In this work we introduce a new methodology to build a kernel matrix from a collection of kernels. The key idea is to build an unique kernel that eliminates spurious differences between kernels. We propose a method based on the Procrustes problems that uses the Alternating Projections method to minimize a certain error measure. The resulting kernel will be used for classification purposes using Support Vector Machines (SVMs). The proposed method has been successfully evaluated against alternative kernel combination techniques.

1. Introduction

Support Vector Machines (SVMs) have proven to be a successful tool for the solution of a wide range of classification problems since their introduction in [2]. The method uses as a primary source of information a kernel matrix \( K \). Using a single kernel may be not enough to solve accurately the problem under consideration. This happens, for instance, when dealing with text mining problems, where analysis results may vary depending on the document similarity measure chosen [9]. Thus, information provided by a single kernel may not be enough for classification purposes, and different kernels should be used to solve the problem. The combination of kernels has been proposed as an alternative to the choice of the best kernel. It is common to use the average of the kernels method (AKM), or to look for a more complex linear combination of the kernels obtained as solution of a semi-definite programming (SDP) [11].

Nevertheless, it is not always clear when the resulting kernel will perform better than its components considered separately. If the individual kernels to be combined achieve similar classification performance (in a SVM context), while their support vectors are different, then the answer to the preceding questions is yes (see [10] for details). In other cases, it is difficult to find a linear combination of kernels that improves the classification performance of the individual kernels. Our method will be develop in this context. The idea is to build a kernel matrix that eliminates the (non-relevant) differences of information between the individual kernels.

To motivate our approximation consider the following example: there is some message transmitted between computers. There are transmission errors that imply noise on the information: differences between the message that is sent from the original computer and the message that is received at the final computer appear. This process is replicated in order to minimize errors importance. Once the receiver have picked up all the information, it is necessary to recover the original message. That is, there are several information matrices that should convey the same information. In such a situation, we would like to eliminate the differences of information between the matrices. We propose a method based on the concepts of Procrustes problems that will use the Alternating Projections algorithm.

The rest of the paper is organized as follows. In Section 2 we summarize the Procrustes methods and propose a new method to modify a kernel matrix to be as similar as possible to a given kernel matrix. Section 3 describes the proposed method for combining kernels. The experimental setup and results on real data sets are described in Section 4. Finally, Section 5 concludes.

2. Procrustes Problems

Let \( K_1, K_2 \in \mathbb{R}^{l \times l} \) two similarity (kernel) matrices for a training data set of \( l \) objects. The Procrustes method [6] modifies one of these matrices in order to get a matrix as similar as possible to the other one in the Frobenius norm. The method uses a transformation matrix \( Q \) to minimize the following expression:

\[
||K_1 - K_2 Q||_F^2. \tag{1}
\]

Procrustes was first used to relate factorial analysis with a labels vector [8] (with a precedent in [12]). In particular, if
Q is constraint to be orthogonal in (1), the problem has been called ‘Orthogonal Procrustes problem’ (OP). The final matrix \( K_2^Q \) is necessarily a Mercer kernel [7]. There is an alternative based on the OP problem which can be used to build Mercer kernels:

\[
\min_Q \| K_1 - Q^T K_2 Q \|_F^2 \\
\text{s.t.} \quad Q \text{ orthogonal.}
\]  

(2)

In this situation, we look for an orthogonal matrix \( Q \) such that \( Q^T K_2 Q \) will be as similar as possible to the original matrix \( K_1 \). Since \( Q \) is orthogonal, \( Q^T K_2 Q \) is a symmetric matrix and it is positive semidefinite (psd) if \( K_2 \) is psd. This method is known as ‘Double Procrustes problem’ (DP) [13]. However, this method can not be used to build a kernel for classification purposes. Consider a new row to be classified, that is a new row (and column) in the matrices \( K_1 \) and \( K_2 \). To know the corresponding value in the new matrix \( Q^T K_2 Q \), it is necessary to solve problem (2) again. We would need to calculate a different hyperplane (different \( Q \) solutions) for each new point in the new sample.

