Abstract—This paper proposes a local linear multi-SVM method based on composite kernel for solving classification tasks in gene function prediction. The proposed method realizes a nonlinear separating boundary by estimating a series of piecewise linear boundaries. Firstly, according to the distribution information of training data, a guided partitioning approach composed of separating boundary detection and clustering technique is used to obtain local subsets, and each subset is utilized to capture prior knowledge of corresponding local linear boundary. Secondly, a composite kernel is introduced to realize the local linear multi-SVM model. Instead of building multiple local SVM models separately, the prior knowledge of local subsets is used to construct a composite kernel, then the local linear multi-SVM model is realized by using the composite kernel exactly in the same way as a single SVM model. Experimental results on benchmark datasets demonstrate that the proposed method improves the classification performance efficiently.

Keywords—Multi-SVM model; local linear; composite kernel; prior knowledge; gene function classification.

I. INTRODUCTION

Support Vector Machines (SVMs) have been widely used in different application areas and become the state of the art. Different from most of conventional methods, SVM implements the Structural Risk Minimization (SRM) principle which seeks to minimize an upper bound of the generalization error rather than minimize the training error [1]. This eventually results in remarkable characteristics such as the good generalization performance, the absence of local minima and the sparse representation of solution.

By mapping the input data onto a higher-dimensional feature space in a non-linear fashion and seeking an optimal separating hyperplane in the feature space, SVMs can deal with linearly inseparable classification problems, and the feature mapping can be done implicitly through the kernel trick. Nonlinear SVMs employ sophisticated kernel functions, such as radial basis function (RBF), polynomial functions, etc., to fit datasets with complex decision surfaces. However, as many other non-linear classification methods, nonlinear kernel SVM models also face the potential overfitting issue when the number of training examples is small due to their large VC dimensions.

In functional genomics, there are some classification tasks (such as FunCat Yeast datasets [2] and Genbase motif-based datasets [3]), the nonlinear (RBF-kernel or other non-linear kernels) SVM models can not obtain good classification results due to overfitting. The linear SVM model was always selected as suboptimal solution in literatures [2], [3]. These tasks are nonlinear problems with characteristics of high noise and large number of input features compared with the relatively small number of training examples, the conventional nonlinear kernel SVM models are severely overfitting at times [2]. In this paper, we propose a local linear multi-SVM model based on composite kernel for solving these classification tasks.

A nonlinear separating boundary is approximately seen as an aggregation of piecewise linear boundaries. The proposed multi-SVM method realizes a nonlinear separating boundary by estimating a series of piecewise linear boundaries. According to the distribution information of training data, a guided partitioning approach composed of separating boundary detection and clustering technique is used to obtain local subsets from training data. By considering sample label changes in neighbor samples, the separating boundary detection can select the training samples which are near to separating boundary. The selected samples are partitioned into local subsets by Affinity Propagation (AP) clustering, and each subset is utilized to capture prior knowledge of corresponding local linear boundary. Then, a composite kernel is introduced to realize the local linear multi-SVM model by incorporating prior knowledge mined from each training subsets. Different from conventional multi-SVM methods with building multiple local SVM models separately, the prior knowledge of local subsets is used to construct a composite kernel, then the local linear multi-SVM model is realized by using the composite kernel exactly in the same way as a single SVM model.

The proposed method falls in the category of multiple linear SVM based on subspace pattern recognition. In some previous works, Ref. [4] presented Localized Support Vector Machine which builds multiple linear SVM models from training date and each model is designed to classify a particular test example. Ref. [5] introduced Mixtures of Linear SVMs by packaging linear SVMs into a probabilistic formulation and embedding them in the mixture of experts model. Different from those methods, in the proposed local linear multi-SVM, first, the subsets for capturing local linear knowledge are only clustered from samples which are near to separating boundaries rather than the original training data.
It can avoid trivial partitioning produced by unsupervised clustering. Second, a composite kernel incorporating prior knowledge is introduced to realize local linear multi-SVM model. Instead of building multiple local SVM models separately, the local subsets is used to estimate a composite kernel, and the composite kernel is utilized to realize the local linear multi-SVM model by implementing the structural risk minimization in the same way as a single standard SVM.

The rest parts of the paper are organized as follows. Section II formulates the linear multi-SVM model based on composite kernel. Section III describes estimation of the proposed local linear multi-SVM method. Section IV presents the Functional Genomics benchmark datasets and the results of experimental performance evaluation. Finally, the conclusions and future work directions are discussed.

