Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2008) 2578–2587
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Classification of gene functions using support vector machine for time-course gene expression data

Changyi Park[a,*], Ja-Yong Koo[b], Sujong Kim[c], Insuk Sohn[b], Jae Won Lee[b]

[a] *Institute of Statistics, Korea University, Seoul 136-701, Republic of Korea*
[b] *Department of Statistics, Korea University, Seoul 136-701, Republic of Korea*
[c] *Department of Biochemistry, Hanyang University, Seoul 133-791, Republic of Korea*

Received 23 August 2006; received in revised form 9 August 2007; accepted 6 September 2007
Available online 15 September 2007

Abstract

Since most biological systems are developmental and dynamic, time-course gene expression profiles provide an important characterization of gene functions. Assigning functions for genes with unknown functions based on time-course gene expressions is an important task in functional genomics. Recently, various methods have been proposed for the classification of gene functions based on time-course gene expression data. In this paper, we consider the classification of gene functions from functional data analysis viewpoint, where a functional support vector machine is adopted. The functional support vector machine can model temporal effects of time-course gene expression data by incorporating the coefficients as well as the basis matrix obtained from a finite expansion of gene expressions on a set of basis functions. We apply the functional support vector machine to both real microarray and simulated data. Our results indicate that the functional support vector machine is effective in discriminating gene functions of time-course gene expressions with predefined functions. The method also provides valuable functional information about interactions between genes and allows the assignment of new functions to genes with unknown functions.

Keywords: B-spline basis; Fourier basis; Functional data classification; Gene function

1. Introduction

Most biological systems are complex and dynamic because they usually consist of multiple molecular pathways with various biological functions. The microarray technique allows us to obtain genome-wide transcriptional profiles at various stages and under different conditions on the biological systems. Using global gene expression data, we are able to identify a set of co-regulated genes and annotate specific biological functions to genes with unknown functions. A common method used in functional analysis of gene expression data is clustering. Clustering is based on the assumption that genes with the same functions will have similar expression profiles over a range of experimental conditions. By taking advantage of the knowledge already established by biologists, it is possible to predefine functional classes of genes and to assign unknown genes to the functional classes or biological pathways.

* Corresponding author. Tel.: +82 2 32901640; fax: +82 2 9249895.
E-mail addresses: park463@korea.ac.kr (C. Park), jykoo@korea.ac.kr (J.-Y. Koo), sundance@amorepacific.com (S. Kim), sis46@korea.ac.kr (I. Sohn), jael@korea.ac.kr (J.W. Lee).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.09.002

There have been increasing interests in statistical methods, particularly in clustering and classification techniques, for time-course gene expression data. Recent works include a clustering method based on a mixed-effects model with B-splines (Luan and Li, 2003), a mixture functional discriminant analysis (MFDA) based on Gaussian mixture models (Gui and Li, 2003; Chudova et al., 2004), a clustering algorithm designed for short time series gene expression data (Ernst et al., 2005), a clustering algorithm based on a dissimilarity measure proposed by Heckman and Zamar (2000) for curve clustering (Bensmail et al., 2005), and a logistic regression based on functional data analysis (Leng and Müller, 2006).

In extending support vector machines (SVM, Cortes and Vapnik, 1995) to deal with functional data, there have been two approaches. One is to pre-smooth the data, obtain an orthonormal representation using functional principal component analysis (FPCA), and apply standard SVM to the coefficients (Lee, 2004). The other is to pre-smooth the data using a basis system and apply the SVM designed for functional data, called FSVM (Rossi and Villa, 2005, 2006). The latter reduces to the former if the adopted basis system is orthonormal.

Leng and Müller (2006) classified time-course gene profiles using logistic regression with the coefficients obtained by FPCA. Henceforth, we call this method as the FPCA logistic regression. FPCA is a valuable tool for exploratory analysis of functional data and has many good features such as dimension reduction and orthogonality of FPCA scores. FPCA can also be adopted in our context. However, we do not follow that direction due to the following reasons. First, correct classification is more important than dimension reduction in gene classification. Second, we may lose the interpretability of data during the process of the classification using FPCA, which consists of three steps: smoothing data, applying FPCA to smoothed data, and classification using the resulting FPCA scores. FSVM is not limited to orthonormal basis systems and can be implemented in two steps: smoothing data and classification using the smoothed data, so that FSVM provides a more transparent and simpler way of classifying functional data.