2.1 Conformal Transformation

As an alternative to the Procrustes methods we propose the Conformal Transformation (CT) method:

\[
\min_W \| K_1 - W K_2 W^T \|_F^2 \\
\text{s.t.} \quad W \text{ diagonal.}
\]  

(3)

We look for the highest similarity between two matrices by using a new diagonal matrix \( W \). Since \( W \) is diagonal, it is clear that \( W K_2 W \) is psd if \( K_2 \) is psd. We propose a simple and iterative method to solve (3); to modify the elements of a original \( W \) (for instance, \( W = I \)) iteratively, only when it improves the approximation between the matrices \( W K_2 W \) and \( K_1 \), until convergence. It is direct to prove that for a new point \( x \) in the sample, the corresponding diagonal element of \( W \) \((w_x = W(x, x))\) is:

\[
w_x = \frac{\sum_{i=1}^I w_i K_1(x, i) K_2(x, i)}{\sum_{i=1}^I (w_i K_2(x, i))^2}.
\]  

(4)

That is, we have a transformation of a matrix \( K_2 \) defined by \( K_2(i, j) \leftarrow K_2(i, j) w_i w_j \) to be as similar as possible to other matrix \( K_1 \). Next, we propose the Kernel Procrustes method as a generalization of the CT method for the case of more than two kernel matrices.

3 Kernel Procrustes

Let \( K_1, \ldots, K_M \) a set of \( M \) psd matrices in \( \mathbb{R}^{I \times I} \). We want to find a conformal transformation of each matrix in the set, such that the distance between the new matrices is minimum:

\[
\min_{W_1, \ldots, W_M} \sum_{m=1}^M \| K_m - W_m K_m W_m^T \|_F^2 \\
\text{s.t.} \quad W_m \text{ diagonal, } \forall m = 1, \ldots, M.
\]  

(5)

The objective function in (5) can be written as deviations to an unique matrix as follows:

\[
M \sum_{m=1}^M \| \tilde{K}_{YY} - W_m K_m W_m^T \|_F^2,
\]  

(6)

where \( \tilde{K}_{YY} (V = \{ W_m \}_{m=1}^M) \) is the mean of the original matrices \( K_1, \ldots, K_M \), after transforming them by \( W_1, \ldots, W_M \), respectively:

\[
\tilde{K}_{YY} = M^{-1} \sum_{m=1}^M W_m K_m W_m.
\]  

(7)

It is necessary to define a constrain on \( W_m \) to avoid the trivial solution \( W_m = 0 \), \( \forall m = 1, \ldots, M \) in (5). We suggest to add a constraint on the average of the matrices: the diagonal of the final matrix should be equal to the diagonal of the original average of the matrices. If we are dealing with similarity matrices, in general, the diagonal elements equate 1 (0 if the matrices under consideration are distances between observations). To keep unchanged the diagonal controls the differences between all the elements of the matrices. The problem to be solved, called ‘Kernel Procrustes problem’ (KP), is:

\[
\min_{W_1, \ldots, W_M} \sum_{m=1}^M \| \tilde{K}_{YY} - W_m K_m W_m \|_F^2 \\
\text{s.t.} \quad \text{diag}(\tilde{K}_{YY}) = \text{diag}(\tilde{K}_F), \quad W_m \text{ diagonal, } \forall m = 1, \ldots, M,
\]  

(8)

where \( \tilde{K}_F \) is the average of the original matrices.

