ABSTRACT

We applied a multiple kernel learning (MKL) method based on information-theoretic optimization to speaker recognition. Most of the kernel methods applied to speaker recognition systems require a suitable kernel function and its parameters to be determined for a given data set. In contrast, MKL eliminates the need for strict determination of the kernel function and parameters by using a convex combination of element kernels. In the present paper, we describe an MKL algorithm based on conditional entropy minimization (MCEM). We experimentally verified the effectiveness of MCEM for speaker classification; this method reduced the speaker error rate as compared to conventional methods.

Index Terms— Multiple kernel learning, MCEM, speaker recognition

1. INTRODUCTION

Kernel-based methods such as support vector machines (SVMs) are frequently applied to speaker recognition systems[1, 2]. However, these methods require a suitable kernel function and kernel parameters to be determined for a given data set. This is indeed troublesome work. One of the approaches to solve this problem is to adopt multiple kernel learning (MKL)[3, 4], in which a combination of several kernels are used for a given data set. Recently, an MKL method that combines complementary derivative and parametric kernels has been applied for speaker verification[5].

Most MKL methods, including that presented in [5], use maximum-margin-based schemes for learning both the parameter of a linear classification function and the combination coefficients of kernels. We have proposed an MKL algorithm based on conditional entropy minimization (MCEM), which in contrast minimizes class-conditional entropy with respect to the aforementioned parameters[6]. We found that MCEM outperforms the margin-optimization-based MKL methods in experiments using simulated data.

In the present study, we verify the effectiveness of MCEM in solving the speaker classification problem by comparing the performance of an MCEM-based system with that of a conventional maximum-margin-based system, using the same kernel functions and kernel parameters. The information-theoretic optimization methods such as MCEM can be used for analyzing qualitative properties such as convergence, robustness, efficiency, and consistency, on the basis of information geometry. Furthermore, MCEM may improve the robustness of speaker recognition systems to the noise around discriminative planes, because this method does not focus on constructing a discriminative plane itself but attempts to find a subspace in which the data can be compactly aggregated for each class.

The rest of this present paper is organized as follows. Section 2 briefly reviews the existing MKL methods. In Sect. 3, the adequacy of applying the conditional entropy minimization criterion to MKL and the MCEM algorithm are described. In Sect. 4, MCEM is experimentally evaluated in terms of speaker recognition accuracy. Finally, in Sect. 5, some concluding remarks and future works are presented.

2. MULTIPLE KERNEL LEARNING

In this section, the existing MKL methods are briefly reviewed. Assume that a set of observed samples \( D = \{ x_i \}_{i=1}^N \) and their class labels \( \{ y_i \}_{i=1}^N, y_i \in \{ \pm 1 \} \) are given. With a classification function of the form

\[
f(x) = \sum_{i=1}^N \alpha_i k(x, x_i),
\]

where \( k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle \) is a kernel function associated with a feature map \( \Phi \), a typical kernel-based learning method is represented as follows:

\[
\min_{\alpha} \left[ \sum_{i=1}^N \mathcal{L}(y_i, \sum_{j=1}^N \alpha_j k(x_i, x_j)) + \| \alpha \|_K^2 \right],
\]

where \( \mathcal{L} \) is the loss function, e.g., hinge loss \( \mathcal{L}(y, t) = \max(0, 1 - yt) \) for SVMs. Here, the kernel function is represented in terms of a kernel matrix, i.e., \( [K]_{ij} = k(x_i, x_j) \). In MKL, the kernel matrix \( K \) is replaced with \( \sum_{s=1}^S \beta_s K_s, \sum_{s=1}^S \beta_s = 1, \beta_s \geq 0 \), using \( S \) kernel matrices. In this case, an attempt is made to optimize the loss over \( \beta = \{ \beta_s \}_{s=1}^S \) in addition to \( \alpha \) as follows:

\[
\min_{\alpha, \beta} \left[ \sum_{i=1}^N \mathcal{L}(y_i, \sum_{j=1}^N \alpha_s \sum_{s=1}^S \beta_s [K_s]_{ij}) + \| \alpha \|_K^2 + \| \beta \|_\infty^2 \right].
\]

In one of the first studies on MKL, a framework was proposed to combine multiple kernel functions for SVMs with several types of loss functions[3]. In this method, the margin of the SVM classifier was simultaneously maximized with respect to \( \alpha \) and \( \beta \), by semidefinite programming (SDP). There are some computationally efficient MKL methods such as SimpleMKL[7], but their classification accuracy is comparable to SDP-MKL. Recently, a novel MKL
method (RMKL) was proposed[4]. This method takes into account the fact that the theoretical error bound of an SVM depends on both the margin and the radius of the smallest sphere that contains all the training samples. In the present study, we only compare the proposed MKL to RMKL, because RMKL is computationally efficient as SimpleMKL and more accurate than SDP-MKL. It should be noted that most existing MKL methods are based on maximum-margin-based schemes and do not explicitly provide information-theoretic interpretation.

