Research Article

Cross-Corpus Speech Emotion Recognition Based on Multiple Kernel Learning of Joint Sample and Feature Matching

Ping Yang

College of Big Data and Information Engineering, Guizhou University, Guiyang 550002, China

Correspondence should be addressed to Ping Yang; pyang3@gzu.edu.cn

Cross-corpus speech emotion recognition, which learns an accurate classifier for new test data using old and labeled training data, has shown promising value in speech emotion recognition research. Most previous works have explored two learning strategies independently for cross-corpus speech emotion recognition: feature matching and sample reweighting. In this paper, we show that both strategies are important and inevitable when the distribution difference is substantially large for training and test data. We therefore put forward a novel multiple kernel learning of joint sample and feature matching (JSFM-MKL) to model them in a unified optimization problem. Experimental results demonstrate that the proposed JSFM-MKL outperforms the competitive algorithms for cross-corpus speech emotion recognition.

1. Introduction

In cross-corpus speech emotion recognition, there is a descent in the recognition performance of many algorithms [1–3]. This is because the lacking of robust features representation and important properties training samples. To address the above issue, researchers use the matched feature selection and sample reweighting [4, 5]. Feature selection or extraction algorithm discovers the shared feature representation for reducing the distribution mismatch between the training and test data. Sample reweighting also aims at reducing this distribution mismatch by reweighting the training samples and then training a robust recognizer on the reweighted training samples. In cross-corpus speech emotion recognition, as well known, there will always exist some training samples that are not relevant to the test samples even in the feature matching subspace [6]. Recent works have also exploited matching feature leaning and sample reweighting individually for improving the performance of cross-corpus speech emotion recognition [4, 5]. However, it is natural to combine the benefits of the two categorical learning strategies in cross-corpus speech emotion recognition. In this work, we extend the idea of feature extraction and sample reweighting to multiple kernel learning (MKL) and

propose a novel multiple kernel learning of joint sample and feature matching (JSFM-MKL) to model them in a unified optimization problem. We test the proposed JSFM-MKL on FAU Aibo speech emotion corpus, which was used in the Interspeech 2009 Emotion Challenge. Experimental results show that the proposed JSFM-MKL outperforms MKL [7] and adaptive multiple kernel learning (A-MKL) [8] and significantly improves the baseline performance of the Emotion Challenge.

2. MKL of Joint Sample and Feature Matching

2.1. Problem Definition. We are given the training D^{tr} and test data D^{te}, respectively. The training data D^{tr} is fully labeled and represented as $D^{\text{tr}} = \{(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_{n_{\text{tr}}}, y_{n_{\text{tr}}})\} \subset R^d \times Y$, where y_i is the label of x_i. The test data is divided into labeled $D_l^{\text{te}} = \{(x_1, y_1), \ldots, (x_{n_{\text{te}}}, y_{n_{\text{te}}})\} \subset R^d \times Y$ and unlabelled $D_u^{\text{te}} = \{(x_1, y_1), \ldots, (x_{n_{\text{te}}}, y_{n_{\text{te}}})\} \subset R^d \times Y$ parts. The training and test data have the equal dimensionality of feature representation d. Our goal is design a robust recognizer to predict label on the unlabelled test data. The proposed recognizer is based on MKL framework [8], in which the sample reweighting and feature matching schemes

are modeled in a unified optimization problem of MKL. Specifically, the learning framework of joint sample and feature matching MKL (JSFM-MKL) can be formulated as

$$
[k, f] = \underset{k,f}{\arg\min} \quad \Omega\left(\text{DIST}_k^2\left(D^{\text{tr}}, D^{\text{te}}\right)\right)
$$
$$
+ \theta R\left(f, k, D\right), \tag{1}
$$

where $\Omega(\bullet)$ is any increasing monotonic function and $\theta > 0$ is the trade-off between the distribution mismatch and the structural risk function $R(f, k, D)$ on the labeled data.

