# Brief Papers

## Behavior-Constrained Support Vector Machines for fMRI Data Analysis

Danmei Chen, Sheng Li, Zoe Kourtzi, and Si Wu

*Abstract*—**Statistical learning methods are emerging as a valuable tool for decoding information from neural imaging data. The noisy signal and the limited number of training patterns that are typically recorded from functional brain imaging experiments pose a challenge for the application of statistical learning methods in the analysis of brain data. To overcome this difficulty, we propose using prior knowledge based on the behavioral performance of human observers to enhance the training of support vector machines (SVMs). We collect behavioral responses from human observers performing a categorization task during functional magnetic resonance imaging scanning. We use the psychometric function generated based on the observers behavioral choices as a distance constraint for training an SVM. We call this method behavior-constrained SVM (BCSVM). Our findings confirm that BCSVM outperforms SVM consistently.**

*Index Terms*—**Functional magnetic resonance imaging (fMRI), pattern classification, psychometric function, support vector machine (SVM).**

### I. Introduction

Statistical learning methods, such as support vector machines (SVM) [1]–[7], are emerging as a valuable tool for the analysis of functional brain imaging data. In particular, they have been successfully applied to extract information for visual features (e.g., orientation and position) and object categories from functional magnetic resonance imaging (fMRI) signals [8]–[13]. In a typical experimental setting, a pattern classifier is first trained with labeled stimulus-activation pairs. After the training, the classifier's performance in predicting stimuli from independent fMRI patterns is tested. Compared with other approaches, such as general linear model (GLM) [12] or calculation of mutual information [13], statistical learning methods have the advantage of being more accurate or requiring less amount of data.

Despite this initial success, a substantial technical obstacle to the application of statistical learning methods is that functional brain imaging data is highly noisy and the number of training examples is often very small compared to the dimensionality of activity patterns, leading to the so-called curse of dimensionality in statistical learning [14]. Thus, it is possible that although there is adequate information for decoding stimuli from fMRI data, an inefficiently-trained pattern classifier is unable to read them out. To improve the training of a pattern classifier, efforts have been made to optimize the pre-processing of fMRI signals (see [7]). However, since the dimensionality of fMRI data is still very high, the curse of dimensionality problem has not been properly solved. To overcome this difficulty, a promising solution is to take into account prior knowledge about the data. In [15], the authors used the knowledge about the spatial correlation of fMRI signals to improve the training of SVM and got encouraging results.

In this paper, we propose a method using prior knowledge based on the observers' behavior to enhance the training of SVM. Our approach is motivated by observations that in many classification tasks the performance of human observers is similar to that of optimal pattern classifiers. This is supported by a recent psychophysical paper [16], in which a gender classification task based on 2-D face images was carried out by both human observers and a linear SVM. Comparing the psychometric function of the human observers (measuring the probability of classification error) and the discrimination function of the SVM (measuring the distance of a pattern from the separating boundary) showed that the two quantities agree with each other very well. This result suggests that there exists a strong link between optimal pattern classifiers and human behavior (i.e., discrimination of critical stimulus features for a given task).

Inspired by this finding, we tested whether taking into account behavioral performance in a categorization task would improve the performance of SVM in discriminating the perceived stimulus categories. In our experiment, human observers were first trained to perform a visual categorization task between two stimulus categories (radial vs. concentric Glass patterns) until they reached stable performance. After the training, the observers performed a categorization task during fMRI scanning. We used the psychometric function generated from each observer's behavioral responses across different stimulus conditions as a distance constraint to modify the conventional training of SVM. We call this method behavior-constrained SVM (BCSVM). We predicted that BCSVM would outperform SVM as it utilizes extra information about the distance of data points from the separating boundary. We applied BCSVM to fMRI data and confirmed that BCSVM works very well. To our knowledge, this approach is the first one that utilizes the behavioral performance of human observers to improve the performance of pattern classification.

### II. Behavior-Constrained SVM

In the categorization task we employ, stimuli are divided into two classes. Our goal is to train a pattern classifier

Manuscript received May 5, 2010; revised July 13, 2010; accepted July 14, 2010. Date of publication September 8, 2010; date of current version October 6, 2010. The work of Z. Kourtzi and S. Wu was supported by BBSRC Cognitive Foresight Initiative Grant BB/E027436/1. D. Chen and S. Li contributed equally to this work; correspondence should be addressed to S. Wu.

D. Chen is with the Department of Informatics, University of Sussex, Brighton BN1 9RH, U.K.

S. Li is with the Department of Psychology, Peking University, Beijing 100871, China, and also with the School of Psychology, University of Birmingham, Birmingham B15 2TT, U.K.