II. LOCAL LINEAR MULTI-SVM MODEL BASED ON COMPOSITE KERNEL

In this paper, we focus our attention on binary classification problems. The classifier is built from a given labeled training dataset of N samples

\[(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_N, y_N)\]  

where \(x_i \in \mathbb{R}^d\) is the input vector corresponding to the \(i\)-th sample labeled by \(y_i \in \{-1, +1\}\) depending on its class.

As showed in Fig.1, a non-linear separating boundary can be seen as an aggregation of \(M\) piecewise linear local boundaries \(\Omega_j x + b_j, j = 1, \ldots, M\). According to the piecewise linear approximation method, the piecewise linear model \(f_P(x)\) can be written compactly as follow.

\[f_P(x) = \sum_{j=1}^{M} (\Omega_j^T x + b_j) R_j(x) + b\]  

where \(R_j(x)\)’s are the basis function, \(\Omega_j\)’s are the coordinate parameter vectors of local linear boundaries. The role of basis function is similar to that of a functional space basis. In some particular situations, they do constitute a functional basis. Typical examples are wavelet basis and RBF basis.

The overall performance of the the piecewise linear model is obtained via an interpolation using the basis function \(R(x)\). It also implies that the piecewise linear model \(f_P(x)\) can describe any sufficiently smooth nonlinear separating hyperplane function on a compact interval arbitrarily well by merely increasing the value of \(M\).

Two parameter vectors \(\Phi(x)\) and \(\Theta\) are defined as follows.

\[\Phi(x) = [R_1(x), x R_1(x), \ldots, R_M(x), x R_M(x)]^T\]

\[\Theta = [b_1, \Omega_1^T, \ldots, b_M, \Omega_M^T]^T\]  

Introducing parameter vectors \(\Phi(x)\) and \(\Theta\) (Eq. 3) into Eq. 2, the piecewise linear model \(f_P(x)\) can be rewritten as:

\[f_P(x) = \Theta^T \Phi(x) + b\]  

For the piecewise linear model \(f_P(x)\) (Eq. 4), we introduce the Structural Risk Minimization principle, it can be written as the QP optimization problem as

\[
\begin{align*}
\min_{\Theta, b, \xi} J_P & = \frac{1}{2} \Theta^T \Theta + c \sum_{k=1}^{N} \xi_k \\
\text{s.t.} & \quad y_k[\Theta^T \Phi(x_k) + b] \geq 1 - \xi_k, k = 1, \ldots, N \\
& \quad \xi_k \geq 0, k = 1, \ldots, N
\end{align*}
\]

The Lagrangian is constructed:

\[\mathcal{L}(\Theta, b, \xi; \alpha, v) = \mathcal{J}_P(\Theta, \xi) - \sum_{k=1}^{N} (\alpha_k y_k \Theta^T \Phi(x_k) + b - 1 - \xi_k) - \sum_{k=1}^{N} v_k \xi_k\]

with Lagrange multipliers \(\alpha_k \geq 0, v_k \geq 0\) for \(k = 1, \ldots, N\). The solution is given by the saddle point of the Lagrangian:

\[
\max_{\alpha, v} \min_{\Theta, b, \xi} \mathcal{L}(\Theta, b, \xi; \alpha, v)
\]

This leads to

\[
\begin{align*}
\frac{\partial \mathcal{L}}{\partial \Theta} & = 0 \implies \Theta = \sum_{k=1}^{N} \alpha_k y_k \Phi(x_k) \\
\frac{\partial \mathcal{L}}{\partial b} & = 0 \implies \sum_{k=1}^{N} \alpha_k y_k = 0 \\
\frac{\partial \mathcal{L}}{\partial \xi_k} & = 0 \implies 0 < \alpha_k < \infty, k = 1, \ldots, N
\end{align*}
\]

The dual problem becomes

\[
\max_{\alpha} J_D(\alpha) = \frac{1}{2} \sum_{k,l=1}^{N} \alpha_k \alpha_l y_k y_l \sum_{k=1}^{M} R_j(x_k) R_j(x_l) \]

\[
\text{s.t.} \quad \sum_{k=1}^{N} \alpha_k y_k = 0 \\
0 \leq \alpha_k \leq \infty, k = 1, \ldots, N
\]