3. MULTIPLE KERNEL LEARNING BASED ON CONDITIONAL ENTROPY MINIMIZATION

In this section, we introduce the MKL algorithm based on conditional entropy minimization (MCEM), which was proposed in [6]. We verify the adequacy of applying the conditional entropy minimization criterion to MKL and briefly review the MCEM algorithm.

3.1. Adequacy of Conditional Entropy Minimization

As stated in [6], we can interpret the classification as supervised dimensionality reduction, because the classification function \(f: \mathbb{R}^n \rightarrow \hat{z}\), \(x \in \mathbb{R}^n, \hat{z} \in \mathbb{R}\) is regarded as a projection function onto a one-dimensional classification axis (e.g., binary or one-versus-rest classifiers). In supervised dimensionality reduction, the projected data should be compactly aggregated in each class. From an information-theoretic viewpoint, the compactness in each class corresponds to small conditional entropy[8]. Thus, it is reasonable to adopt the conditional entropy minimization criterion for supervised MKL.

3.2. MCEM Algorithm

From the above discussion, we formulated MCEM as follows:

\[
\min_{\alpha, \beta} H(f(X; \alpha, \beta) \mid Y)
\]

s. t. \(H(f(X; \alpha, \beta)) = \text{const.}, \sum_{s=1}^{S} \beta_s = 1, \beta_s \geq 0,\)

where \(f(x; \alpha, \beta)\) denotes the classification function expressing the dependency on \(\alpha\) and \(\beta\) explicitly. We used \(H(f(X; \alpha, \beta))\) as a regularization function for minimizing \(H(f(X; \alpha, \beta) \mid Y)\) to avoid trivial solutions and overfitting. In MCEM, the kernel Fisher discriminant analysis (KFDA)[9] is used for the optimization of MKL.

3.2.1. KFDA and Its Application to MCEM

The KFDA finds a linear projection of the observed samples in a supervised manner. Here, we describe the adequacy of applying the KFDA to MCEM. Suppose an observed sample \(x \in \mathbb{R}^n\) is mapped onto \(n\)-dimensional feature space \(\mathbb{R}^n\) by the feature mapping \(\Phi: \mathbb{R}^n \rightarrow \mathbb{R}^n\). Consider a classification function in terms of \(f(x) = \alpha^T \Phi(x), \alpha' \in \mathbb{R}^n\). In kernel methods, with an appropriate regularity condition, we can apply the representer theorem to obtain the expression \(\alpha' = \sum_{i=1}^{N} \alpha_i \Phi(x_i)\) with real-valued weight parameters \(\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)\). Let \(K \in \mathbb{R}^{N \times N}\) be the Gram matrix of the given data set \(D_t\) and \(k_i\) be the \(i\)-th column vector of \(K\). The sample mean vector of each class and that of all the data are computed by \(\bar{k}_y = \frac{1}{N_y} \sum_{y \in D_y} k_i\) and \(\bar{k} = \frac{1}{N} \sum_{t \in D_t} k_i\), respectively, where \(D_y\) denotes a set of observed samples that belong to the class \(y; N_y = |D_y|\), the number of samples in the class \(y\); and \(N = \sum_y N_y\), the total number of samples. The within-class covariance matrix \(\Sigma_w\) and between-class covariance matrix \(\Sigma_b\) in the feature space are computed as follows:

\[
\Sigma_w = \frac{1}{N} \sum_{y \in \{\pm 1\}} \sum_{i \in D_y} (k_i - \bar{k}_y)(k_i - \bar{k}_y)^T, \quad (5)
\]

\[
\Sigma_b = \frac{1}{N} \sum_{y \in \{\pm 1\}} N_y (\bar{k}_y - \bar{k})(\bar{k}_y - \bar{k})^T. \quad (6)
\]

The objective of the KFDA is to minimize the within-class data scattering and maximize the between-class data scattering, which are measured by \(\alpha^T \Sigma_w \alpha\) and \(\alpha^T \Sigma_b \alpha\), respectively. Hence, the problem of the KFDA is formulated as a minimization problem, as follows:

\[
\min_{\alpha} \left[ \alpha^T (\Sigma_w + \zeta \Sigma_b) \alpha \right] \quad \text{s. t. } \alpha^T \Sigma_b \alpha = \text{const.} \quad (7)
\]

In this case, we regularize \(\Sigma_w \) as \(\Sigma_w + \zeta \Sigma_b\), \(\zeta \geq 0\) to avoid overfitting.