Z. Kourtzi is with the School of Psychology, University of Birmingham, Birmingham B15 2TT, U.K.

S. Wu is with the Laboratory of Neural Information Processing, Institute of Neuroscience, Chinese Academy of Sciences, Shanghai 200031, China (e-mail: siwu@ion.ac.cn).

Digital Object Identifier 10.1109/TNN.2010.2060353

to predict the class labels of the stimuli based on their evoked voxel activities. Two types of data are acquired in the experiment. One is the voxel activity evoked by the stimuli, which we denote as $\mathbf{x}_i$, for $i = 1, \ldots, N$, where $N$ is the total number of trials, and $y_i = 1$ or $-1$ represents the corresponding class label. The other is the psychometric function of human observers, which we denote as $p(s_j)$ for the stimulus condition $s_j$, for $j = 1, \ldots, M$, with $M$ the number of stimulus conditions. In the categorization task we employ, we choose $M = 7$, i.e., seven different stimulus conditions, and for each stimulus condition, noise is added to generate different implementations of the same stimulus. Each implementation of the stimulus generates one voxel pattern.

The psychometric function measures the classification errors of human observers for each stimulus condition. Typically, the closer a pattern is to the separating boundary, the higher the probability that human observers will make a classification mistake. Thus, the psychometric function can provide useful information about the potential distance of a pattern from the boundary. In this paper, we convert the psychometric function into the distance to the boundary according to the following rule (we also tried other rules that showed similar performances):

$$\begin{aligned} d_i &= |2p(s_j) - 1| \quad \text{for } \mathbf{x}_i \in s_j \quad 0 < p(s_j) < 1 \\ d_i &\geq 1 \qquad \text{for } \mathbf{x}_i \in s_j \, p(s_j) = 1 \text{ or } p(s_j) = 0. \end{aligned} \tag{1}$$

Here, the condition $\mathbf{x}_i \in s_j$, means that the voxel pattern $\mathbf{x}_i$ is generated by the stimulus condition $s_j$.

Thus, training examples $\{\mathbf{x}_i\}$, for $i = 1, \ldots, N$, are divided into two types. One is those whose distances from the boundary are smaller than one. We call them critical examples. They will form the equality constraints in BCSVM and will become support vectors after training. We denote $N_1$ as the number of critical examples. The other type is the training examples whose distances from the boundary are larger than one. They correspond to the stimuli for which human observers show no classification error, implying that they are far away from the boundary. Since there is no further information on the distances of these examples to the boundary, they satisfy inequality constraints in BCSVM. For explanatory purposes, we align all examples and let the first $N_1$ be the critical ones. This has no effect on the training performance, since the cost function is quadratic and has global minima.

We apply the distance information in the training of the classifier. The discrimination function of BCSVM is written as

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b. \tag{2}$$

Its parameters are optimized through minimizing the following cost function:

$$\min_{\mathbf{w},b,\sigma,\xi} \quad \tfrac{1}{2}||\mathbf{w}||^2 + \tfrac{D}{2}\textstyle\sum_{i=1}^{N_1} \sigma_i^2 + C\sum_{i=N_1+1}^{N} \xi_i$$
$$\text{s.t.} \quad \begin{cases} y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = d_i - \sigma_i & \text{for } 1 \leq i \leq N_1 \\ y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i & \text{for } N_1 < i \leq N \\ \xi_i \geq 0 & \text{for } N_1 < i \leq N \end{cases} \tag{3}$$

where $0 < d_i < 1$ for $1 \leq i \leq N_1$.

Compared with the conventional SVM, BCSVM has extra equality constraints for $1 \leq i \leq N_1$, i.e., to those critical examples, and a term, $D/2\sum_{i=1}^{N_1} \sigma_i^2$, in the cost function, penalizing the violation of these constraints. Note that the parameter $\sigma_i$ is not required to be positive, since noise has equal probability in increasing or decreasing the distance.

By applying the standard Lagrangian method, we obtain the Lagrangian function as follows:

$$\begin{aligned} \mathcal{L}(\mathbf{w}, b, \xi, \alpha, \beta) = {} & \frac{1}{2}||\mathbf{w}||^2 + \frac{D}{2}\sum_{i=1}^{N_1} \sigma_i^2 + C\sum_{i=N_1+1}^{N} \xi_i \\ & - \sum_{i=1}^{N_1} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - d_i + \sigma_i] \\ & - \sum_{i=N_1+1}^{N} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=N_1+1}^{N} \beta_i \xi_i \end{aligned} \tag{4}$$

where $\alpha_i \geq 0$ and $\beta_i \geq 0$ when $i > N_1$.