In this quadratic form, the kernel trick is applied

\[K(x_k, x_l) = \Phi(x_k)^T \Phi(x_l) = (1 + x_k^T x_l) \sum_{j=1}^{M} R_j(x_k) R_j(x_l)\]

for \(k = 1, \ldots, N\). Hence, the piecewise linear model \(f_P(x)\) is reduced to a standard SVM based on a composite kernel (Eq. 10). Finally the nonlinear SVM classifier takes the form

\[y = \text{sign}\left[\sum_{k=1}^{N} \alpha_k y_k K(x_k, x) + b\right]\]

with \(\alpha_k\) positive real constants which are the solution to a QP problem.

From the optimal solution, we can find that it is not necessary to know the parameter vector \(\Phi(x)\) in expansion (Eq. 3) explicitly, as the solution only depends on the inner product which defines composite kernel \(K(x_k, x_l)\) (Eq. 10). In order to represent an inner product, the kernel is required to satisfy Mercer’s condition [1]. In Eq. 10, factors \(R_j(x_k), R_j(x_l)\) are outputs of basis functions. It follows that the kernel \(K_c(x_k, x_l)\) satisfies Mercer’s condition because it consists of a sum of products of Mercer kernels, cf. [1].
III. ESTIMATION OF THE LOCAL LINEAR MULTI-SVM MODEL

A. Affinity Propagation Clustering

Affinity Propagation (AP) [7] is a recently introduced clustering method which takes as input a set of measures of similarity between pairs of data points and outputs a set of clusters of the points with their corresponding exemplars. The algorithm takes a matrix of similarity measures between each pair of points $s(i, k)$ as input. Instead of requiring that the number of clusters be predetermined, the AP takes as input a real number $s(k, k)$ for each data point $k$. These values, which are called preferences, are a measure of how likely each point is to be chosen as exemplar.

The algorithm works by exchanging messages between the points until a stop condition is satisfied. There are two types of messages to be exchanged between data points. The responsibility $r(i, k)$, sent from data point $i$ to candidate exemplar point $k$, reflects the accumulated evidence for how well-suited point $k$ is to serve as the exemplar for point $i$, taking into account other potential exemplars for point $i$. The availability $a(i, k)$, sent from candidate exemplar point $k$ to point $i$, reflects the accumulated evidence for how appropriate it would be for point $i$ to choose point $k$ as its exemplar, taking into account the support from other points that point $k$ should be an exemplar.

The availabilities are initialized to zero: $a(i, k) = 0$. Then, the parameters are computed and updated using the rules as follows:

$$r(i, k) \leftarrow s(i, k) - \max_{k' \neq k, k' \neq k} \{a(i, k') + s(i, k')\} \quad (12)$$

$$a(i, k) \leftarrow \min\{0, r(k, k) + \sum_{i' \neq i, i' \neq k} \max\{0, r(i', k)\}\} \quad (13)$$

$$a(k, k) \leftarrow \sum_{i' \neq i, i' \neq k} \max\{0, r(i', k)\} \quad (14)$$

The AP method has been praised because of its ability to efficiently and quickly handle clustering problems.

B. Guided Partitioning Based on Separating Boundary Detection

In order to estimate parameters of the proposed model according to the distribution information of training data, clustering techniques can be used to partition training data into subsets. But our partitioning task is somewhat different from conventional unsupervised clustering. In order to capture the distribution information of local linear separating boundaries properly, the partitioning approach should ensure each subset contains training examples from two classes simultaneously. But conventional clustering techniques consider only the proximity between examples and often end up grouping training examples from the same class into the same cluster, because such clusters tend to be pure.

A guided partitioning approach composed of separating boundary detection and clustering technique is proposed to partition training data into subsets. A separating boundary detection considering sample label changes in neighbor area is used to select samples which are near to the separating boundary firstly. Then, the AP clustering is used to partition the selected samples into subsets for estimating parameters of the local linear multi-SVM model.

An example on synthetic data is used to formulate the proposed guided partitioning method. As showed in Fig. 2(a), a two-class dataset is constructed comprising 339 points in each class, the points are uniformly drawn from some shifted cosine signals in the two-dimensional space and perturbed with Gaussian noise in the vertical direction. Implementing unsupervised clustering technique onto the original data may produce trivial partitioning. The clustering results of implementing AP clustering (the cluster number is set as 3 empirically) onto the synthetic data (678 samples) is showed in Fig. 2(b), where samples from different clusters are plotted in different colors. We can find that the partitioning tends to grouping samples from the same class into the same cluster. The partitioned subsets can not be used to capture distribution information of local separating boundaries properly.