Our work JSFM-MKL is motivated by the following two aspects: matching feature selection and sample reweighting. The training data may be less representative with the testing data for cross-corpus speech emotion recognition. More specifically, $p_{\text{tr}}(Y \mid X = x)$ is different from $p_{\text{te}}(Y \mid X = x)$. This indicates that some features may behave differently between the training and test data. A recognizer that heavily relies on these features in training data may be not perform well in the recognition tasks of the unlabelled test data. Thus, one key computational problem is to reduce the distribution mismatch between $p_{\text{tr}}(Y|X = x)$ and $p_{\text{te}}(Y \mid X = x)$ [9]. However, it is not a nontrivial problem to intermediately estimate the probability density. To avoid this problem, we resort to the empirical Maximum Mean Discrepancy (MMD) [10], which is an effective nonparametric distance measure to compare data distribution in RKHS. Using the training and test data, the MMD can be formulated as follows:

$$
\text{DIST}_k^2\left(D^{\text{tr}}, D^{\text{te}}\right) = \left\| \frac{1}{n_{\text{tr}}} \sum_{i=1}^{n_{\text{tr}}} \phi\left(x_i\right) - \frac{1}{N_{\text{te}}} \sum_{i=n_{\text{tr}}+1}^{n_{\text{tr}}+n_{\text{te}}} \phi\left(x_i\right) \right\|^2. \tag{2}
$$

Let $\Phi^{\text{tr}} = [\phi(x_1^{\text{tr}}), \ldots, \phi(x_{n_{\text{tr}}}^{\text{tr}})]$ be the feature mapping matrix of training data and $\Phi^{\text{te}} = [\phi(x_1^{\text{te}}), \ldots, \phi(x_{n_{\text{te}}}^{\text{te}})]$ be the feature mapping matrix of test data. In addition, we define two column vectors s^{tr} and s^{te}, respectively. s^{tr} has n_{tr} entries by setting each entry as $1/n_{\text{tr}}$, and n^{te} has n_{tr} entries by setting each entry as $1/n_{\text{te}}$. Then (2) can be rewritten as

$$
\text{DIST}_k^2\left(D^{\text{tr}}, D^{\text{te}}\right) = \left\| [\Phi^{\text{tr}} \quad \Phi^{\text{te}}] \begin{bmatrix} s^{\text{tr}} \\ s^{\text{te}} \end{bmatrix} \right\|^2
$$
$$
= [s_{\text{tr}}' \quad s_{\text{te}}'] \begin{bmatrix} K^{\text{tr,tr}} & K^{\text{tr,te}} \\ K^{\text{te,tr}} & K^{\text{te,te}} \end{bmatrix} \begin{bmatrix} s_{\text{tr}} \\ s_{\text{te}} \end{bmatrix}. \tag{3}
$$

Instead of learning a kernel matrix, following [8], we assume a kernel is a linear combination of base kernels, namely,

$$
k = \sum_{m=1}^{M} d_m k_m, \tag{4}
$$

where $d_m \geq 0$, $\sum_{m=1}^{M} d_m = 1$. We furthermore assume that the first objective $\Omega(\text{DIST}_k^2(D^{\text{tr}}, D^{\text{te}}))$ in (1) is

$$
\Omega\left(\text{DIST}_k^2\left(D^{\text{tr}}, D^{\text{te}}\right)\right)
$$
$$
= \frac{1}{2} \left(\sum_{p=1}^{P} d_p [s_{\text{tr}}' \quad s_{\text{te}}'] \begin{bmatrix} K_p^{\text{tr,tr}} & K_p^{\text{tr,te}} \\ K_p^{\text{te,tr}} & K_p^{\text{te,te}} \end{bmatrix} \begin{bmatrix} s_{\text{tr}} \\ s_{\text{te}} \end{bmatrix} \right)^2. \tag{5}
$$