We study the classification of time-course gene expressions from functional data analysis (see e.g., Ramsay and Silverman, 1997) viewpoint, where each time-course gene profile can be seen as a curve observed on some discrete time points, possibly non-uniform time points due to experimental conditions or missing observations. In this situation, applying multivariate classification methods to the original data may not be desirable because these methods ignore potential temporal effects. The logistic regression proposed in Leng and Müller (2006) has the advantage of estimating conditional class probabilities. However, their method may have a room for improvements in its predictive performance. FSVM can be effective in discriminating time-course gene profiles. In this paper, we illustrate the performance of FSVM through real data analysis and a simulation study in discriminating sequentially observed gene profiles. To the best of our knowledge, this paper is the first study applying FSVM in the classification of time-course gene expressions.

This paper is organized as follows. Section 2 introduces FSVM designed for functional data classification. Section 3 compares the performance of FSVM and other classification methods using both real and simulated data sets. Some concluding remarks are given in Section 4.

2. Support vector machine for functional data

2.1. Support vector machine

Let (X, Y) be a pair of random variables with $X \in \mathbb{R}^d$ and $Y \in \{+1, -1\}$. Suppose that $\{(x_i, y_i)\}_{i=1}^n$ is a set of training data, independently drawn from the distribution of (X, Y). For presentational purpose, we consider only a linear classification problem. Denote an input vector as $x = (x_1, \ldots, x_d)'$ and a coefficient vector as $w = (w_1, \ldots, w_d)' \in \mathbb{R}^d$. The bias term is denoted by $b \in \mathbb{R}$.

For separable cases, SVM maximizes the geometric margin, or minimizes $2/\|w\|^2$, with respect to linear hyperplanes subject to the constraints $y_i(\langle w, x_i\rangle + b) \geqslant 1, i = 1, \ldots, n$, where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{R}^d$. For non-separable cases, a soft-margin SVM is introduced to minimize the objective function

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \tag{1}$$

subject to the constraints $y_i(\langle w, x_i\rangle + b) \geqslant 1 - \xi_i$ and $\xi_i \geqslant 0$ for $i = 1, \ldots, n$, where $C > 0$ is a penalization parameter and $\{\xi_i\}_{i=1}^n$ are called the slack variables. The dual form for the above primal optimization problem (1) is to

minimize

$$\sum_{i=1}^{n} \alpha_i - \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \tag{2}$$

subject to $\sum_{i=1}^{n} \alpha_i y_i = 0$ and $0 \leqslant \alpha_i \leqslant C$ for $i = 1, \ldots, n$. Let $\hat{\alpha}_i$, $i = 1, \ldots, n$, be the solution of (2). The solution function is

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i y_i \langle x_i, x \rangle + \hat{b},$$

where $\hat{b}$ is determined via the Karush–Kuhn–Tucker boundary conditions. If $\hat{f}(x) \geqslant 0$ for an instance x, then the class label of x is predicted as $+1$. Otherwise, the class label is predicted as -1. Since $\hat{w}$ can be represented in terms of non-zero $\hat{\alpha}_i$'s alone, those training data points with non-zero $\hat{\alpha}_i$'s are called the support vectors. Nonlinear classifications can be implemented using nonlinear kernels such as a radial basis function (RBF) kernel. See Vapnik (1996) for more details.

2.2. Functional support vector machine

Consider square integrable functions x_i defined on $L^2[0, T]$ for $i = 1, \ldots, n$. Let $\mathbf{x}_i = (x_i(ta_{i,1}), \ldots, x_i(t_{i,m_i}))'$ be the vector of observations for ith curve observed at time points $t_{i,j}$ for $j = 1, \ldots, m_i$ and $i = 1, \ldots, n$. Note that time points may not be equally spaced. Denote $y_i \in \{-1, +1\}$, the state of ith curve. Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ be the training set. Suppose that each x_i can be expressed as $x_i = \sum_{l=1}^{\infty} c_{i,l} \phi_l$ for a given basis system $\{\phi_l\}_{l=1}^{\infty}$ on $L^2[0, T]$. For some finite integer L, we approximate x_i by $\hat{x}_i = \sum_{l=1}^{L} c_{i,l} \phi_l$, where L can be determined so that the misclassification error estimated by cross validation (CV) is minimized. Let $\mathbf{c}_i = (c_{i,1}, \ldots, c_{i,L})' \in \mathbb{R}^L$ for $i = 1, \ldots, n$.

