Local Preserving Projections and Within-class Scatter Based Semi-supervised Support Vector Machines

Jun Gao  
The School of Information Engineering  
Yancheng Institute of Technology  
Yancheng, China  
gj0104211@163.com

Lili Xiang  
The School of Humanity  
Yancheng Institute of Technology  
Yancheng, China  
xllgj@ycit.cn

Abstract—The support vector machines (SVMs), as one of special regularization methods, has been used successfully in the field of pattern recognition. However, the traditional SVMs, a supervised learning method, gets the normal vector of the decision boundary mainly according to the largest interval law but has not taken the underlying geometric structure and the discriminant information into full consideration. Therefore, a local preserving projection and within-class scatter based semi-supervised support vector machine: LWSSVM, is presented in this paper by incorporating the basic theories of the locality preserving projections (LPP) and the linear discriminant analysis (LDA) in the SVMs. This method inherits the characteristics of the traditional SVMs, fully considers the global and local geometric structure between the samples and shows the global and local underlying discriminant information so that the classification accuracy can be increased. The tests on the face recognition datasets show the above mentioned advantages of the LWSSVM method.

Keywords—support vector machines; locality preserving projection; linear discriminant analysis; semi-supervised

I. INTRODUCTION

Support Vector Machine (SVM) is a new method in the field of data mining. It deals with machine learning problems by using statistical learning theories and optimization and has been successfully used in fields of pattern recognition. However, classical SVMs, supervised learning methods, can only be applied to a few of labeled samples, which leads to insufficient learning to certain extent. And since they do not take the underlying geometric structure and the discriminant information into full consideration, the recognition capability is not so effective to certain extent. Therefore, we incorporate Linear Discriminant Analysis (LDA) and Locality Preserving Projections (LPP) in SVMs and propose an efficient method: Locally Preserving Projections and Within-class Scatter Based Semi-supervise Support Vector Machines (LWSSVM), which has the advantages as follows: (1) it not only inherits the characteristics of classical SVMs but also correct the defect of insufficient learning to certain extent; (2) it is the first time that LDA and LPP are incorporated in SVMs. So the underlying local geometric structure of samples can be maintained to certain extent, showing local discriminant information in samples and at the same time, due to the advantages of LDA, it also keeps global geometric structure of samples, showing global discriminant information in samples; (3) it is easy to make nonlinear embedding and get a nonlinear method: Ker-LWSSVM, to find out manifold structure with high-dimension nonlinear characteristics.

II. LOCAL PRESERVING PROJECTIONS AND WITHIN-CLASS SCATTER BASED SEMI-SUPERVISED SUPPORT VECTOR MACHINES

From Reference works [4], we know that within-class scatters in the LDA method can maintain to certain extent global geometric structure and discriminant information of samples. So it is reasonable to incorporated within-class scatter matrix of LDA in LWSSVM proposed in this paper in order to maintain internal global geometric structure to certain extent.

Definition 1[2]: Suppose sample set \( D = \{ x_1, \ldots, x_l \} \) consists of \( l \) samples, \( \forall x_i \in R^n \). They belong to 2 different classes: \( C, C' \), in which sample subset \( D_i \) with \( l_i \) samples belongs to class \( k \). Let \( \omega \) be direction vector of classifying decision boundary and \( \omega'S,\omega \) be within-class scatter. We define

\[
S_k = \sum_{i=1}^{l_k} (x_i - u_k)(x_i - u_k)^T
\]

as within-class matrix

\[
u_k = \frac{1}{l_k} \sum_{i=1}^{l_k} x_i(k=1,2)\text{as mean of } D_i
\]

Definition 2: Let us take \( X = \{ x_1, \ldots, x_n \} \) as a sample set which consists of \( l \) labeled samples belonging to 2 different classes: \( C, C' \) and \( u \) unlabeled samples. Let \( \omega' S,\omega \) be within-class scatter corresponding to \( l \) labeled samples and \( L \) be Laplacian matrix of data set \( X \). We define the optimization problem of LWSSVM as:

\[
\min_{w,\beta} \sum_{i=1}^{l} \xi_i + \frac{\gamma_+}{2} S_w \omega \omega' S_w \omega + \frac{\gamma_-}{2} \omega' X \omega + \beta^T \left( \gamma \right)
\]

s.t. \( y_i ((\omega, x_i) + b) \geq 1 - \xi_i, i = 1, \ldots, l \)

where \( \gamma_+ \geq 0, \gamma_- \geq 0 \)

