Compute the kernel matrix, for the test data, $K_{te} = [K_{ti}]_{n_t\times n}$, where $K_{ti} = K(\mathbf{z}_t, \mathbf{z}_i)$.
2. Call KPLS algorithm to find k component directions [7]:
   (a) for $i = 1, \ldots, k$
   (b) initialize $\mathbf{u}^i$
   (c) $\mathbf{t}^i = \Phi\Phi'\mathbf{u}^i = K\mathbf{u}^i$, $\mathbf{t}^i \leftarrow \mathbf{t}^i/||\mathbf{t}^i||$
   (d) $\mathbf{c}^i = \mathbf{y^i t^i}$
   (e) $\mathbf{u^i} = \mathbf{yc^i}$, $\mathbf{u^i} \leftarrow \mathbf{u^i}/||\mathbf{u^i}||$
   (f) repeat steps (b) -(e) until convergence
   (g) deflate $K, \mathbf{y}$ by $K \leftarrow (I - \mathbf{t^i t^{i\prime}})K(I - \mathbf{t^i t^{i\prime}})$ and $\mathbf{y} \leftarrow \mathbf{y} - \mathbf{t}^i\mathbf{t}^{i\prime}\mathbf{y}$
   (h) obtain component matrix $U = [\mathbf{u^1}, \ldots, \mathbf{u^k}]$
3. Find the projections $\mathbf{V} = KU$ and $\mathbf{V}_{te} = K_{te}U$ for the training and test data, respectively.
4. Build a logistic regression model using $\mathbf{V}$ and $\{y_i\}_{i=1}^n$ and test the model performance using $\mathbf{V}_{te}$ and $\{y_t\}_{t=1}^{n_t}$.

The following are among the popular kernel functions:

- First norm exponential kernel

$$K(\mathbf{z}_i, \mathbf{z}_j) = \exp(-\beta||\mathbf{z}_i - \mathbf{z}_j||)$$

- Radial basis function kernel (RBF)

$$K(\mathbf{z}_i, \mathbf{z}_j) = \exp\Big(-\frac{|\mathbf{z}_i - \mathbf{z}_j|^2}{\sigma^2}\Big)$$

- Power exponential kernel (a generalization of RBF kernel)

$$K(\mathbf{z}_i, \mathbf{z}_j) = \exp\Big[-\Big(\frac{|\mathbf{z}_i - \mathbf{z}_j|^2}{r^2}\Big)^\beta\Big]$$

- Sigmoid kernel

$$K(\mathbf{z}_i, \mathbf{z}_j) = \tanh(\beta\mathbf{z}_i'\mathbf{z}_j)$$

- Polynomial kernel

$$K(\mathbf{z}_i, \mathbf{z}_j) = (\mathbf{z}_i'\mathbf{z}_j + p_2)^{p_1}$$

- Linear kernel

$$K(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i'\mathbf{z}_j$$

We can show that the above LKPLS classification algorithm is a nonlinear version of the logistic regression. In fact, it follows from our KPLS classification algorithm that the probability of the label y given the projection $\mathbf{v}$ is expressed as

$$P(y|\mathbf{w}, \mathbf{v}) = g\Big(b + \sum_{i=1}^k w_i v_i\Big), \tag{1}$$

where the coefficients $\mathbf{w}$ are adjustable parameters and g is the logistic function

$$g(u) = (1 + \exp(-u))^{-1}.$$

Given a data point $\Phi(\mathbf{z})$ in the transformed feature space, its projection v_i $(i = 1, ..., k)$ can be written as

$$v_i = \Phi(\mathbf{z})\Phi'\mathbf{u^i} = \sum_{j=1}^n u_j^i K(\mathbf{z}_j, \mathbf{z})$$

Therefore, from equation (1), we have

$$P(y|\mathbf{w}, \mathbf{v}) = g\Big(b + \sum_{j=1}^N c_j K(\mathbf{z}_j, \mathbf{z})\Big), \tag{2}$$

where

$$c_j = \sum_{i=1}^k w_i u_j^i, \;\; j = 1, \cdots, n.$$

When $K(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i'\mathbf{z}_j$, equation (2) becomes a logistic regression. Therefore, LKPLS classification algorithm is a generalization of logistic regression.

Similar to support vector machines, LKPLS algorithm was originally designed for two class classification. However, we can deal with the multi-class classification problem through the popular 'one against all others' scheme. What we do is that we perform classification for all the two-class problems and then send each sequence to the class with the highest probability.

4 Results

Ovarian Cancer

First we evaluate the performance of the proposed algorithm on the ovarian cancer data. This cancer dataset was downloaded directly from the web site: http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp. The sample set includes 91 controls and 162 ovarian cancer cases. To evaluate the performance of the algorithm, we merged the control and cancer data together and split the the data with a ten-fold validation scheme. The data were divided randomly into ten roughly equal subsets, and then we applied the algorithm 10 times, each time with 9 subsets used for training and the remaining subset for performance evaluation. In this example, linear kernels were used. The averaged error over the 10 times was reported as an overall performance. The output is given in Table 1. This table clearly shows that LKPLS performs better with DWT feature selection methods. The number of features in Table 1 is the number of components used in the LKPLS algorithm.