However, (5) does not consider the role of each feature on reducing the mismatch of conditional distribution. Therefore, it is natural to select the features that can reduce the mismatch of conditional distribution. Although the previous MKL can perform feature selection by the corresponding kernel weights, it generally regards the all features from the same distribution. In other words, it did not address this problem of cross-corpus feature selection as we do [7]. To address this problem, we construct each type of feature with k different kernel choices and formulate the weight of kernel as the matrix D. The entry d_{mp} is the weight of the mth type feature corresponding to the pth kernel. As to feature selection, we impose $l_{2,1}$ norm constraint on D, which shrinks the entries of some rows to zero. This $l_{2,1}$ norm constraint is defined as the summation of the l_2 norm of row of D. Then, (4) can be reformulated as follows:

$$
\frac{1}{2} \left(\sum_{mp} d_{mp} [s_{\text{tr}}' \quad s_{\text{te}}'] \begin{bmatrix} K_{\text{tr,tr}} & K_{\text{tr,te}} \\ K_{\text{te,tr}} & K_{\text{te,te}} \end{bmatrix} \begin{bmatrix} s_{\text{tr}} \\ s_{\text{te}} \end{bmatrix} \right)^2 + \|D\|_{2,1}
$$
$$
\sum_{mp} d_{mp} = 1, \quad d_{mp} \geq 0, \tag{6}
$$

where $D = [d_{mp}] \in R^{M \times P}$ is the weight matrix of base kernels. The mixed $l_{2,1}$ norm constraint creates the sparsity between different features, while the values of d_{mp} for the same feature need not sparsity. This will make that a different property of selected features able to be represented by more than one kernel.

However, matching feature selection based on the MMD minimization is not good enough for cross-corpus speech emotion recognition, since it only reduces the mismatch of conditional distribution by high order moments of probability distribution. Then the distribution mismatch is far away perfect. In fact, there are some training samples that are irrelevant to the test samples. Therefore, a sample reweighting procedure should be combined with the matching feature selection to deal with this difficult setting. Following the previous works, Kernel mean matching (KMM) [5] is introduced to weight the training data by minimizing the difference between the means of weighted-training and test data distribution in RKHS. Different from the previous works, the sample reweighting procedure and matching feature selection are modeled in a unified optimization problem. Thus the optimization problem can be rewritten as

$$
\sum_{mp} d_{mp} \left[\frac{1}{n_{\text{tr}}} \beta' \quad s_{\text{te}}' \right] \begin{bmatrix} K_{mp}^{\text{tr,tr}} & K_{mp}^{\text{tr,te}} \\ K_{mp}^{\text{te,tr}} & K_{mp}^{\text{te,te}} \end{bmatrix} \begin{bmatrix} \frac{1}{n_{\text{tr}}} \beta \\ s_{\text{te}} \end{bmatrix}
$$
$$
+ \|D\|_{2,1},
$$
$$
\text{s.t.} \quad \sum_{mp} d_{mp} = 1, \quad d_{mp} \geq 0 \tag{7}
$$
$$
0 \leq \beta_i \leq B,
$$
$$
n_{\text{tr}} (1 - \varepsilon) \leq \sum_{i=1}^{n_{\text{tr}}} \beta_i \leq n_{\text{tr}} (1 + \varepsilon).
$$

Letting $B(\beta_{mp}) = [(1/n_{tr}^2)\beta' K_{mp}^{tr,tr}\beta + (1/n_{tr})(s_{te}' K^{te,tr}\beta + \beta' K^{tr,te}s_{te}) + s_{te}' K^{te,te}s_{te}] \in R^{M\times P}$,
(6) can be rewritten as follows:

$$\Omega\left(\text{DIST}_k^2\left(D^{tr}, D^{te}\right)\right)$$

$$= \frac{1}{2}\left(\text{tr}\left(\text{DB}\left(\beta_{mp}\right)\right)\right)^2 + \|D\|_{2,1}$$

$$\sum_{mp} d_{mp} = 1, \quad d_{mp} > 0 \tag{8}$$

$$0 \le \beta_i \le B,$$

$$n_{tr}(1-\varepsilon) \le \sum_{i=1}^{n_{tr}}\beta_i \le n_{tr}(1+\varepsilon).$$