Theorem 1: The dual problems for original optimization problems (3) and (4) of linear LWSSVM can be defined as follows:

\[
\min_{\beta} \frac{1}{2} \beta^T H \beta - \beta^T \beta
\]
Based on Karush-Kuhn-Tucker (KKT) [5] conditions and Lagrange function, we can obtain director vector and bias of the decision boundary as follows, respectively:

\[
\frac{\partial L}{\partial \omega} = 0
\]
\[
\Rightarrow \omega = \frac{1}{l} \sum_{j=1}^{l} \beta_j \gamma_j S_{w_j} + \gamma_j x_j \L^T x_j \right) \omega
\]

\[
b = \frac{1}{l} \sum_{j=1}^{l} \gamma_j x_j \left( \gamma_j S_{w_j} + \gamma_j x_j \L^T x_j \right) x_j \right)
\]

where \((.)^{-1}\) represents generalized inverse matrix.

Therefore, we get following linear LWSSVM algorithm:

**Input:** \(l\) labeled samples \(\{x_i, y_i\}_{i=1}^{l}\), \(u\) unlabeled samples \(\{x_k\}_{k=1}^{u}\).

**Output:** classifying decision function \(g(x)\).

**Step 1:** Construct the within-class scatter matrix \(S_w\) by substituting \(\{x_i, y_i\}_{i=1}^{l}\) into function in definition (1).

**Step 2:** Construct the adjacency graph \(G\) with \(l+u\) nodes by using \(k\)-NN and compute to get a weight matrix \(W\).

**Step 3:** Compute Laplacian matrix \(\mathbf{L} = \mathbf{D} - \mathbf{W}\) in which \(\mathbf{D}\) is a diagonal matrix and \(D_{ii} = \sum_{j=1}^{l} W_{ij}\).

**Step 4:** Define parameters \(\gamma, \gamma_i\) and compute Lagrange coefficient \(\beta\) by using Theorem 1.

**Step 5:** Compute director vector \(\omega\) of decision boundary by using equation (7).

**Step 6:** Compute bias \(b\) by using equation (8).

**Step 7:** Get classifying function \(g(x) = \omega^T x + b\).

Note that, from equation (7), it can be found that the director vector \(\omega\) of decision boundary depend on not only labeled samples but also unlabeled ones, which is of great significance in constructing nonlinear LWSSVM algorithm.

III. NONLINEAR LWSSVM: KER-LWSSVM

As to samples with internal geometric structure being high-dimension nonlinear manifolds, they cannot be dealt with by linear LWSSVM. In order to solve this problem, we propose a nonlinear Ker-LWSSVM algorithm. After analysis, we find it necessary to deal with within-class scatter matrix correspondingly in the process of developing the linear LWSSVM algorithm into a non-linear one. Therefore, we convert the within-class scatter matrix \(S_w\) into a Graph Laplacian matrix by the method in the reference works [4].

According to Definition 1, given \(l\) labeled samples \(x_i = \{x_i\}_{i=1}^{l}\) belong to 2 different classes in which the number of samples in each class is \(l_i\), the weight matrix \(W\) of the \(k\) nearest neighbor adjacency graph \(G\) corresponding to \(l\) samples can be defined as follows:

\[
W' = \begin{cases} 
1/k & \text{if } x_i \text{ and } x_j \text{ both belong to } k \text{ class} \\
0 & \text{otherwise}
\end{cases}
\]

Then the Graph Laplacian matrix of \(S_w\) is \([4]\):

\[
S_w = \mathbf{X}_l \L \mathbf{X}_l^T
\]

where \(\mathbf{L} = \mathbf{I} - \mathbf{W}, \mathbf{I}\) is identity matrix.

We can construct original optimization problem of the nonlinear Ker-LWSSVM algorithm based on above analysis.

**Theorem 2:** Following are the optimization problems in the non-linear Ker-LWSSVM algorithm:

\[
\min_{a, b, \xi, \lambda} \frac{1}{2} \alpha^T K \alpha + \gamma \sum_{i=1}^{l} \xi_i + \frac{1}{2} \lambda^T K \lambda
\]

\[
s.t. \; y_i \sum_{j=1}^{l} a_j K(x_i, x_j) + b \geq 1 - \xi_i, i = 1, \ldots, l
\]

where \(K(x, y)\) denotes Mercer kernel function. Given \(l\) labeled samples \(x_i, u\) unlabeled samples and \(x_i = (X, X)\), we denote the projections of the above data onto the feature space by using a non-linear function \(\phi\) as \(X, X\) and \(X = (X, X)\), then \(K = \phi(X, X)\) and \(K = \phi(X, X)\).