At the saddle point, we have

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 = \mathbf{w} - \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \tag{5}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 = \sum_{i=1}^{N} \alpha_i y_i \tag{6}$$

$$\frac{\partial \mathcal{L}}{\partial \sigma_i} = 0 = D\sigma_i - \alpha_i \quad \text{for } 1 \leq i \leq N_1 \tag{7}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 = C - \alpha_i - \beta_i \quad \text{for } N_1 < i \leq N. \tag{8}$$

Using the above relationships, we get the dual optimization problem as follows:

$$\begin{aligned} \max_{\alpha} W(\alpha) = {} & -\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j + \sum_{i=1}^{N_1} \alpha_i d_i \\ & + \sum_{i=N_1+1}^{N} \alpha_i - \frac{1}{2D}\sum_{i=1}^{N_1} \alpha_i^2 \\ \text{s.t.} \quad & \begin{cases} -\infty < \alpha_i < \infty & \text{for } 1 < i \leq N_1 \\ 0 \leq \alpha_i \leq C & \text{for } N_1 < i \leq N \\ \sum_{i=1}^{N} \alpha_i y_i = 0. \end{cases} \end{aligned} \tag{9}$$

Again, compared with the dual formulation of SVM, the differences are on the constraints for $\alpha_i$ when $i \leq N_1$ and the corresponding penalty terms in the cost function.

Finally, the solution of BCSVM is written as

$$f(\mathbf{x}) = \sum_{i \in SV} \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b \tag{10}$$

where the summation runs over all support vectors, which are $\alpha_i$ for $i \leq N_1$ and those non-vanishing $\alpha_i$ for $i > N_1$.

The determination of the parameter $b$ is different from that in SVM, since for BCSVM there are many equality constraints. We find that if the number of critical examples is sufficiently large, the following strategy works very well:

$$b = \frac{1}{N_1}\sum_{i=1}^{N_1} (\mathbf{w} \cdot \mathbf{x}_i - d_i y_i). \tag{11}$$

Obviously, since both the formulation and the final solution for BCSVM depend only on the inner product between data points, the kernel trick can be used [17], and we can easily generalize BCSVM to non-linear cases, that is

$$f(\mathbf{x}) = \sum_{i \in SV} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \tag{12}$$

where $K(\mathbf{x}, \mathbf{x}')$ is the kernel function.

We can, in principle, solve the optimization problem (9) by using the standard quadratic programming algorithm. However, this is too slow for a large data set. To speed up the calculation, we develop an algorithm mimicking the fast SMO approach in SVM [18]. The computational speed of BCSVM is therefore about the same as that of SVM when SMO is used. The detail of the algorithm is presented in the Appendix.

One may consider the possibility of using the SVM regression approach that takes into account the behavior of human observers, e.g., to only include equality constraints in (2). This method, however, is not the most efficient way to utilize the observers' performance. For training examples with perfect behavioral performance [i.e., $p(s) = 1$ or 0 for the psychometric function], we have no further information about their distance from the boundary. Therefore, we set these training examples so that they satisfy the inequality constraints. These examples are still valuable in determining the optimal position of the classification boundary and can not be simply omitted. This is particularly true, when the number of training examples is much smaller than the dimensionality of data points. Thus, it is best to include both equality and inequality constraints, as in BCSVM for fully utilizing the behavioral information.

## III. BCSVM for fMRI Data Analysis

We apply BCSVM to decode information from the data collected in an fMRI experiment, in which human observers categorized two classes of visual stimuli [Glass patterns: radial vs. concentric [19] (Fig. 1)] according to pre-defined boundaries.

### A. fMRI Experiment

In the experiment, white dot pairs (dipoles) were displayed within a square aperture (7.7° × 7.7°) on a black background (100% contrast). We first generated concentric and radial Glass patterns by placing dipoles tangentially (concentric stimulus) or orthogonal (radial stimulus) to the circumference of a circle centered at the fixation dot. We then generated intermediate patterns between these two Glass pattern types by manipulating continuously the spiral angle of the pattern from 0° (radial) to 90° (concentric). For each dot dipole, the spiral angle was defined as the angle between the dot dipole orientation and the radius from the center of the dipole to the center of the stimulus aperture. For a given pattern of 100% signal, dot dipoles were all aligned to the specified spiral angle. Noise was introduced by varying the orientation of a portion of dot dipoles randomly, e.g., a stimulus of 60% signal means 60% dot dipoles are aligned to a given spiral angle while the rest have random orientations. The separating boundary was set to be either at the spiral angle 30° or 60° (Fig. 1). This means that for a stimulus pattern, if its spiral angle is smaller than the spiral angle of the boundary, then it belongs to the radial category, otherwise, it belongs to the concentric one. To carry out this discriminative task, human subjects, and so does a learning machine, must extract the global features (i.e., spiral angle), rather than local cues, of Glass patterns.