After obtaining $\Omega(\text{DIST}_k^2(D^{tr}, D^{te}))$, we use the objective function of MKL to model the second objective function $R(f, k, D)$. Thus, the optimization problem JSFM- MKL can be written as

$$\min_{D,\beta} \quad \frac{1}{2}\left(\text{tr}\left(\text{DB}\left(\beta\right)\right)\right)^2 + \|D\|_{2,1} + J(D)$$

$$0 \le \beta_i \le B, \tag{9}$$

$$n_{tr}(1-\varepsilon) \le \sum_{i=1}^{n_{tr}}\beta_i \le n_{tr}(1+\varepsilon),$$

where

$$J(D) = \frac{1}{2}\sum_{m,p=1}^{M,k}\frac{1}{d_{m,p}}\left\|\phi_{m,p}\left(\omega_{m,j}\right)\right\|^2 + C\sum_{i=1}^{n_{tr}}\xi_i$$

$$\text{s.t.} \quad y_i\left(\sum_{m,j}^{M,P}\left\langle\phi_{mj}\left(\omega_{mj}\right), \phi_{mj}\left(x_i\right)\right\rangle + b\right) + \xi_i \ge 1 \tag{10}$$

$$\xi_i \ge 0, \quad \text{for } i = 1, \ldots, n_{tr}.$$

By introducing the Lagrange multiplier α, the dual form of the optimization of JSFM- MKL can be formulated as

$$\min_{D,\beta} \quad \frac{1}{2}\left(\text{tr}\left(\text{DB}\left(\beta\right)\right)\right)^2 + \|D\|_{2,1} + J(D)$$

$$\sum_{mp} d_{mp} = 1, \quad d_{mp} > 0$$

$$0 \le \beta_i \le B, \tag{11}$$

$$n_{tr}(1-\varepsilon) \le \sum_{i=1}^{n_{tr}}\beta_i \le n_{tr}(1+\varepsilon),$$

where

$$J(D) = \max \quad 1_n'\alpha - \frac{1}{2}(\alpha \circ y)'\left(\sum d_{mp}K_{mp}\right)(\alpha \circ y),$$

$$\sum_{mp} d_{mp} = 1, \quad d_{mp} > 0$$

TABLE 1: Acoustic features and statistical functionals.

Raw acoustic features + deltas	Statistical functional
Pitch	Mean, stand deviation, kurtosis
Root mean square energy	Skewness, minimum, maximum
Zero cross rate	Relative position .range
Harmonic to noise ratio Mel-frequency Cepstral coefficients mean square error of (1–12 MFCC)	Two liner regression coefficients Mean square error of liner regression

$$0 \le \beta_i \le B,$$

$$n_{tr}(1-\varepsilon) \le \sum_{i=1}^{n_{tr}}\beta_i \le n_{tr}(1+\varepsilon). \tag{12}$$

In this work, we employ alternate optimization algorithm [8] to iteratively update the dual variable α, the weighting matrix D, and the weighting vector β. Specifically, we update the dual variable α with the fixed weighting matrix D and the weighting vector β; then we update the weighting matrix D and the weighting vector β with fixed variable α.

3. Experiments

In this work, we evaluate the proposed JSFM-MKL using the spontaneous FAU Aibo Emotion Corpus [11]. This corpus was an integral part of Interspeech 2009 Emotion Challenge [12]. It contains recordings of 51 children at the age of 10–13 years interacting with Sony's dog-like Aibo robot. The children were asked to treat the robot as a real dog and were led to believe that the robot was responding to their spoken commands. In this recognition task, we use these utterances including 5-class emotion: angry, emphatic, positive, neutral, and rest. The evaluation measure of all experimental results is the average unweighted recall, which is defined as the accuracy per class averaged by total number of classes and is more suitable for imbalanced data [12]. To achieve good average unweighted recall, we arrange multiple recognizers into the binary decision tree structure proposed by Lee et al. [13]. In addition, we use synthetic minority oversampling [14] to reduce the imbalance of classes during each recognizer training phrase. For acoustic feature extraction, we use a "brute force" approach based on a baseline feature set without any attempt to select a smaller subset of well-performing features. Specifically, we use the OpenEar toolkit [15] to extract acoustic features from each utterance.