A linear kernel for FSVM is defined as

$$K(x_i, x_j) = \langle \hat{x}_i, \hat{x}_j \rangle = \mathbf{c}_i' \Phi \mathbf{c}_j,$$

where $\Phi = (\int_0^T \phi_i(t) \phi_j(t)\, \mathrm{d}t)_{i,j=1,\ldots,L}$ is a basis matrix. For orthonormal bases, FSVM with the linear kernel reduces to a standard SVM using the coefficients $\mathbf{c}_i$'s because Φ is an identity matrix. In general cases, the basis matrix Φ can be obtained through numerical integration. The rest is essentially the same as standard SVM, i.e., to minimize

$$\sum_{i=1}^{n} \alpha_i - \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{3}$$

subject to $\sum_{i=1}^{n} \alpha_i y_i = 0$ and $0 \leqslant \alpha_i \leqslant C$ for $i = 1, \ldots, n$. Denote the solution of (3) as $\hat{\alpha}_i$'s. The solution function is defined as

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i y_i K(x_i, x) + \hat{b},$$

where $\hat{b}$ is determined via the Karush–Kuhn–Tucker boundary conditions. Given an instance $\mathbf{x}$ with the corresponding curve x, its class label is predicted as $\mathrm{sign}(\hat{f}(x))$.

One may use nonlinear kernels such as the RBF kernel as in Rossi and Villa (2006). We consider the linear kernel alone because the use of the RBF kernel may result in over-smoothing. We investigate the effect of different basis systems, especially Fourier and B-spline bases. Ramsay and Silverman (1997) provide general guidelines for the choice of an appropriate basis system. A Fourier basis system fits data exhibiting periodicity with simple temporal patterns. In the presence of local features, splines seem to be a better choice. Although Fourier basis systems may be useful for data with simple temporal patterns, B-splines are generally expected to work better than Fourier bases for a wide range of data.

3. Data analysis

In this section, we compare the performances of linear FSVM, logistic regression, FPCA logistic regression, and MFDA. For linear FSVM and logistic regression, we compare the performances of Fourier basis and cubic B-splines with equally spaced knots. To investigate the functional nature of data, we also applied standard SVM with the RBF kernel to original data, where gene expression levels at each time point are regarded as observations of a variable.

The logistic regression, using the `multinom` function from the `mass` package in `R`, is performed on coefficients $\mathbf{c}_i$, $i = 1, \ldots, n$. The basis matrix for FSVM can be obtained by the `inprod` function provided in the `fda` package of `R`. To save computing time, we can store the basis matrix Φ for each fixed L and solve the dual form (3) incorporated with the stored Φ.

The number of basis functions L was determined by misclassification error rates estimated by leave-one-out CV (LOO-CV) under the restriction that $L \leqslant \min_{i=1,\ldots,n} m_i$. To determine the tuning parameter C for linear FSVM, and the tuning parameter C and the scaling parameter for standard SVM with the RBF kernel, we have used grid search on $2^{-5}, 2^{-4}, \ldots, 2^5$ using LOO-CV.

3.1. Dictyostelium data

Iranfar et al. (2001) investigated gene expression patterns of cell-type-specific genes in Dictyostelium. Our training data set consists of 35 genes with 14 prestalk and 21 prespore type genes, observed at 11 time points, whose temporal patterns are shown in Fig. 1(a). Note that prestalk type genes have their peaks earlier in the observation period while peaks are observed later for prespore type genes. Although a linear B-splines may be sufficient for this data set due to the small number of time points, we used cubic B-splines for consistency with the other numerical examples.

Table 1 shows misclassification error rates estimated by LOO-CV. The number of basis functions and tuning parameters have been determined so that LOO-CV error rate is minimized. FSVM with B-splines yielded the smallest error rate, but the gain seems to be marginal.

To compare relative performances of the methods further, we randomly partitioned 35 genes into 25 training and 10 test data. This process of random partitioning was repeated 200 times. The results are summarized in Table 2. In terms of test error rates, FSVM with B-splines performed the best. Since standard SVM was outperformed by the other methods, there might be some functional nature in this data set. P-value in Table 2 has been obtained from one-sided Wilcoxon test on the classification results for each method against those from FSVM with B-splines.

3.2. Yeast cell cycle data

Eisen et al. (1998) analyzed 2467 genes based on their time-course gene expressions measured over 14 time points during the cell cycle process. Among these genes, we analyzed 264 genes assigned to one of the five functional classes; tricarboxylic acid (TCA) with 10 genes, respiration (Resp) with 6 genes, cytoplasmic ribosomes (Ribo) with 189 genes, proteasome (Prot) with 40 genes and histone (His) with 9 genes. The functional classes were determined based on the Comprehensive Yeast Genome Database (CYGD) (Mewes et al., 2002). Fig. 1(b) shows temporal patterns of those genes; gene expression levels for TCA increase slowly over the time period, Resp genes decay slowly over time, expression levels for Ribo genes increase up to time 2 and then decrease slowly after that, Prot genes have similar patterns as Resp genes, and Hist genes have similar patterns, but with larger amplitude, as Ribo genes. Since there are some local features in their temporal patterns, B-splines would be more appropriate than a Fourier basis.