The scanning experiments were conducted at the Birmingham University Imaging Center with 3T Philips Achieva scanner. EPI data (Gradient echo-pulse sequences) were acquired from 24 slices (whole-brain coverage, TR, 1500 ms, TE, 35 ms, flip-angle, 73°, 2.5 × 2.5 × 4 mm resolution). Eight observers participated in the discrimination task. They were first trained to get familiar with the task in a psychophysical training before scanning. Afterward, they carried out two scanning sessions during which they performed the categorization task on the Glass pattern stimuli on each of the two boundaries (30° and 60° spiral angles). For each observer, we collected data from 7 to 8 event-related runs in each session, and each run consists of 128 trials. For the 30° boundary, the stimulus conditions comprised Glass patterns of 0°, 15°, 25°, 30°, 35°, 60°, and 90° spiral angles. This gives $M = 7$ stimulus conditions. For the 60° boundary, the stimulus conditions comprised Glass patterns of 0°, 30°, 55°, 60°, 65°, 75°, and 90° spiral angles.

### B. Data Preprocessing

Pre-processing of the fMRI data included slice-scan time correction, head movement correction, temporal high-pass filtering (three cycles) and removal of linear trends. Spatial smoothing was not performed for data used for the pattern classification analysis. Trials with head motion larger than 1 mm of translation or 1° of rotation were excluded from the analysis. We averaged signals from each set of eight trials from the same condition in each run to generate one training example, resulting in 96 training examples in total for each session. 100 voxels from each cortical area were selected as input data for classification.

The classification performance of human subjects was recorded during the scanning sessions, which generated the psychometric functions as shown in Fig. 2(a). The values of the psychometric functions were then converted into the corresponding distance constraints according to (1).

### C. Analyze fMRI Data with BCSVM

After acquiring behavioral performance and fMRI data, we applied BCSVM to decode the perceived stimulus categories. The prediction accuracy of BCSVM was compared with that of the conventional SVM that does not utilize the behavioral performance of human observers.

We estimated the classification accuracies of BCSVM and SVM using an N-fold cross-validation procedure (N is number of the runs in each session). The distance constraint $d$ was determined only based on the performance of an observer on the training trials (i.e., the behavioral data on the testing trials were excluded for computing the psychometric function). To ensure fair comparison, the parameter $C$ in the standard SVM

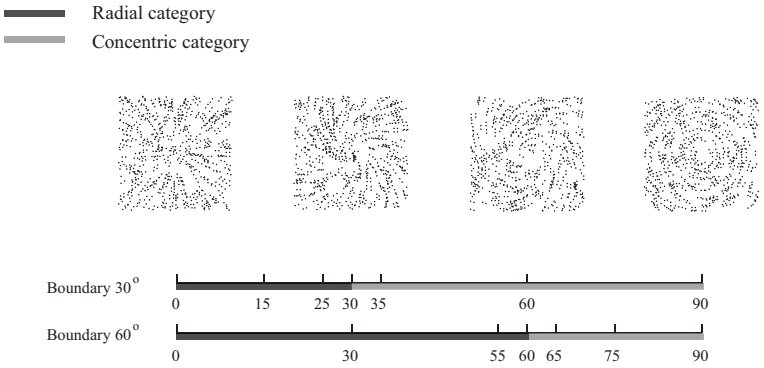


Fig. 1. Stimulus and category space. Four examples of Glass pattern stimuli (100% signal) at spiral angles of 0°, 30°, 60°, and 90° are shown. Categorical boundaries and spiral angles tested during the scanning sessions are also shown (black bar: stimuli that resemble radial, grey bar: stimuli that resemble concentric) that indicate the categorical membership of the stimuli for each boundary in the experiment.

was first optimized, which turned out to be 0.8. We then optimized the parameter $D$ in BCSVM, which turned to be 0.09. The same $C$ and $D$ values were used across areas.

The fMRI patterns were labeled according to their stimulus category in physical space rather than the category perceived by the observers as indicated by the observers' behavioral choice. The logic of this classification scheme is that information about the observers' behavior would facilitate classification of the fMRI signals evoked by the different stimulus categories. This is especially useful when the stimulus patterns are close to the category boundary which implies more uncertainty in the categorical decision and fMRI signals.

For each observer, we performed this analysis separately for each session and averaged classification performance across sessions. We compared classification accuracies (averaged across 16 sessions from 8 observers) for BCSVM and SVM in three representative cortical areas: 1) V1, known to encode the basic local visual feature; 2) the higher occipitotemporal cortex lateral occipital (LO), known to have preferential responses for radial vs. concentric stimuli [20]; and 3) the dorsal lateral prefrontal cortex (DLPFC), known to play a fundamental role in categorization [10], [23], [24].