To solve (8) we propose the use of the Alternating Projections method (see [3] for details). Define the sets:

\[
S = \{ K \in \mathbb{R}^{I \times I} : K = K^T, K = \tilde{K}_{YY}, W = \{ W_m \}_{m=1}^M \},
\]

\[
U = \{ K \in \mathbb{R}^{I \times I} : K = K^T, \text{diag}(K) = \text{diag}(\tilde{K}_F) \},
\]

where each \( W_m \) in \( S \) is obtained as solution of the following CT problem:

\[
\min_{W_m} \| K - W_m K_m W_m^T \|_F^2 \\
\text{s.t.} \quad W_m \text{ diagonal,}
\]  

(9)

where \( K \) is a given matrix, and \( U \) is the set of symmetric matrices with diagonal elements equal to the diagonal elements of the original average of kernels. The solution to (8)
is a matrix in the intersection of $S$ and $U$. To find it, we project on these sets respectively until convergency:

$$K \leftarrow P_U(P_S(K)).$$  \hfill (10)

Given a symmetric matrix $K$, $P_U(K)$ implies a change in the diagonal elements of $K$ to be equal to $\text{diag}(\bar{K})$ and $P_S(K) = M^{-1}\sum_{m=1}^{M} W_m K_m W_m$ where $W_m$ is the solution to (9).

We propose algorithm 1 to solve the KP problem.

<table>
<thead>
<tr>
<th>Algorithm 1: Kernel Procrustes.</th>
</tr>
</thead>
</table>
| Let $K_1, \ldots, K_M$ a set of $M$ psd matrices in $\mathbb{R}^{t \times t}$. Initialize $\Delta S_0 = 0$, $Y_0 = K$. Repeat $k = k + 1$, $R_k = Y_{k-1} - \Delta S_{k-1}$, where $\Delta S_{k-1}$ is Dykstra’s correction, $X_k = P_S(R_k)$, that is:
| for $m=1,\ldots,M$, evaluate $W_m$
| $\min_{W_m} \| R_k - W_m K_m W_m \|_F^2$
| $s.t.$ $W_m$ diagonal.
| end estimate $X_k = M^{-1}\sum_{m=1}^{M} W_m K_m W_m.$
| $\Delta S_k = X_k - R_k.$
| $Y_k = P_U(X_k)$.
| Until convergency, that is, the changes in $X_k$ and $Y_k$ are not significative. |

4. Experiments

4.1 Human Skulls

The data set consists of 150 observations corresponding to human skulls from the ancient Egypt [14]. There are 4 morphological variables: skull width ($X_1$), skull height from base ($X_2$), skull length ($X_3$) and nasal height ($X_4$).

We design the following experiment: consider 4 RBF kernel functions, $K_m(x,z) = \exp(-\|x - z\|^2/c)$, $m = 1, \ldots, 4$ with $c = 0.01$. Each kernel is defined using three out of the four variables: the first kernel function, $K_1$, is designed using the information on $X_2$, $X_3$ and $X_4$, the second kernel, $K_2$, is defined on $X_1$, $X_3$ and $X_4$, etc. It is necessary to use at least two kernels to take into account all the variables in the sample.

We apply the Procrustes and KP methods to get an unique matrix from the four kernel matrices in the example. Let $K_m^*$, $m = 1, \ldots, 4$, be the transformed $K_m$ matrix, that is: $K_m^* = K_m$ if the AKM method has been used ($Q_m = I$), $K_m^* = K_m Q_m$ if we use the OP method, and $K_m^* = Q_m^T K_m Q_m$ if we use the DP method, and $K_m^* = M^{-1}\sum_{m=1}^{M} K_m^*$ and $K^* = M^{-1}\sum_{m=1}^{M} K_m^*$. We suggest to stop the algorithm when the following quantities: $\text{error} = \sum_{m=1}^{M} \| \bar{K}^* - K_m^* \|^2_F$, where \( \bar{K}^* = M^{-1}\sum_{m=1}^{M} K_m^* \) and $\text{relative error} = \frac{\text{error}}{\| K^* \|^2_F}.$

Let $K$ the RBF kernel matrix defined using all the variables in the data set. This is the ‘solution’ matrix we would like to recover once we have the information from the four kernel functions. Define the ‘approximation’ to this matrix as:

$$\| K - \bar{K}^* \|_F^2,$$

and the ‘relative approximation’ as:

$$\sqrt{\| K \|_F^2 / \| \bar{K}^* \|_F^2}.$$

The results on 10 replications of the experiment are presented in Table 1. In each replication, we select a random sample of 75 observations.