Proof: Through analyzing the direction vector \(\omega\) in linear LWSSVM algorithm, we know that it depends on not only labeled samples but unlabeled ones. And by incorporating Representative Theorems [5], the direction vector of the nonlinear decision boundary in feature space can be defined as \(\omega^* = \sum_{i} a_i \phi(x)\) where \(a = (a_1, \ldots, a_u)^T\) is weight vector. We can denote regularized units as following:

\[
\omega^T S_w \omega^* = a^T X_{l} X_{l}^T X_{l} a = a^T K L_a K a
\]

\[
\omega^T \mathbf{L}_a \mathbf{X}_a^T \omega^* = a^T \mathbf{X}_a^T \mathbf{L}_a \mathbf{X}_a^T \omega^* = a^T K \mathbf{L}_a K \mathbf{a}
\]

where \(\mathbf{K}\) is a \((l+u)\times(l+u)\) matrix and \(\mathbf{K}\) is \((l+u)\times(l+u)\).

So this theorem proves reasonable by instituting the above equation (13) into the equations (3) and (4).

**Theorem 3:** the dual problems of the original optimization problems (11) and (12) in the nonlinear Ker-LWSSVM algorithm are

\[
\min \frac{1}{2} \beta^T H_{\beta} \beta - \mathbf{1}^T \beta
\]

\[
s.t. \; \sum_{j=1}^{l} y_j \beta_j = 0, \; 0 \leq \beta_j \leq \frac{1}{l}
\]

where \(H_{\beta} = (K \mathbf{L} K + \gamma K \mathbf{L} K) \mathbf{K} Y\),

\(Y = \mathbf{diag}(y_1, \ldots, y_l)\) and \(\mathbf{1} = (1, \ldots, 1)\).

Compute director vector and bias of the nonlinear decision boundary by using KKT conditions and Lagrange function:

\[
\frac{\partial L}{\partial \alpha} = 0 \Rightarrow \alpha = (\gamma K \mathbf{L} K + \gamma K \mathbf{L} K) \mathbf{K} Y \beta
\]
Therefore, we obtain nonlinear Ker-LWSSVM through Theorem 2 and Theorem 3:

\[
b_w = \frac{1}{l} \sum_{i=1}^{l} (y_i - \sum_{j=1}^{u} \alpha_j K(x_i, x_j))
\]

(17)

IV. Experiments

In order to show the performance of LWSSVM methods proposed in this paper in terms of keeping the internally global and local features, we tested on biological data sets (http://www.ics.uci.edu/~mlearn/MLRepository.html) and face recognition data sets (http://www.cs.uiuc.edu/homes/dengcai2) and compared the outcomes of them with those of other methods. In the procedure of testing, we used cross-validation to choose parameters aiming at obtaining the most reasonable outcomes.

A. UCI datasets

Sonar and Heart sets, biological datasets, in UCI database are often used to testify classification accuracy of classifiers. We performed experiments on both Sonar and Heart datasets to test linear and non-linear SDA, linear and non-linear Lap-SVM, LWSSVM and Ker-LWSSVM. The former consists of 135 training samples with 20 labeled samples and 115 unlabeled ones and 70 forecast samples. The latter 180 training samples with 60 labeled samples and 120 unlabeled ones and 90 forecast samples. At the same, we used 5-fold cross-validation to choose parameters. That is, the range of the corresponding parameters of Lap-SVMs, LWSSVM and Ker-LWSSVM was \( k = [1, 5, 15, 40, 60] \) and \( \sigma = \gamma = \gamma = 2^{-7}, 2^{-3}, 3, 5, 7, 2^{10} \), while that of the \( k \) nearest neighbor parameter in SDA was \( k = [1, 5, 15, 40, 60] \) and \( \sigma = \alpha = 2^{-7}, 2^{-3}, 3, 5, 7, 2^{10} \). Also SDA used 1-NN classifiers. Table I shows the outcomes of the experiments.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>SDA</th>
<th>Lap-SVMs</th>
<th>LWSSVM</th>
<th>Ker-LWSSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear</td>
<td>Non-linear</td>
<td>Linear</td>
<td>Non-linear</td>
</tr>
<tr>
<td>heart</td>
<td>0.6667</td>
<td>0.7667</td>
<td>0.7889</td>
<td>0.8778</td>
</tr>
<tr>
<td>sonar</td>
<td>0.6857</td>
<td>0.7285</td>
<td>0.7</td>
<td>0.82357</td>
</tr>
</tbody>
</table>

Table I indicates that LWSSVM and Ker-LWSSGM proposed in this paper are of better classification effectiveness compared with SDA and Lap-SVMS, which proves that the methods in this paper take underlying global structure between the samples into full consideration and show global discriminant information when keeping the underlying local discriminant information.