Fig. 2(a) shows the psychometric function of an observer recorded during the scanning sessions. Fig. 2(b) compares the classification performances of BCSVM and SVM. We observe that BCSVM outperforms SVM as indicated by a significant difference between the performance of the two classifiers [$F(1, 15) = 16.3$, $p = 0.001$]. In particular, BCSVM significantly outperformed SVM in LO [$t(15) = 2.82$, $p = 0.013$] and DLPFC [$t(15) = 6.18$, $p < 0.001$]. This is consistent with the previous findings showing that areas in the higher occipitotemporal and the frontal cortex are involved in categorization decision tasks, and neural activity in these regions is correlated with discriminative behavioral performances [10], [23], [25], [26]. Thus, BCSVM can outperform SVM in these areas by utilizing the valuable behavioral information. In contrast, neural activity in V1 is known to represent only local stimulus features (e.g., the orientation of dipoles), and is independent of the observer's behavioral performance. Therefore, the behavioral performance has no significant effect in improving the accuracy of BCSVM in V1 [$t(15) = 1.56$, $p = 0.139$].

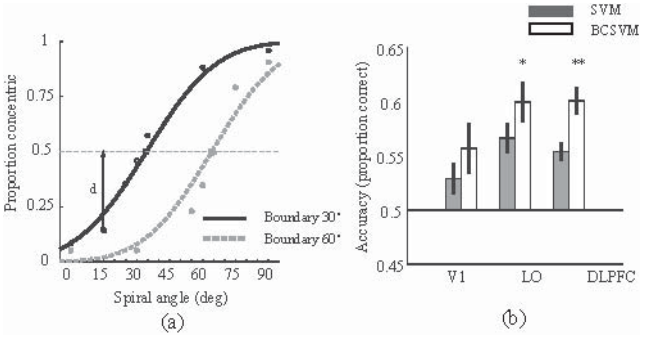


Fig. 2. (a) Psychometric function. Behavioral data collected during the scanning sessions (circles) are shown for each boundary. Lines indicate the cumulative Gaussian fits of the behavioral data. Error bars indicate the 95% confidence interval at 50% concentric threshold following a bootstrap procedure. (b) Classification performances. Classification accuracies of BCSVM and SVM in three representative cortical areas: V1, LO, and DLPFC. Error bars indicate the standard error of mean across observers and sessions (* $p < 0.05$, ** $p < 0.001$).

## IV. Conclusion

In this paper, we have proposed a novel approach, BCSVM, to improve the performance of pattern classification in fMRI data analysis. BCSVM aims to overcome the limitations often imposed by a small number of training examples and highly noisy data in neuroimaging. In contrast to the conventional application of SVM in the classification of fMRI data (see [6]–[9]), BCSVM utilizes the probabilistic behavioral performance of human observers. This additional information is used as a distance constraint in BCSVM and helps to improve the accuracy of the classifier when fMRI signals in a cortical area are associated with a cognitive task. Our experimental studies confirm that BCSVM outperforms SVM consistently. We expect that BCSVM may serve as a general tool to enhance the applications of statistical learning methods when the number of training examples is limited and highly noisy. Nevertheless, further work is needed to establish the utility of BCSVM. In particular, theoretical quantification of the BCSVM performance is necessary when the distance information is uncertain. This is important since the performance of human observers may fluctuate highly. Furthermore, the utility of BCSVM can be tested for predicting a larger range of

stimulus features and multimodal brain imaging signals, e.g., EEG, MEG, and optical imaging. Finally, we would like to point out that although the present paper only applied BCSVM to classifying fMRI data, it can be applied to general situations whenever prior knowledge about the distance information of training patterns to the boundary is available.

## APPENDIX A
## THE IMPLEMENTATION OF BCSVM

We use an iterative method to maximize the cost function in (9). Here the main concern is to ensure the equality constraint is not violated when updating the parameters. We borrow the idea from the SMO algorithm in SVM, that is, we only update two parameters in each step of adjusting, and enforce the equality constraint satisfied strictly. Since only two parameters are involved each time, a closed-form on the maximum changes of the parameters are analytically available, the actual converging speed of the algorithm is fast. In below, we introduce the detail of the algorithm.