Note that Leng and Müller (2006) have used 90 genes whose cell cycles were identified by Spellman et al. (1998) while we have used 264 genes whose gene function classes were identified by Eisen et al. (1998). Both studies analyzed different subsets of the yeast cell cycle data. A reason why we have analyzed those 264 genes with different functional classes is that the classification of gene functions seems to be biologically more meaningful than the classification of cell cycles.

The classification of these 264 genes is a multiclass classification problem. Most popular approaches to multiclass SVM are one-against-all and one-against-one approaches. Although one-against-all approach has an advantage of faster computation than one-against-one approach, its predictive performance is generally worse than that of one-against-one approach (Hsu and Lin, 2002). So we adopted one-against-one approach to implement the multiclass classification for both FSVM and standard SVM.

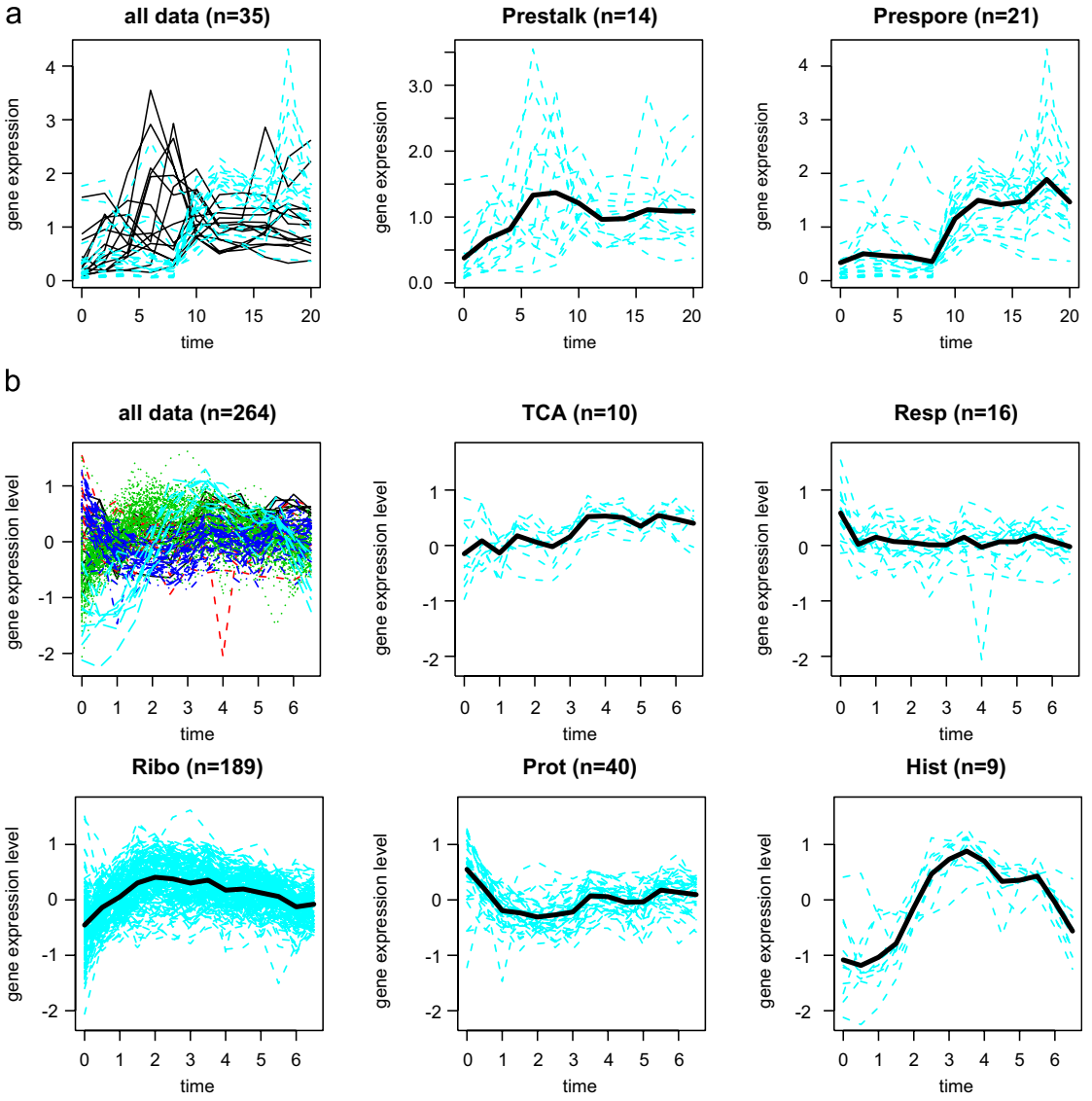

Fig. 1. Temporal patterns of microarray data, (a) Dictyostelium data and (b) Yeast cell cycle data. The first panel provides the overall gene profiles. The other panels show gene profiles for each class. The thick solid line indicates the mean curve for each group.