B. Facial Expression Recognition

The face recognition datasets have obvious non-linear manifold structures. So they are used to test a lot of manifold learning methods [6-10]. We compare our methods with nonlinear SDA and Lap-SVM to show the effectiveness of the methods in this paper when face recognition datasets with manifold structure are classified.

IV. Experiments

In order to show the performance of LWSSVM methods proposed in this paper in terms of keeping the internally global and local features, we tested on biological data sets (http://www.ics.uci.edu/~mlearn/MLRepository.html) and face recognition data sets (http://www.cs.uiuc.edu/homes/dengcai2) and compared the outcomes of them with those of other methods. In the procedure of testing, we used cross-validation to choose parameters aiming at obtaining the most reasonable outcomes.

A. UCI datasets

Sonar and Heart sets, biological datasets, in UCI database are often used to testify classification accuracy of classifiers. We performed experiments on both Sonar and Heart datasets to test linear and non-linear SDA, linear and non-linear Lap-SVM, LWSSVM and Ker-LWSSVM. The former consists of 135 training samples with 20 labeled samples and 115 unlabeled ones and 70 forecast samples. The latter 180 training samples with 60 labeled samples and 120 unlabeled ones and 90 forecast samples. At the same, we used 5-fold cross-validation to choose parameters. That is, the range of the corresponding parameters of Lap-SVMs, LWSSVM and Ker-LWSSVM was \( k = [1, 5, 15, 40, 60] \) and \( \sigma = \gamma = \gamma = 2^{-7}, 2^{-3}, 3, 5, 7, 2^{10} \), while that of the \( k \) nearest neighbor parameter in SDA was \( k = [1, 5, 15, 40, 60] \) and \( \sigma = \alpha = 2^{-7}, 2^{-3}, 3, 5, 7, 2^{10} \). Also SDA used 1-NN classifiers. Table I shows the outcomes of the experiments.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>SDA</th>
<th>Lap-SVMs</th>
<th>LWSSVM</th>
<th>Ker-LWSSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear</td>
<td>Non-linear</td>
<td>Linear</td>
<td>Non-linear</td>
</tr>
<tr>
<td>heart</td>
<td>0.6667</td>
<td>0.7667</td>
<td>0.7889</td>
<td>0.8778</td>
</tr>
<tr>
<td>sonar</td>
<td>0.6857</td>
<td>0.7285</td>
<td>0.7</td>
<td>0.82357</td>
</tr>
</tbody>
</table>

Table I indicates that LWSSVM and Ker-LWSSGM proposed in this paper are of better classification effectiveness compared with SDA and Lap-SVMS, which proves that the methods in this paper take underlying global structure between the samples into full consideration and show global discriminant information when keeping the underlying local discriminant information.

B. Facial Expression Recognition

The face recognition datasets have obvious non-linear manifold structures. So they are used to test a lot of manifold learning methods [6-10]. We compare our methods with nonlinear SDA and Lap-SVM to show the effectiveness of the methods in this paper when face recognition datasets with manifold structure are classified.
Table II indicates that Ker-LWSSVM proposed in this paper are of better classification effectiveness which proves that the method in this paper, when dealing with face recognition datasets of obvious manifold structure, not only keep local manifold structure to certain extent but also make sure that the within-class scatter is minimized, thus maintaining the underlying global and local discriminant information of face recognition datasets.

V. CONCLUSION AND FUTURE WORK

We propose a global and local preserving based semi-supervised support vector machine in this paper by incorporating the basic theories of LDA and LPP after analyzing the shortcomings of classical SVMs. It overcomes the shortcoming of traditional SVMs, fully considers the global and local geometric structure between the samples and shows the global and local underlying discriminant information. However, each method has its own advantages and disadvantages. Since within-class scatter is incorporated in it, it calls for certain amount of labeled samples when dealing with real problems and increases the time and space complexity to certain extent, which will be our future work.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper in advance. This work was partly supported by the Jiangsu Provincial Innovation Project of Graduate Students of China (CX09B-175Z) and Yancheng Institute of Technology Project of Applied Basic Research (XKY2009070).

REFERENCES