Without loss of generality, we consider two parameters, $\alpha_i$ and $\alpha_j$, are updated. To satisfy the equality constraint, we impose the following condition:

$$y_i\alpha_i^{new} + y_j\alpha_j^{new} = y_i\alpha_i^{old} + y_j\alpha_j^{old}. \tag{13}$$

This leads to

$$\alpha_i^{new} = \begin{cases} \alpha_i^{old} - \alpha_j^{old} + \alpha_j^{new} & \text{if } y_i \neq y_j \\ \alpha_i^{old} + \alpha_j^{old} - \alpha_j^{new} & \text{if } y_i = y_j. \end{cases} \tag{14}$$

Since $\alpha_i$ and $\alpha_j$ are constrained, the above relationship restricts the lower and upper bounds of $\alpha_i^{new}$, which we denote as $U$ and $V$, respectively, i.e., $U \leq \alpha_i^{new} \leq V$. These two bounds will be later used to truncate the updating value of $\alpha_i^{new}$ in the optimizing process. We distinguish eight different cases as follows.

1) For $y_i = y_j$.
   a) If $1 \leq i, j \leq N_1$
   $$\begin{cases} U = -\infty \\ V = +\infty. \end{cases} \tag{15}$$
   b) If $1 \leq i \leq N_1$ and $j > N_1$
   $$\begin{cases} U = \alpha_i^{old} + \alpha_j^{old} - C \\ V = \alpha_i^{old} + \alpha_j^{old}. \end{cases} \tag{16}$$
   c) If $i, j > N_1$
   $$\begin{cases} U = \max(0, \alpha_i^{old} + \alpha_j^{old} - C) \\ V = \min(U, C, \alpha_i^{old} + \alpha_j^{old}). \end{cases} \tag{17}$$
   d) If $i > N_1$ and $1 \leq j \leq N_1$
   $$\begin{cases} U = 0 \\ V = C. \end{cases} \tag{18}$$
2) For $y_i \neq y_j$.
   a) If $1 \leq i, j \leq N_1$
   $$\begin{cases} U = -\infty \\ V = +\infty. \end{cases} \tag{19}$$
   b) If $1 \leq i \leq N_1$ and $j > N_1$
   $$\begin{cases} U = \alpha_i^{old} - \alpha_j^{old} \\ V = \alpha_i^{old} - \alpha_j^{old} + C. \end{cases} \tag{20}$$
   c) If $i, j > N_1$
   $$\begin{cases} U = \max(0, \alpha_i^{old} - \alpha_j^{old}) \\ V = \min(U, C, \alpha_i^{old} - \alpha_j^{old} + C). \end{cases} \tag{21}$$
   d) If $i > N_1$ and $1 \leq j \leq N_1$
   $$\begin{cases} U = 0 \\ V = C. \end{cases} \tag{22}$$

Now let us optimize the cost function with respect to $\alpha_i$ and $\alpha_j$ when all other parameters are fixed and are under the condition as follows:

$$s\alpha_i^{new} + \alpha_j^{new} = s\alpha_i^{old} + \alpha_j^{old} = \gamma \tag{23}$$

where $s = y_i y_j$ and $\gamma$ is a constant at each time of updating. We write down $W$ as a functional of $\alpha_i$ and $\alpha_j$ as follows:

$$\begin{aligned} W(\alpha_i, \alpha_j) = &-\frac{1}{2}K_{ii}\alpha_i^2 - \frac{1}{2}K_{jj}\alpha_j^2 - sK_{ij}\alpha_i\alpha_j - y_i\alpha_i v_i \\ &- y_j\alpha_j v_j + G(\alpha_i) + G(\alpha_j) + constant \end{aligned} \tag{24}$$

where

$$K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) \tag{25}$$

$$v_i = \sum_{l \neq i,j} y_l\alpha_l K_{i,l} \tag{26}$$

$$G(\alpha_i) = \begin{cases} \alpha_i & \text{if } i > N_1 \\ d_i\alpha_i - \alpha_i^2/(2D) & \text{if } 1 \leq i \leq N_1. \end{cases} \tag{27}$$

By using the relationship, $\alpha_j = \gamma - s\alpha_i$, we get the following:

$$\begin{aligned} W(\alpha_i) = &-\frac{1}{2}K_{ii}\alpha_i^2 - \frac{1}{2}K_{jj}(\gamma - s\alpha_i)^2 - sK_{ij}\alpha_i(\gamma - s\alpha_i) \\ &- y_i\alpha_i v_i - y_j(\gamma - s\alpha_i)v_j + G(\alpha_i) + G(\gamma - s\alpha_i) \\ &+ constant \end{aligned} \tag{28}$$

and, hence

$$\begin{aligned} \frac{\partial W(\alpha_i)}{\partial \alpha_i} &= -K_{ii}\alpha_i + sK_{jj}(\gamma - s\alpha_i) \\ &\quad + K_{ij}\alpha_i - sK_{ij}(\gamma - s\alpha_i) - y_i v_i + y_i v_j \\ &\quad + G'(\alpha_i) + G'(\gamma - s\alpha_i) \\ &= \alpha_i(2K_{ij} - K_{ii} - K_{jj}) + \gamma s(K_{jj} - K_{ij}) \\ &\quad + y_i(v_j - v_i) + G'(\alpha_i) + G'(\gamma - s\alpha_i) \\ &= 0. \end{aligned} \tag{29}$$