The feature set includes 16 low level descriptors consisting of prosodic, spectral envelope, and voice quality features listed in Table 1. These low level descriptors are zero crossing rate, root mean square energy, pitch, harmonics-to-noise ratio, and 12 mel-frequency cepstral coefficients and their deltas. Then 12 statistical functionals were computed for every low level descriptor per utterance in the Aibo database:

TABLE 2: Experimental results under 0 labeled test samples.

Methods	Testing on Aibo-Mont		Testing on Aibo-Ohm	
	2-class	5-class	2-class	5-class
uLSIF	67.07	39.54	66.82	38.25
KMM	68.54	38.31	66.15	37.69
KLIEP	67.07	39.75	66.07	39.14
MKL	68.98	40.06	67.89	39.40
A-MKL	70.12	40.23	68.58	40.51
JSFS-MKL	**71.45**	**41.45**	**69.54**	**41.58**

TABLE 3: Experimental results under 50 labeled test samples.

Methods	Testing on Aibo-Mont		Testing on Aibo-Ohm	
	2-class	5-class	2-class	5-class
uLSIF	71.34	42.65	70.65	42.32
KMM	72.13	43.50	71.57	42.25
KLIEP	71.97	46.06	71.87	43.78
MKL	73.64	45.96	74.16	73.00
A-MKL	74.56	46.82	75.13	47.64
JSFS-MKL	**76.48**	**47.13**	**77.13**	**46.80**

TABLE 4: Experimental results under 100 labeled test samples.

Methods	Testing on Aibo-Mont		Testing on Aibo-Ohm	
	2-class	5-class	2-class	5-class
uLSIF	75.46	46.63	76.41	47.35
KMM	76.31	47.31	76.82	46.03
KLIEP	75.60	46.32	77.24	47.00
MKL	77.21	47.80	78.52	49.87
A-MKL	79.96	49.62	79.00	50.34
JSFS-MKL	**81.44**	**50.47**	**82.74**	**51.73**

kurtosis, skewness, minimum, maximum, relative position, range, two linear regression coefficients, mean, standard deviation, and their respective mean square error. This results in a collection of 384 acoustic features for per utterance. Then they were normalized between 0 and 1.

We systematically compare the proposed algorithm JSFM-MKL with the baseline MKL and other cross-corpus speech emotion algorithms including unconstrained least-squares importance fitting (uLSIF), kernel mean matching (KMM), and Kullback-Leibler importance estimation procedure (KLIEP). The kernel bandwidth in SVM and the penalty factor C are determined by cross-validation (5-fold) method over labeled training set. For the MKL, we construct 10 base kernels with different bandwidths, whose values are $\{2^{-3} \cdot \mu, 2^{-2} \cdot \mu, \ldots, 2^6 \cdot \mu\}$; the μ value is determined by the mean of the Euclidean distance between each pair of training samples. This work lets the amount of labeled test samples vary from 0 to 200. For each setting with labeled test samples, we ran 10 experiments with different, randomly chosen, labeled test samples. Specifically, the number of labeled test samples is 0, 50, 100, 150, and 200, respectively. Correspondingly, the average results

of all algorithms are presented in Tables 2, 3, 4, 5, and 6.