Table 3 summarizes the results. As expected, B-spline basis system outperformed a Fourier basis system. The performance of FSVM with B-splines was slightly better than those of other methods except for MFDA. FSVM with B-splines seems to be promising. The functional nature of data seems to be rather weak because the standard SVM performed well on this data set.

We did not consider random partitioning for the comparison of these methods due to the following reasons. First, the classification is a multiclass problem. Second, there is a serious class imbalance problem in this data set. Note that there are 10, 16, 189, 40, and 9 genes in TCA, Resp, Ribo, Prot, and His groups, respectively. In addition to the effects of the choice of classification methods, there may be other effects due to those problems.

3.3. Simulation study

We compared the predictive performance of linear FSVM with those of other methods based on three different simulation scenarios. In the first scenario, classes have different amplitudes. In biology, genes with expression level

Table 1
Misclassification error rates estimated by LOO-CV for Dictyostelium data

Method	Bases	Error rate
Standard SVM		0.2571
MFDA	5	0.2000
FPCA logistic	5	0.2285
Fourier logistic	7	0.2571
B-spline logistic	8	0.2285
Fourier FSVM	5	0.1428
B-spline FSVM	7	0.1142

Table 2
Average error rates and *P*-value from one-sided Wilcoxon test against B-spline FSVM for Dictyostelium data over 200 random partitions

Method	Average error rate	Standard error	*P*-value
Standard SVM	0.3250	0.0087	0.0001
MFDA	0.2880	0.0150	0.9810
FPCA logistic	0.2940	0.0078	0.0703
Fourier logistic	0.2920	0.0080	0.1500
B-spline logistic	0.2890	0.0094	0.2839
Fourier FSVM	0.2960	0.0079	0.0983
B-spline FSVM	0.2815	0.0086	

Table 3
Misclassification error rates estimated by LOO-CV for yeast cell cycle data

Method	Bases	Error rate
Standard SVM		0.1060
MFDA	6	0.1666
FPCA logistic	7	0.1136
Fourier logistic	3	0.1287
B-spline logistic	3	0.1174
Fourier FSVM	7	0.1136
B-spline FSVM	13	0.0984

over a certain threshold is considered to be significant (DeRisi et al., 1997). The second scenario has two classes with different phases. The purpose of this scenario is to see which classifica'tion method is effective in discriminating periodically expressed genes during the cell-cycle. In the third scenario, two classes follow the same patterns. However, one class has local features at fixed times points. This makes the discrimination of two classes more difficult. See Fig. 2 for temporal patterns of these scenarios. Since generated patterns are very simple for the first two scenarios, we expect that a small number of Fourier basis functions will be sufficient. The third scenario has significant local patterns, for which B-splines would be a reasonable choice.

We have generated 100 data sets according to the following simulation schemes. Each data set is composed of 50 training and 100 test curves. The following are descriptions of the simulation schemes for three scenarios.

- Simulation 1: amplitude variation
  This simulation is a modification of the simulation in Biau et al. (2005). Let the observation time points are $t = 0, \frac{1}{10}, \ldots, \frac{10}{10}$. For each $i \leqslant 50$,