To solve above problem, in the proposed guided partition-
Figure 2. An example of guided partitioning approach. (a) the synthetic data (678 samples) constructed by shifted cosine signals with Gaussian noise; (b) the clustering results of implementing AP clustering onto the original synthetic data; (c) the selected samples (219 samples) by implementing the separation boundary detection; (d) the clustering results of implementing AP clustering onto the selected samples.

Figure 2. An example of guided partitioning approach. (a) the synthetic data (678 samples) constructed by shifted cosine signals with Gaussian noise; (b) the clustering results of implementing AP clustering onto the original synthetic data; (c) the selected samples (219 samples) by implementing the separation boundary detection; (d) the clustering results of implementing AP clustering onto the selected samples.

C. Estimation of the Local Linear Multi-SVM Model

According to the definition of the proposed local linear multi-SVM with composite kernel described in previous, basis functions is used to interpolate the piecewise linear hyperplanes, and each basis function corresponds to a local linear hyperplane. Therefore, the basis functions should abstract the distribution information of dataset as accurate as possible. We select a RBF gaussian function as the basis function to capture distribution information of obtained partitioning subsets.

\[ R(x) = e^{-\frac{(x - \mu)^2}{2\sigma^2}} \]  

where \( \mu \) is the subset center, \( \sigma \) the radius of subset and \( \lambda \) the scale parameter.

Another important parameter in the proposed model is the number of basis function \( M \). It is decided by the number of partitioning subsets (clusters). Theoretically, the proposed model \( f_P(x) \) (Eq. 2) can describe any sufficiently smooth nonlinear separating hyperplane function on a compact interval arbitrarily well by merely increasing the value of \( M \). If a proper parameter \( M \) is selected according to the data distribution (for example, select \( M = 3 \) for the cosine synthetic data showed in Fig. 2(a)), performance of the proposed model will be improved. In practice, the AP clustering used in the proposed model is an adaptive clustering approach, and it also can work by user-specified number of clusters.
IV. EXPERIMENTS AND RESULTS

A. Experiment DateSets

In functional genomics, an important problem is predicting the functions of genes (proteins). Currently, the number of protein sequences being stored in central protein databases from labs all over the world is constantly increasing. From these proteins only a fraction has been experimentally analyzed in order to detect their structure and hence their function in the corresponding organism. The reason is that experimental determination of structure is labor-intensive and quite time-consuming. Therefore there is the need for automated tools that can classify new proteins to structural families.

Generally, gene function predictions always are tasks of multi-label classification [8] or hierarchical multi-label classification [2], and SVMs is often used as efficient basis classifier because of their good generalization ability [2], [8], [9]. The purpose of this paper is to improve the basis classifier performance for each label (class), so we only implement per-class (binary classification) experiments for evaluating the proposed method. Two yeast datasets (“Sequence” and “All microarray”) from FunCat are used to evaluate the proposed method. The different datasets describe different aspects of the genes in the yeast genome, and the different sources of data highlight different aspects of gene function.

The properties of experiment datasets, including instance number $D$ and attribute number, are listed in Tab.I, the detailed description of each dataset can be referred from Ref. [10], [11]. The datasets are downloaded from the following webpage: http://dtai.cs.kuleuven.be/clus/hmcdatasets/. We select five classes (labels) in “protein fate” FunCat class (FunCat ID = 14) to implement per-class (binary classification) experiments, the descriptions of labels are listed in Tab.II.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Attribute</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence (seq)</td>
<td>478</td>
<td>2580</td>
<td>1339</td>
</tr>
<tr>
<td>All microarray (expr)</td>
<td>551</td>
<td>2488</td>
<td>1291</td>
</tr>
</tbody>
</table>

B. Evaluation Metrics

Three classical evaluation metrics of Precision, Recall and F-score are used to evaluate the efficiency of the proposed method. The three metrics are traditionally defined for a binary classification task with positive and negative classes. Precision is the proportion of positive predictions that are correct, and recall is the proportion of positive samples that are correctly predicted positive. That is:

$$
\text{Precision} = \frac{TP}{TP + FP} \\
\text{Recall} = \frac{TP}{TP + FN} \\
F - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
$$

with $TP$ the number of true positives (correctly predicted positive samples), $FP$ the number of false positives (positive predictions that are incorrect), and $FN$ the number of false negatives (positive samples that are incorrectly predicted negative). Note that these measures ignore the number of correctly predicted negative samples because we only concern positive samples in gene function prediction.