From Tables 2–6, we can see that the MKL-based recognition algorithms outperform the SVM-based ones, which indicate the data has a better representative ability in this space spanned by multiple kernel functions. The number of labeled test samples especially is zero; the best JSFS-MKL best UA of 71.45% is achieved by the JSFS-MKL algorithm, compared to 70.2% for the best contributor Interspeech 2009 Emotion Challenge [12]. JSFS-MKL significantly outperforms uLSIF, KMM, and KLIEP, which are cross-corpus speech emotion recognition algorithms based sample reweighting or matching. However, as we have justified in this paper, only sample reweighting or matching is not good enough for cross-corpus adaptation when the corpus difference is substantially large, since there will always be some samples which are not similar to the target samples. Existing feature selection methods, for example, MKL, can perform better than uLSIF, KMM, and KLIEP. However, MKL as a feature selection strategy is not effective as JSFS-MKL, which is a joint sample reweighting and feature selection algorithm for cross-corpus speech emotion recognition.

TABLE 5: Experimental results under 150 labeled test samples.

Methods	Testing on Aibo-Mont		Testing on Aibo-Ohm	
	2-class	5-class	2-class	5-class
uLSIF	78.64	50.71	78.79	49.60
KMM	78.54	50.13	79.45	50.34
KLIEP	79.64	49.30	78.69	49.21
MKL	80.45	51.45	79.35	51.27
A-MKL	81.67	52.64	81.04	52.28
JSFS-MKL	**82.40**	**53.09**	**82.33**	**53.47**

TABLE 6: Experimental results under 200 labeled test samples.

Methods	Testing on Aibo-Mont		Testing on Aibo-Ohm	
	2-class	5-class	2-class	5-class
uLSIF	81.23	52.44	82.73	51.39
KMM	80.21	52.06	82.69	51.87
KLIEP	81.03	51.47	81.33	50.80
MKL	83.45	52.37	83.98	52.74
A-MKL	84.90	53.06	84.78	53.00
JSFS-MKL	**85.69**	**54.02**	**86.72**	**54.04**

4. Conclusion

In this paper, we have proposed a novel multiple kernel learning of joint sample and feature matching (JSFM-MKL) for cross-corpus speech emotion. The proposed JSFS-MKL aims to jointly match features and reweight instances across domains in a multiple kernel learning procedure. An important advantage of JSFM-MKL is that it is robust to both the distribution difference and the irrelevant instances. Comprehensive experimental results show that JSFS-MKL is effective for a variety of cross-corpus speech emotion and can significantly outperform state-of-the-art adaptation method.

Conflicts of Interest

The author declares that they have no conflicts of interest.

Acknowledgments

References

[1] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[2] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Proceedings of the Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, IEEE, 2011.

[3] P. Song, Y. Jin, L. Zhao, and M. Xin, "Speech emotion recognition using transfer learning," *IEICE Transaction on Information and Systems*, vol. 97, no. 9, pp. 2530–2532, 2014.

[4] Schuller, Björn et al., "Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization," in *Proceedings of the Afeka-AVIOS Speech Processing Conference*, Tel Aviv, Israel, 2011.

[5] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 585–589, 2016.

[6] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: compensating for covariate shift," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 7, pp. 1458–1468, 2013.

[7] F. R. Bach, G. RG. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proceedings of the twenty-first international conference on Machine learning. ACM'04*, 2004.

[8] L. Duan, D. Xu, I. W. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1667–1680, 2012.

[9] M. Chen, K. Q. Weinberger, and B. John, *Co-Training for Domain Adaptation*, Advances in Neural Information Processing Systems, 2011.

[10] K. M. Borgwardt, M. Karsten et al., "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[11] S. Stefan, *Automatic cLassification of Emotion Related User States in Spontaneous Children's Speech*, University of Erlangen-Nuremberg, Erlangen, Germany, 2009.

[12] Schuller, Björn, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proceedings of the Tenth Annual Conference of the International Speech Communication Association*, 2009.

[13] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[14] V. Chawla Nitesh et al., "SMOTE: synthetic minority oversampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[15] F. Eyben, Florian, M. Wollmer, and B. Schuller, "OpenEAR-Introducing the munich open-source emotion and affect recognition toolkit," in *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII '09)*, pp. 1–6, IEEE, Amsterdam, The Netherlands, September 2009.