<table>
<thead>
<tr>
<th>Method</th>
<th>Error</th>
<th>Relative Error</th>
<th>Approx. Error</th>
<th>Relative Approx. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKM</td>
<td>1758.65</td>
<td>1.390</td>
<td>67.98</td>
<td>0.064</td>
</tr>
<tr>
<td>OP</td>
<td>1327.89</td>
<td>1.027</td>
<td>75.06</td>
<td>0.070</td>
</tr>
<tr>
<td>DP</td>
<td>103.54</td>
<td>0.075</td>
<td>86.67</td>
<td>0.079</td>
</tr>
<tr>
<td>KP</td>
<td>1140.94</td>
<td>0.987</td>
<td>38.20</td>
<td>0.038</td>
</tr>
</tbody>
</table>

4.2 A Handwritten Digit Recognition Problem

The experiment in this section concerns a binary classification problem: the recognition of digits ‘7’ and ‘9’ from
the Alpaydin and Kaynak database [1]. The data set is made up by 1128 records, represented by 32 × 32 binary images. We have employed three different methods to specify features in order to describe the images. The first one is the 4 × 4 method: features are defined as the number of ones in each of the 64 squares of dimension 4 × 4. The second method was introduced by Frey and Slate [4]: 16 attributes are derived from the image, related to the horizontal/vertical position, width, height, etc. The last method under consideration was designed by Fukushima and Imagawa [5]: features are defined as a collection of 12 different representations in a 4 × 4 square. We have used these representations to calculate three kernels (\(K_1\), \(K_2\), \(K_3\)) using the Euclidean distance. This is a typical example with several sources of information.

We compare the individual support vectors from the SVMs obtained using each kernel (Table 2). A priori, it is not expected a great improvement on the classification results by using a linear combination of the kernels. The relative errors, defined in (13), are 0.339 and 0.327 for the AKM and KP methods, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>Support Vectors</th>
<th>Common S.V.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(K_1)</td>
<td>28.0</td>
<td>11.0 (2.7)</td>
</tr>
<tr>
<td>(K_2)</td>
<td>75.0</td>
<td>21.0 (2.0)</td>
</tr>
<tr>
<td>(K_3)</td>
<td>54.0</td>
<td>21.0 (5.3)</td>
</tr>
</tbody>
</table>

The classification performance for all the methods is tabulated in Table 3. The best classification results are achieved by the KP method. The KP method is the only one that improves the best of the original kernels.

### 5 Conclusions

A new technique to combine kernel matrices has been presented based on the ideas of the Procrustes problem. The Kernel Procrustes method is based on the use of a proper transformation of the original kernel matrices to get kernels as similar as possible to the average kernel matrix. Experimental results support that the method can be a good alternative to other combination of kernels methods. Further research will focus on theoretical properties of the method and extensions.

#### Table 3. Classification errors and support vectors (%) for the handwritten digit recognition data set.

<table>
<thead>
<tr>
<th>Method</th>
<th>Train</th>
<th>Test</th>
<th>S.V.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(K_1)</td>
<td>0.00</td>
<td>3.62</td>
<td>3.64</td>
</tr>
<tr>
<td>(K_2)</td>
<td>5.46</td>
<td>11.14</td>
<td>9.75</td>
</tr>
<tr>
<td>(K_3)</td>
<td>0.00</td>
<td>4.46</td>
<td>7.02</td>
</tr>
<tr>
<td>AKM</td>
<td>0.00</td>
<td>4.46</td>
<td>6.11</td>
</tr>
<tr>
<td>SDP</td>
<td>0.00</td>
<td>3.62</td>
<td>6.24</td>
</tr>
<tr>
<td>KP</td>
<td>0.26</td>
<td>3.34</td>
<td>16.64</td>
</tr>
</tbody>
</table>

Acknowledgments. This work was partially supported by Spanish grant SEJ 2004-03303.

### References