Since class-conditional entropy is upper bounded as follows[6]:

\[
H(f(X) \mid Y) \leq (2\pi\eta)^{1/2} + \frac{1}{2} \log(\alpha^T \Sigma_w \alpha), \quad (8)
\]

the KFDA is regarded as a minimization problem for an upper bound of the class-conditional entropy of a value on the projected axis, because the KFDA objective function is essentially the same as the term on the right-hand side of Eq. 8. Considering this fact, we have formulated MCEM as supervised dimensionality reduction, which aims at finding a projection \(f\) that minimizes the class-conditional entropy \(H(f(X) \mid Y)\).

3.2.2. Optimization of \(\alpha\) and \(\beta\)

The MCEM algorithm is summarized in Algorithm 1. The direct simultaneous optimization of Eq. 4 with respect to both \(\alpha\) and \(\beta\) is difficult. Therefore, we adopted an alternating optimization approach, i.e., we iterate the following two steps: optimizing \(\alpha\) for a fixed \(\beta\) and optimizing \(\beta\) using a fixed \(\alpha\) estimated in the previous step. The estimates of \(\alpha\) and \(\beta\) at the \(t\)-th iteration are denoted by \(\alpha(t)\) and \(\beta(t)\), respectively.

The KFDA is used for optimizing class-conditional entropy with respect to \(\alpha\). From Eq. 8, the optimal upper-bound solution \(\alpha\) is given by the KFDA for a fixed \(\beta\). Such an optimization of \(\alpha\) is represented by Eq. 9. For the optimization of \(\beta\), we define the minimization problem of the class-conditional entropy, as described by Eq. 10, where \(H(f(X; \alpha, \beta) \mid Y)\) denotes a regularization term with a tuning parameter \(\eta > 0\). In the present study, we had adopted the random search algorithm in [6]. We sampled \(P\) candidates of \(\beta\) from a Gaussian distribution with a mean vector \(\beta(\cdot - 1)\) and an identity covariance matrix. Then, we computed \(H(f(X; \alpha(t); \beta(t)) \mid Y)\) with \(\beta(\cdot = 1)\) and selected \(\beta\) that minimized the objective function described in Eq. 10. In the present study, we estimated entropy and class-conditional entropy using the efficient method proposed in [10].

After the \(\beta\) optimization step, we obtain a new kernel matrix with the updated coefficient \(\beta(\cdot)\). With this new kernel matrix, we can update \(\Sigma_w(\beta(\cdot))\) and \(\Sigma_b(\beta(\cdot))\), which are explicitly dependent on \(\beta(\cdot)\). Then, we again minimize the updated objective function of the KFDA with respect to \(\alpha\). We iterate these two steps until \(\alpha\) and \(\beta\) converge or until some predetermined stop criterion is satisfied.
Algorithm 1 MCEM: Multiple kernel learning algorithm based on conditional entropy minimization

Input: A set of kernel functions \( \{k_s(x, x')\}^S_{s=1} \), training data \( D = \{(x_1, y_1), \ldots, (x_N, y_N)\} \in \mathbb{R}^n \), class labels \( \{y_1, \ldots, y_N\} \in \{-1, 1\} \), and a regularization parameter \( \zeta \) for KDFA.

Initialization: Calculate kernel matrices \( \{K_s\}^S_{s=1} \) using \( D \). Initialize the combination coefficients \( \beta^{(0)} = \{\beta^{(0)}_1, \ldots, \beta^{(0)}_S\} \) of element kernels at random.