$$x_i(t) = u_i \exp(-u_i t) + \varepsilon_i(t),$$

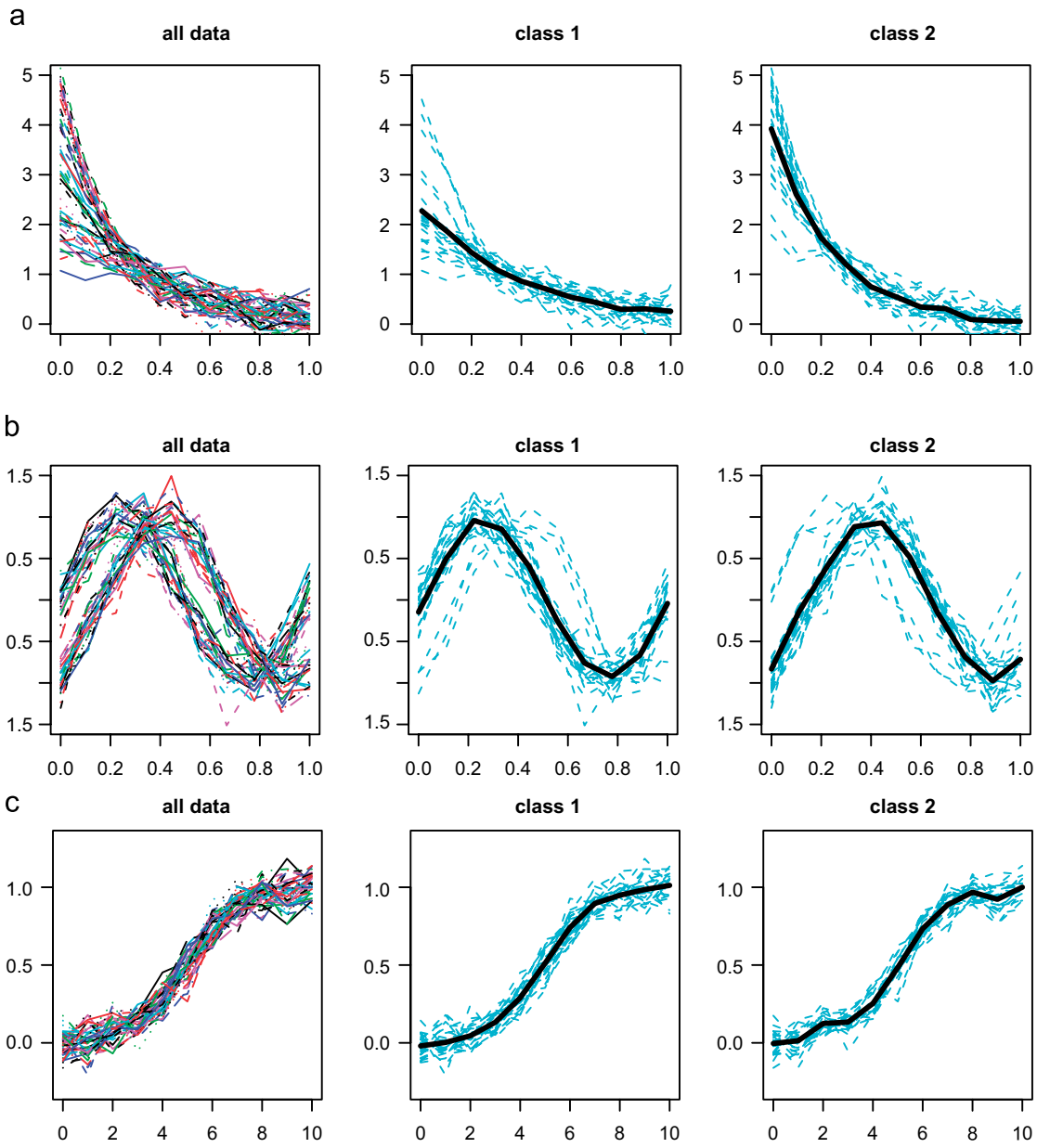

Fig. 2. Temporal patterns of simulated data sets with amplitude variation (a), phase variation (b), and local features (c). The first panel provides overall curve, and the second and third panels show curves for classes 1 and 2, respectively. The thick solid line indicates the mean curve for each class.

where u_i is a uniform random number on $[1, 5]$, $t \in [0, 1]$, and $\varepsilon_i(t)$ is a normal random variable with mean 0 and variance 0.04 for given t. y_i is generated from Bernoulli(0.2) distribution if $u_i < 3$ and from the Bernoulli(0.8) distribution otherwise. Without the noise term $\varepsilon_i(t)$, the Bayes error rate would be 0.2 as indicated in Biau et al. (2005). The optimal error rate in this simulation study is larger than 0.2.

- Simulation 2: phase variation
  Let $\varepsilon_i(t) \sim \mathrm{N}(0, 0.2)$ for $t = 0, \frac{1}{10}, \ldots, \frac{10}{10}$ and $i = 1, \ldots, 50$. For $i \leqslant 25$,

$$x_i(t) = \sin(2\pi t) + \varepsilon_i(t)$$

Table 4
Average error rate, standard error, and *P*-value from one-side Wilcoxon test against the best classifier for three simulation schemes