Define $\kappa = K_{ii} + K_{jj} - 2K_{ij}$, we have the following:

$$\kappa\alpha_i - G'(\alpha_i) - G'(\gamma - s\alpha_i) = y_i[f(\mathbf{x}_j) - f(\mathbf{x}_i)] + \alpha_i\kappa. \tag{30}$$

We distinguish four different cases as follows.

1) When $i, j > N_1$
$$G'(\alpha_i) + G'(\gamma - s\alpha_i) = 1 - s \tag{31}$$
$$\alpha_i^{new} = \alpha_i^{old} + \frac{y_i[f(\mathbf{x}_j) - f(\mathbf{x}_i) - y_j + y_i]}{\kappa}. \tag{32}$$

2) When $1 \le i \le N_1$ and $j > N_1$

$$G'(\alpha_i) + G'(\gamma - s\alpha_i) = d_i - s - \alpha_i/D \quad (33)$$

$$\alpha_i^{new} = \frac{D\kappa\alpha_i^{old}}{D\kappa+1} + \frac{Dy_i[f(\mathbf{x}_j) - f(\mathbf{x}_i) - y_j + d_i y_i]}{D\kappa+1}. \quad (34)$$

3) When $1 \le i, j \le N_1$

$$G'(\alpha_i) + G'(\gamma - s\alpha_i) = d_i - d_j s + s\gamma/D - 2\alpha_i/D \quad (35)$$

$$\alpha_i^{new} = \frac{D\kappa\alpha_i^{old}}{D\kappa+2} + \frac{Dy_i[f(\mathbf{x}_j) - f(\mathbf{x}_i) - d_j y_j + d_i y_i + y_j\gamma/D]}{D\kappa+2}. \quad (36)$$

4) When $1 \le j \le N_1$ and $i > N_1$

$$G'(\alpha_i) + G'(\gamma - s\alpha_i) = 1 - d_j s + s\gamma/D - \alpha_i/D \quad (37)$$

$$\alpha_i^{new} = \frac{D\kappa\alpha_i^{old}}{D\kappa+1} + \frac{Dy_i[f(\mathbf{x}_j) - f(\mathbf{x}_i) - d_j y_j + y_i + y_j\gamma/D]}{D\kappa+1}. \quad (38)$$

Finally, we clip $\alpha_i^{new}$ to ensure it is in the range of $[U, V]$. The value of $\alpha_j^{new}$ is given by

$$\alpha_j^{new} = \alpha_j^{old} + y_i y_j(\alpha_i^{old} - \alpha_i^{new}). \quad (39)$$

## REFERENCES


[1] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

[2] P. Williams, S. Li, J. Feng, and S. Wu, "A geometrical method to improve performance of the support vector machine," *IEEE Trans. Neural Netw.*, vol. 18, no. 3, pp. 942–947, May 2007.

[3] K. Kobayashi and F. Komaki, "Information criteria for support vector machines," *IEEE Trans. Neural Netw.*, vol. 17, no. 3, pp. 571–577, May 2006.

[4] Q. Tao, D. Chu, and J. Wang, "Recursive support vector machines for dimensionality reduction," *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 189–193, Jan. 2008.

[5] S. Decherchi, S. Ridella, R. Zunino, P. Gastaldo, and D. Anguita, "Using unsupervised analysis to constrain generalization bounds for support vector classifiers," *IEEE Trans. Neural Netw.*, vol. 21, no. 3, pp. 424–438, Mar. 2010.

[6] D. D. Cox and R. L. Savoy, "Functional magnetic resonance imaging (fMRI) 'brain reading': Detecting and classifying distributed patterns of fMRI activity in human visual cortex," *Neuroimage*, vol. 19, no. 2, pp. 261–270, 2003.

[7] S. LaConte, S. Strother, V. Cherkassky, J. Anderson, and X. Hu, "Support vector machines for temporal classification of block design fMRI data," *Neuroimage*, vol. 26, no. 2, pp. 317–329, 2005.

[8] S. Li, D. Ostwald, M. Giese, and Z. Kourtzi, "Flexible coding for categorical decisions in the human brain," *J. Neurosci.*, vol. 27, no. 45, pp. 12321–12330, 2007.

[9] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby, "Beyond mind-reading: Multi-voxel pattern analysis of fMRI data," *Trends Cogn. Sci.*, vol. 10, no. 9, pp. 424–430, Sep. 2006.