Repetition: Until convergence, from \( t = 1 \):

\( \alpha \) optimization step: Solve a KDFA minimization problem with fixed \( \beta^{(t-1)} \) to obtain \( \alpha^{(t)} \):

\[
\min_{\alpha} \left[ \alpha^T \left( V_w(\beta^{(t-1)}) + \zeta K \right) \alpha \right] \quad \text{s.t.} \quad \alpha^T \left( V_b(\beta^{(t-1)}) \right) \alpha = \text{const}. \tag{9}
\]

\( \beta \) optimization step: Minimize the conditional entropy of \( f(x; \alpha^{(t)}, \beta) \) with \( \alpha^{(t)} \) to obtain \( \beta^{(t)} \):

\[
\min_{\beta} \left[ H(f(X; \alpha^{(t)}, \beta)) - \eta H(f(X; \alpha^{(t)}, \beta)) \right] \quad \text{s.t.} \quad \sum_{s=1}^S \beta_s = 1, \quad \beta_s \geq 0. \tag{10}
\]

Output: Converged parameter \( \alpha \) and \( \beta \), used to construct the classification function as follows:

\[
f(x; \alpha, \beta) = \frac{1}{\sum_{i=1}^N \alpha_i} \sum_{s=1}^S \beta_s k_s(x, x). \tag{11}
\]

4. SPEAKER RECOGNITION EXPERIMENT

Experimental comparisons were made for text-independent speaker recognition. The first point to be investigated is the discrimination capability of MCEM. For this purpose, pairwise comparisons were performed for each pair of 20 speakers (Exp. 1), which involved a total of 190 comparisons. The second investigation involves using the comparison results to classify the 20 speakers (Exp. 2).

4.1. Experimental Condition

4.1.1. Evaluation Parameters

In the present study, we evaluated the performance of speaker recognition systems using the following classifiers: a Gaussian mixture model (GMM), two SVMs using multiple kernels that were combined by RMKL[4] and MCEM, and six SVMs each using one of the element kernels of the aforementioned multiple kernels. Table 2 lists the element kernels used. Table 3 lists the classifiers compared.

4.1.2. Speech Samples

The speech samples used consists of a 4320 Japanese phoneme-balanced word database, which consists of 216 types of isolated spoken words uttered by the 20 male speakers. These speech utterances were recorded using close-talking microphones, sampled at 16 kHz and quantized into 16-bit data. In the present study, 2-fold cross validation (CV) experiments were carried out. For each speaker, 108 words were used for training, and the remaining 108 words were used for evaluation.

4.1.3. Acoustic Feature Extraction

The acoustic feature parameters used in the experiment were 24-dimensional parameters consisting of 12-dimensional mel-frequency cepstral coefficients (MFCCs) and 12-dimensional ΔMFCCs. The experimental conditions for acoustic feature extraction are listed in Table 1.

4.1.4. GMM-based System

In the GMM-based system, we used a 128-mixture Gaussian distribution with diagonal covariances for each speaker, which was trained by the EM algorithm using framewise 24-dimensional vectors.

At the classification stage in Exp. 1, the likelihoods of each utterance spoken by two speakers were calculated using the corresponding two speaker models. In Exp. 2, the likelihoods of each utterance spoken by 20 speakers were calculated using the corresponding 20 speaker models. Then, the best matching speaker was determined for each utterance.

4.1.5. SVM-based System

We used the sequential kernel[11] to handle speech utterances with different frame length. For the two utterances, \( \mathbf{X} = (x_1, \ldots, x_T) \in \mathbb{R}^{n \times T} \) and \( \mathbf{X}' = (x'_1, \ldots, x'_{T'}) \in \mathbb{R}^{n \times T'} \), where \( n = 24 \) in this case, we defined the sequential kernel as

\[
K(\mathbf{X}, \mathbf{X}') = \frac{1}{T \cdot T'} \sum_{t=1}^T \sum_{t'=1}^{T'} k(x_t, x'_{t'}). \tag{12}
\]

For RMKL and MCEM, \( k(x_t, x'_{t'}) \) used in the present study were three RBF kernels, two polynomial kernels, and a linear kernel, as listed in Table 2. We set the regularization parameter \( \zeta \) in Eq. 9 at \( \zeta = 0.001 \) for all the experiments. In RMKL, which is based on soft-margin SVMs, we set the soft-margin parameter to one. In MCEM, we used \( \eta = 0.5 \) for the trade-off parameter and sampled 500 candidates of \( \beta \) (i.e., \( P = 500 \)) in the \( \beta \) optimization step. These parameters were determined such that the best performance could be achieved in the preliminary experiment. After the optimal map \( f \) (i.e., the optimal \( \alpha \) and \( \beta \)) was obtained, we computed projections onto a one-dimensional classification axis using Eq. 11. The iteration process for optimizing \( \alpha \) and \( \beta \) was stopped when the rate of decrease in conditional entropy reduced below 0.0001 or when conditional entropy increased.