Method	Average error rate	Standard error	*P*-value
(a) Amplitude variation			
Standard SVM	0.2914	0.0110	0.6682
MFDA	0.3031	0.0059	0.0000
FPCA logistic	0.3074	0.0058	0.0001
Fourier logistic	0.2921	0.0057	0.0052
B-spline logistic	0.3419	0.0055	0.0000
Fourier FSVM*	0.2755	0.0068	
B-spline FSVM	0.2808	0.0067	0.2408
(b) Phase variation			
Standard SVM	0.1297	0.0099	0.0065
MFDA	0.1321	0.0054	0.0053
FPCA logistic	0.1080	0.0034	0.0000
Fourier logistic	0.1086	0.0036	0.0083
B-spline logistic	0.1855	0.0056	0.0000
Fourier FSVM*	0.0963	0.0031	
B-spline FSVM	0.1018	0.0035	0.1671
(c) Local feature			
Standard SVM	0.3336	0.0086	0.0000
MFDA	0.3193	0.0063	0.0005
FPCA logistic	0.3213	0.0059	0.0004
Fourier logistic	0.3319	0.0060	0.0076
B-spline logistic	0.3047	0.0065	0.0009
Fourier FSVM	0.3227	0.0057	0.0298
B-spline FSVM*	0.2857	0.0058	

and y_i is generated from Bernoulli (0.9) distribution; for $25 < i \leqslant 50$,

$$x_i(t) = \sin(2\pi(t - 1/6)) + \varepsilon_i(t)$$

and y_i is a random variate from Bernoulli (0.1) distribution. In this case, the Bayes error rate is larger than 0.1.

- Simulation 3: local feature
  We have modified the simulation scheme in James and Sood (2006) slightly. Let the observation time points be $t = 0, 1, \ldots, 10$. For $i \leqslant 25$, $y_i = 1$ and x_i is a logistic type curve defined as

$$x_i(t) = \frac{1}{1 + \exp(-t + 5)} + \varepsilon_i(t),$$

where $\varepsilon_i(t)$ have an independent $N(0, 0.08^2)$ distribution. For $25 < i \leqslant 50$, $y_i = -1$ and x_i is a logistic type curve with blips at time points 1 and 9, defined as

$$x_i(t) = \frac{1}{1 + \exp(-t + 5)} + \frac{1}{10} I(t = 1) - \frac{1}{20} I(t = 9) + \varepsilon_i(t).$$

Table 4 summarizes the results for the simulations. For the first two scenarios, logistic regression seems to favor orthogonal (Fourier and FPCA) basis systems. The reason appears to be that the logistic regression is applied directly to the coefficients. For non-orthonormal basis systems such as B-splines, logistic regression may not be able to incorporate temporal effects properly, so that the performance of logistic regression degraded for B-splines.

In contrast, FSVM seems to be less sensitive to the choice of a basis system than logistic regression. For the first two simulations, a Fourier basis seems to be slightly better than B-splines for FSVM. A possible explanation is that their temporal patterns are pretty simple and there is no local feature. For the third simulation with local features, B-splines seem to be better than a Fourier basis both for FSVM and logistic regression.

Now let us compare standard SVM with the other methods. For the first simulation, the performances of all the methods were similar. The second simulation is a little bit more complicated. The performance of standard SVM was worse than those of logistic regression with orthogonal basis systems and FSVM. The relative difference seems to become larger for the third simulation.

4. Discussion

Most biological processes such as development, growth and responses to environmental changes are time dependent. It is important to extract valuable functional information from temporal gene expression data for elucidating the molecular mechanism underlying those processes. Such efforts seem to be challenging, because many genes are involved in more than one biological pathways and genes with the same biological function may have different time-course expression profiles.

In this paper, we compared the predictive performance of FSVM with those of other competing methods through both real and simulated data analysis. For real microarray data, misclassification error rates estimated by LOO-CV for FSVM were lower than those for the state-of-art classification method, the logistic regression, in time-course microarray data analysis. Although we have encouraging results, the gain in the predictive performance for real data sets was not so evident. We compared those methods using simulations under several different scenarios: amplitude variation, phase variation, and local feature. For these scenarios, FSVM performed the best.

Moreover, the method provided valuable functional information about interactions between genes and assigned new functions to genes. For example, SIS1 (YNL007C) was assigned in ribosomes (Ribo), based on the Comprehensive Yeast Genome Database (CYGD) (Mewes et al., 2002); however, FSVM classified this gene into the proteasome (Prot) category. Several studies have shown that SIS1 is a member of a pair of chaperon specially involved in efficient protein turnover in the yeast, whose overexpression suppressed the growth defects caused by the proteasome mutations (Ohba, 1997). We believe FSVM can be useful in assigning new functions to genes with unknown functions based on time-course expression profiles, which may lead to important scientific discoveries.

In addition, we investigated the effects of different basis systems, B-spline and Fourier, because the quality of classification seems to be mainly dependent on the choice of a basis system. If temporal patterns are simple and periodic as in the first and second simulations, then a Fourier basis appears to be sufficient. Since non-negligible local features can be present in many cases, B-splines may be preferable on a variety of applications as illustrated in the third simulation. Alternatively, one might want to consider wavelets, because they are known to be powerful for detecting local features. When time points for gene experiments are not equally spaced and the number of time points is small, there may be some problems in applying wavelet basis functions.