[10] T. Mitchell, R. Hutchinson, R. Niculescu, F. Pereira, X. Wang, M. Just, and S. Neman, "Learning to decode cognitive states from brain images," *Mach. Learn.*, vol. 57, nos. 1–2, pp. 145–175, Oct.–Nov. 2004.

[11] Z. Wang, A. Childress, J. Wang, and J. Detre, "Support vector machine learning based fMRI data group analysis," *Neuroimage*, vol. 36, no. 4, pp. 1139–1151, Jul. 2007.

[12] K. J. Friston, A. P. Holmes, K. J. Worsley, J. Poline, C. D. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: A general linear approach," *Human Brain Map.*, vol. 2, no. 11, pp. 189–210, 1995.

[13] M. Montemurro, R. Senatore, and S. Panzeri, "Tight data-robust bounds to mutual information combining shuffling and model selection techniques," *Neural Computat.*, vol. 19, no. 11, pp. 2913–2957, Nov. 2007.

[14] C. Bishop, *Neural Networks for Pattern Recognition*. New York: Oxford Univ. Press, 1995.

[15] L. Liang, V. Cherkassky, and D. Rottenberg, "Spatial SVM for feature selection and fMRI activation detection," in *Proc. Int. Joint Conf. Neural Netw.*, 2006, pp. 1463–1469.

[16] A. Graf, F. Wichmann, H. Bulthöff, and B. Schölkopf, "Classification of faces in man and machine," *Neural Computat.*, vol. 18, no. 1, pp. 143–165, Jan. 2006.

[17] B. Schölkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge, MA: MIT Press, 2002.

[18] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C. Burgess, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999, pp. 185–208.

[19] L. Glass, "Moire effect from random dots," *Nature*, vol. 223, no. 5206, pp. 578–580, Aug. 1969.

[20] D. Ostwald, J. M. Lam, S. Li, and Z. Kourtzi, "Neural coding of global form in the human visual cortex," *J. Neurophysiol.*, vol. 99, no. 5, pp. 2456–2469, May 2008.

[21] J. Duncan, "An adaptive coding model of neural function in prefrontal cortex," *Nat. Rev. Neurosci.*, vol. 2, no. 11, pp. 820–829, Nov. 2001.

[22] E. K. Miller, "The prefrontal cortex and cognitive control," *Nat. Rev. Neurosci.*, vol. 1, no. 1, pp. 59–65, Oct. 2000.

[23] N. Sigala and N. K. Logothetis, "Visual categorization shapes feature selectivity in the primate temporal cortex," *Nature*, vol. 415, no. 6869, pp. 318–320, Jan. 2002.

[24] S. Li, S. D. Mayhew, and Z. Kourtzi, "Learning shapes the representation of behavioral choice in the human brain," *Neuron*, vol. 62, no. 3, pp. 441–452, May 2009.


## Facial Expression Recognition in JAFFE Dataset Based on Gaussian Process Classification

Fei Cheng, Jiangsheng Yu, and Huilin Xiong

***Abstract*— The Gaussian process (GP) approaches to classification synthesize Bayesian methods and kernel techniques, which are developed for the purpose of small sample analysis. Here we propose a GP model and investigate it for the facial expression recognition in the Japanese female facial expression dataset. By the strategy of leave-one-out cross validation, the accuracy of the GP classifiers reaches 93.43% without any feature selection/extraction. Even when tested on all expressions of any particular expressor, the GP classifier trained by the other samples outperforms some frequently used classifiers significantly. In order to survey the robustness of this novel method, the random trial of 10-fold cross validations is repeated many times to provide an overview of recognition rates. The experimental results demonstrate a promising performance of this application.**

Manuscript received October 26, 2009; revised July 21, 2010; accepted July 29, 2010. Date of publication August 19, 2010; date of current version October 6, 2010. This work was supported in part by Peking University through the 985 Project, in part by the Beijing Natural Science Foundation under Grant 048SG/46810707-001 and Grant 4032013, and in part by the Natural Science Foundation of China through Project 60775008.

F. Cheng is with the Department of Mathematics, Beijing Jiaotong University, Beijing 100044, China (e-mail: fcheng@bjtu.edu.cn).

J. S. Yu is with the Department of Computer Science and Technology, Key Laboratory of High Confidence Software Technologies, Ministry of Education, Peking University, Beijing 100871, China (e-mail: yujs@pku.edu.cn).

H. L. Xiong is with the Department of Automation, Shanghai Jiaotong University, Shanghai 200240, China (e-mail: hlxiong@sjtu.edu.cn).

Digital Object Identifier 10.1109/TNN.2010.2064176