Various classification methods using projected samples have been developed. In the present study, we used the one-versus-one
Table 3. Speaker error rate (%) for GMM, SVMs with single kernel, and SVMs with multiple kernels. In Exp. 1, each pair among the 20 speakers was classified. In Exp. 2, the 20 speakers were classified.

<table>
<thead>
<tr>
<th>System</th>
<th>Exp. 1</th>
<th>Exp. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM</td>
<td>3.0</td>
<td>4.2</td>
</tr>
<tr>
<td>SVM (single kernel)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBF ($\sigma = 0.01$)</td>
<td>1.0</td>
<td>0.046</td>
</tr>
<tr>
<td>RBF ($\sigma = 0.05$)</td>
<td>12.0</td>
<td>12.6</td>
</tr>
<tr>
<td>RBF ($\sigma = 0.1$)</td>
<td>27.1</td>
<td>4.4</td>
</tr>
<tr>
<td>Polynomial ($d = 2$)</td>
<td>2.5</td>
<td>0</td>
</tr>
<tr>
<td>Polynomial ($d = 3$)</td>
<td>1.9</td>
<td>2.5</td>
</tr>
<tr>
<td>Linear</td>
<td>5.7</td>
<td>0</td>
</tr>
<tr>
<td>SVM (multiple kernel)</td>
<td>1.6</td>
<td>0</td>
</tr>
<tr>
<td>MCEM (proposed)</td>
<td>1.4</td>
<td>0</td>
</tr>
</tbody>
</table>

SVM classifiers, which classify one speaker with respect to another speaker. In this case, we constructed a total of 190 classifiers for each CV fold. In Exp. 2, we used the winner-takes-all strategy for integrating the one-versus-one SVM classifiers.

4.2. Experimental Results

Table 3 lists the classifiers evaluated in this experiment and the corresponding speaker error rates. The results show that MCEM aids in realizing high-performance speaker recognition systems. In Exp. 1, the MCEM-based system gave an error rate of 1.4%; this system reduced the number of errors by 53.3% and 12.5% as compared to the GMM-based system (3.0%) and RMKL-based system (1.6%), respectively. In Exp. 2, the MCEM- and RMKL-based systems did not give any error, while an error rate of the GMM-based system was 4.2%.

In addition, the MKL-based systems showed equivalent or better performance than the SVM-based systems with a single kernel. In Exp. 1, the SVM-based system that used the RBF kernel with $\sigma = 0.01$ outperformed the MKL-based systems. However, in the case of SVMs with a single kernel, strict tuning of the kernel functions and parameters was required, because the performance of the speaker recognition systems varied significantly according to the selected kernel functions and parameters; for example, the best performance was 1.0%, while the worst was 27.1%. In contrast, the MKL methods achieved high-performance speaker recognition even though these methods did not involve strict determination of the kernel functions and parameters. The reduction in the errors of single-kernel SVM-based systems by the MKL-based systems implied that SVMs with different kernel functions and parameters resulted in different error trends, and thus, their individual weaknesses were reduced by integrating such complementary classifiers. In Exp. 2, the MKL-based systems and some single-kernel SVM-based systems did not give any error. In this case, the classifier integration based on the winner-takes-all strategy may have contributed to the error correction effect.

Figure 1 shows the mean values of the learned combination coefficients $\hat{\beta}$ for element kernels listed in Table 2. The MCEM- and RMKL-based systems showed equivalent performance but these systems induced different trends in terms of $\hat{\beta}$. The MCEM-based system showed a clear trend, i.e., $\hat{\beta}$ for the polynomial kernels and linear kernel was approximately zero, while no such clear trend was observed in the RMKL-based system. It should be noted that the values of $\hat{\beta}$ need not correlate with the performance of the corresponding systems, and $\hat{\beta}$ is determined on the basis of the complementarity of the element classifiers. In the MKL-based systems, in terms of model selection and feature selection, it is important whether or not several estimates of $\beta$ become zero (or approximately zero). The results shown in Fig. 1 imply that MCEM may be more suitable for model selection and feature selection than RMKL.

5. CONCLUSION

We described the framework and algorithm of an MKL method based on conditional entropy minimization and applied this method to speaker recognition. This method eliminates the need for strict determination of a suitable kernel function and its parameters. In addition, this method classifies the data in the discriminative subspace obtained by information-theoretic dimensionality reduction. The proposed method reduced errors by approximately 53% and 13% as compared to the GMM- and RMKL-based methods, respectively. In future, we will evaluate the MCEM-based system using a larger speech database (e.g., NIST-SRE corpus) and assess its robustness to noise.

6. REFERENCES