Finally, let us mention some issues requiring further investigation. For convenience of implementation, we have used equally space knots for the B-splines. It would be worthwhile to consider data-dependent knot placement, designed to detect key features of time-course gene expression profiles such as peaks and troughs. Since time-course gene expressions are typically short time-series, the improvement may be marginal. Another issue is the investigation of other features discriminating functional classes of genes. For stock pricing data in Tarpey and Kinateder (2003), the first derivative $x'(t)$ reflecting the trend worked better than the original curve $x(t)$ in clustering. Another example is the near infrared absorbance spectrum of meat data in Rossi and Villa (2006), where FSVM on the curvature of the spectra $x''(t)$ separated meat samples with high and low fat better in terms of test error rates. As illustrated in Jank and Shmueli (2005), B-splines with penalization, called P-splines, can be a useful basis system in analyzing differentiated curves that reflect temporal dynamics of the underlying data generation mechanism.

Acknowledgments

We would like to appreciate the editor, an associate editor, and anonymous reviewers for their constructive comments, which improved the presentation of the paper greatly. This research was supported by Korea Research Foundation Grant funded by Korea Government (MOEHRD, Basic Research Promotion Fund) (KRF-2005-070-C00020).

References

Bensmail, H., Aruna, B., Semmes, O.J., Haoudi, A., 2005. Functional clustering algorithm for high dimensional protemics data. J. Biomed. Biotechnol. 2, 80–86.

Biau, G., Bunea, F., Wegkamp, M.H., 2005. Functional classification in Hilbert spaces. IEEE Trans. Inform. Theory 51, 2163–2172.

Chudova, D., Hart, C., Mjolsness, E., Smyth, P., 2004. Gene expression clustering with functional mixture models. In: Advances in Neural Information Processing 16, MIT Press, Cambridge, MA.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

DeRisi, J.L., Iyer, V.R., Brown, P.O., 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278, 680–686.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. In: Proceedings of National Academy of Sciences, pp. 14863–14868.

Ernst, J., Nau, G.J., Bar-Joseph, Z., 2005. Clustering short time series gene expression data. Bioinformatics 21, 159–168.

Gui, J., Li, H., 2003. Mixture functional discriminant analysis for gene function classification based on time course gene expression data. In: Proceedings of the Joint Statistical Meeting.

Heckman, N., Zamar, R., 2000. Comparing the shapes of regression function. Biometrika 87, 135–144.

Hsu, C.-W., Lin, C.-J., 2002. A comparison of methods for multi-class support vector machines. IEEE Trans. Neural Networks 13, 415–425.

Iranfar, N., Fuller, D., Sasik, R., Hwa, T., Laub, M., Loomis, W.F., 2001. Expression patterns of cell-type-specific genes in dictyostelium. Mol. Biol. Cell 12, 2590–2600.

James, G., Sood, A., 2006. Performing hypothesis tests on the shape of functional data. Comput. Statist. Data Anal. 50, 1774–1792.

Jank W., Shmueli, G., 2005. Dynamic Profiling of Online Auctions Using Curve Clustering. Department of Decision and Information Technologies, University of Maryland, College Park, MD, 20742.

Lee, H.-J., 2004. Functional data analysis: classification and regression, Ph.D. Thesis, Department of Statistics, Texas A&M University.

Leng, X., Müller, H.-G., 2006. Classification using functional data analysis for temporal gene expression data. Bioinformatics 22, 68–76.

Luan, Y., Li, H., 2003. Clustering of time-course gene expression data using a mixture-effects model with B-splines. Bioinformatics 19, 474–482.

Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., Weil, B., 2002. MIPS: a database for genomes and protein sequences. Nucl. Acids Res. 30, 31–34.

Ohba, M., 1997. Modulation of intracellular protein degradation by SSB1–SIS1 chaperon system in yeast *S. cerevisiae*. FEBS Lett. 409 (2), 307–311.

Ramsay, J.O., Silverman, B.W., 1997. Functional Data Analysis. Springer, New York.

Rossi, F., Villa, N., 2005. Classification in Hilbert spaces with support vector machines. In: Proceedings of XIth International Symposium on Applied Stochastic Models and Data Analysis, ASMDA 2005, Brest, France.

Rossi, F., Villa, N., 2006. Support vector machine for functional data classification. Neurocomputing 69, 730–742.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Bostein, D., Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol. Biol. Cell 9, 3273–3297.

Tarpey, T., Kinateder, K.K.J., 2003. Clustering functional data. J. Classification 20, 93–114.

Vapnik, V.N., 1996. The Nature of Statistical Learning Theory. Springer, New York.