C. Experiment Setting and Results

In our experiments, the LSSVM [12] is taken as a basis. SVM parameters are chosen by cross-validation procedure. In guided partitioning approach, we set neighbor parameter $n = 5$ for separating boundary detection. About the parameters of AP clustering (defined in Ref. [7]), the maximum number of iterations is set as 1000, and early terminate parameter is set as 100 (i.e., the clustering procedure will be terminated if the estimated exemplars stay fixed for continuous 100 iterations); the damping factor, which may be needed if oscillations occur, is set as 0.9; Euclidean distance is used for the similarity metric of samples.

We compared our method with the RBF-kernel SVM and linear SVM. The experimental results for five labels on two datasets are presented in the Tab.III and Tab.IV. We can find that the RBF-kernel SVMs are severely overfitting, most of testing instances are predicted as negative and values of the recall metrics are very low. The experimental results of all evaluation metrics demonstrate that the proposed method improves the classification performance efficiently.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we present a local linear multi-SVM method based on composite kernel incorporating prior knowledge. The proposed method approximates a nonlinear separating boundary by estimating a series of piecewise linear boundaries. A guided partitioning approach composed of separating boundary detection and clustering technique is used to obtain local subsets from the training data. The separating boundary detection can select the training samples which are near to separating boundary by considering sample label changes in neighbor samples. The selected samples are partitioned into local subsets by AP clustering, and each subset is utilized to capture prior knowledge of corresponding...
Table III
RESULTS OF COMPARING WITH THE LINEAR SVM FOR FUN CAT “SEQUENCE” DATASET.

<table>
<thead>
<tr>
<th>FunCat ID</th>
<th>RBF-kernel SVM</th>
<th>Linear SVM</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.01</td>
<td>0.3963</td>
<td>0.0472</td>
<td>0.0843</td>
</tr>
<tr>
<td>14.04</td>
<td>0.5326</td>
<td>0.0825</td>
<td>0.1429</td>
</tr>
<tr>
<td>14.07</td>
<td>0.7253</td>
<td>0.0623</td>
<td>0.1147</td>
</tr>
<tr>
<td>14.10</td>
<td>0.4039</td>
<td>0.0772</td>
<td>0.1296</td>
</tr>
<tr>
<td>14.13</td>
<td>0.4532</td>
<td>0.0451</td>
<td>0.0820</td>
</tr>
</tbody>
</table>

Table IV
RESULTS OF COMPARING WITH THE LINEAR SVM FOR FUN CAT “ALL MICROARRAY” DATASET.

<table>
<thead>
<tr>
<th>FunCat ID</th>
<th>RBF-kernel SVM</th>
<th>Linear SVM</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.01</td>
<td>0.4092</td>
<td>0.1175</td>
<td>0.1826</td>
</tr>
<tr>
<td>14.04</td>
<td>0.4813</td>
<td>0.1281</td>
<td>0.2023</td>
</tr>
<tr>
<td>14.07</td>
<td>0.6947</td>
<td>0.1275</td>
<td>0.2155</td>
</tr>
<tr>
<td>14.10</td>
<td>0.4187</td>
<td>0.0984</td>
<td>0.1594</td>
</tr>
<tr>
<td>14.13</td>
<td>0.4382</td>
<td>0.0946</td>
<td>0.1566</td>
</tr>
</tbody>
</table>

local linear boundary. Then, a composite kernel is introduced to realize the local linear multi-SVM model by incorporating prior knowledge mined from each training subsets.

The experimental results demonstrate that the proposed method can solve the nonlinear classification tasks in gene function prediction, which conventional nonlinear kernel SVMs can not work efficiently. In future, we will implement the local linear multi-SVM model as basis classifier on the FunCat hierarchical multi-label classification datasets.

ACKNOWLEDGMENT

This research was supported by Waseda University Global COE Program 'International Research and Education Center for Ambient SoC’ sponsored by MEXT, Japan.

REFERENCES