Table 1. Performance of LKPLS on ovarian cancer data with different feature selection methods

	T-test	DWT
No. of Features	17	10
Test Error (%)	1.75 ± 1.4	0 ± 0
Sensitivity(%)	99.03 ± 1.48	100 ± 0
Specificity (%)	96.64 ± 2.19	100 ± 0

Prostate Cancer

The prostate cancer data were downloaded from the same web site as the ovarian data. The Surface Enhanced Laser Desorption/Ionization time of flight (SELDI) method and a mass spectra analysis for this data set have been performed in [5]. SELDI process is a relatively new medical technique that measures the content of different proteins in blood samples from patients. This dataset consists of four subsets:

1. 63 samples with no evidence of disease and the prostate specific antigen (PSA) level less than 1(ng/ml).
2. 190 samples with benign prostate and PSA level greater than 4.
3. 26 samples with prostate cancer and PSA levels between 4 and 10.
4. 43 samples with prostate cancer and PSA levels greater than 4.

There are 322 samples in total and we treated them as coming from 3 classes: normal, benign, and cancer. Again, the ten-fold validation was used for the experiments. The 'One against all others' scheme was applied to separate each class against the other two. The performance of LKPLS algorithm and the comparison of LKPLS with Fisher linear discrimination (LD), k nearest neighbor (KNN), and neural networks are given in Table 2 and Table 3, respectively.

Table 2. Performance of LKPLS on prostate cancer data with different feature selection methods

	T-test	DWT
No. of Features	53	28
Test Error (%)	11.8 ± 2.8	1.7 ± 1.4
Sensitivity(%)	87.1 ± 2.94	98.5 ± 1.3
Specificity healthy (%)	96.8 ± 2.69	100 ± 0
specificity benign (%)	83.1 ± 1.72	94.7 ± 3.58

Table 3. Performance comparison of different classification methods

Feature Selection & Classification	Error (%)
PCA & LD	95.3
PCA & Logistic Regression	96.8
PCA & KNN	91.9
PCA & Neural Network (15 nodes)	83.5
KPLS (2nd order poly.) & LD	96.3
KPLS (2nd order poly.) & KNN	93.6
KPLS (2nd order poly.) & Neural Network	85.9
LKPLS (2nd order poly.)	98.3

Table 2 and Table 3 show that the best overall classification performance (98.3%) is achieved by LKPLS. Logistic regression seems to perform better than LD, KNN, and neural network for this example. It is expected that logistic regression normally performs better than linear discrimination when data have outliers, as there is no normality assumption with logistic regression. DWT can be used as a data preprocessing and feature extraction step to achieve high prediction accuracies. Results also indicate that neural networks do not seem to be very efficient for this type of application, at least for this special example. One reason might be that the number of parameters needed to build an accurate model is too large in comparison to the number of available examples.

5 Conclusion

Our limited experiments with two cancer data sets show that the proposed algorithm LKPLS is promising. Mass spectrometry measure together with feature selection, dimension reduction, and classification algorithms could be an effective way to make a highly reliable prognostication of patients possibly suffering from cancer. The preprocessing of MS output is a very crucial step in the overall analysis of MS data. Our proposed DWT data preprocessing step seems to work well for the two datasets. Feature selection and dimension reduction constitute another important step in the analysis. Features selected with the kernel partial least squares algorithm performed better than those selected with either PCA or T-test. This is reasonable since PCA selects features to explain as much information (variance) in the input as possible, while LKPLS strikes a compromise between explaining variance in input and finding correlation with the corresponding labels. The proposed algorithm is easy to be implemented and has a high prediction accuracy. Finally, we note that LKPLS can also be applied to multi-class microarray data classificationcon [4].

Acknowledgements

D. Chen was supported by the National Science Foundation grant CCR-0311252.

Note: The opinions expressed herein are those of the authors and do not necessarily represent those of the Uniformed Services University of the Health Sciences and the Department of Defense.

References

[1] Adam, B-L., Qu, Y., Davis, J.W., Ward, M.D., Clements, M.A., et al. Serum protein figerprinting cupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.* 62:3609-3614, 2002.

[2] Burrus, C. S., Gopinath, R. A., and Guo, H. *Introduction to Wavelets and Wavelet Transforms: A Primer*. Prentice-Hall, Inc., 1998.

[3] Lilien, H., Farid, H., and Donald, B. R. Probabilistic Disease Classification of Expression-Dependent Proteomic Data from Mass Spectrometry of Human Serum. *Journal of Computational Biology*, 10(6):925-946, 2003.

[4] Liu, Z., and D. Chen. Gene Expression Data Classification with Revised Kernel Partial Least Squares Algorithm. *Proceedings of the 17th International FLAIRS conference*, 104-108, 2004.

[5] Petricoin, E. F. III, Ornstein, D. K., Paweletz, C. P., Ardekani, A., Hackett, P. S., Hitt, B. A. , et al. Serum proteomic patterns for detection of prostate cancer. *J Natl Cancer Inst*, 94:15768, 2002.

[6] Qu, Y., Adam, B. L., Thornquist, M., Potter, J. D., Thompson, M. L., Yasui, Y., et al. Data Reduction Using a Discrete Wavelet Transform in Discriminant Analysis of Very High Dimensionality Data. *Biometrics*, 59:143-151, 2003.

[7] Rosipal, R. and Trejo, L. J. Kernel partial least squares regression in RKHS, Theory and empirical comparison. Technical report, University of Paisley, UK, 2001.

[8] Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., et al. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19(13): 1636-1643, 2003.