COVARIATE SHIFT ADAPTATION FOR SEMI-SUPERVISED SPEAKER IDENTIFICATION

[1]*Makoto Yamada,* [1]*Masashi Sugiyama, and* [2]*Tomoko Matsui*

[1]Department of Computer Science, Tokyo Institute of Technology
[2]Department of Statistical Modeling, The Institute of Statistical Mathematics

e-mail: yamada@sg.cs.titech.ac.jp sugi@cs.titech.ac.jp tmatsui@ism.ac.jp

ABSTRACT

In this paper, we propose a novel semi-supervised speaker identification method that can alleviate the influence of non-stationarity such as session dependent variation, the recording environment change, and physical condition/emotion. We assume that the utterance variation follows the *covariate shift* model, where only the utterance sample distribution changes in the training and test phases. Our method consists of weighted versions of kernel logistic regression and cross-validation and is theoretically shown to have the capability of alleviating the influence of covariate shift. We experimentally show through text-independent speaker identification simulations that the proposed method is promising in dealing with variations in session dependent utterance variation.

Index Terms— Speaker identification, covariate shift, semi-supervised learning, kernel logistic regression, importance estimation.

1. INTRODUCTION

Speaker identification methods are widely used in various real-world situations such as access control of information service systems and speaker detection in speech dialog and speaker indexing problems with large audio archives. Recently, the speaker identification and indexing problems in meetings attracted a great deal of attention.

Standard methods of text-independent speaker identification includes the *Gaussian mixture model (GMM)* [1] or kernel methods such as the *support vector machine (SVM)* [2]. In these supervised learning methods, it is implicitly assumed that training and test data follow the same distribution. However, the training and test distributions are not necessarily the same in practice since the utterance features vary over time due to session dependent variation, the recording environment change, and physical condition/emotion.

To alleviate the influence of session dependent variation, it is common to use speech samples recorded in several different sessions [3]. However, gathering many speech samples and labeling the speaker ID to the collected data are expensive both in time and cost and therefore not realistic in practice.

A more practical setup would be *semi-supervised learning*, where unlabeled samples are additionally given from the test environment. In semi-supervised learning, it is required that the training and test distributions are related to each other in some sense; otherwise we may not be able to learn anything about the test distribution from the training samples. A popular modeling is called *covariate shift* [4], where the input distributions are different in the training and test phases but the conditional distribution of labels remains unchanged.

In this paper, we formulate the semi-supervised speaker identification problem in the covariate shift framework and propose a method that can cope with utterance variation. Under covariate shift, standard maximum likelihood estimation is no longer consistent—the influence of covariate shift can be asymptotically canceled by weighting the log-likelihood terms according to the *importance* [4]: $w(\mathrm{X}) = p_{te}(\mathrm{X})/p_{tr}(\mathrm{X})$, where $p_{te}(\mathrm{X})$ and $p_{tr}(\mathrm{X})$ are test and training input densities. The importance weight $w(\mathrm{X})$ is unknown in practice and needs to be estimated from data. For weight estimation, we utilize the *Kullback-Leibler importance estimation procedure (KLIEP)* [5] due to its superior performance. The regularized kernel logistic regression model contains two tuning parameters: the kernel width and the regularization parameter [3]. Usually these tuning parameters are optimized based on *cross validation (CV)*. However, CV is no longer unbiased due to covariate shift and therefore is not reliable as a model selection method. To cope with this problem, we use an importance-weighted version of CV (IWCV) [6] for unbiased model selection. The validity of our approach is experimentally shown through text-independent speaker identification simulations.

2. PROBLEM FORMULATION

In this section, we formulate the speaker identification problem based on the kernel logistic regression (KLR) model.

Text-independent Speaker Identification: An utterance sample X pronounced by a speaker is expressed as a set of N *mel-frequency cepstrum coefficient (MFCC)* [7] vectors of

d dimensions:

$$\mathrm{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{d \times N}.$$

For training, we are given n labeled utterance samples

$$\mathcal{Z} = \{(\mathrm{X}_i, y_i)\}_{i=1}^n,$$

where $y_i \in \{1, \ldots, K\}$ denotes the index of the speaker who pronounced X_i. The goal of speaker identification is to predict the speaker index of a test utterance sample X based on the training samples. We predict the speaker index c of the test sample X following the *Bayes decision rule*:

$$\max_c p(y = c \,|\, \mathrm{X}).$$

For approximating the class-posterior probability, we use

$$p(y = c \,|\, \mathrm{X}; \mathrm{V}) = \frac{\exp f_{\boldsymbol{v}_c}(\mathrm{X})}{\sum_{l=1}^K \exp f_{\boldsymbol{v}_l}(\mathrm{X})},$$

where $\mathrm{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K]^\top \in \mathbb{R}^{K \times n}$ is the parameter, $^\top$ denotes the transpose, and $f_{\boldsymbol{v}_l}$ is a discriminant function corresponding to speaker l. This form is known as the *softmax* function and widely used in multiclass logistic regression. We use the following kernel regression model as the discriminant function $f_{\boldsymbol{v}_l}$:

$$f_{\boldsymbol{v}_l}(\mathrm{X}) = \sum_{i=1}^n v_{l,i} \mathcal{K}(\mathrm{X}, \mathrm{X}_i) \quad l = 1, \ldots, K,$$

where $\boldsymbol{v}_l = (v_{l,1}, \ldots, v_{l,n})^\top \in \mathbb{R}^n$ are parameters corresponding to speaker l and $\mathcal{K}(\mathrm{X}, \mathrm{X}')$ is a kernel function. In this paper, we use the *sequence kernel* [2] as the kernel function since it allows us to handle features with different size; for two utterance samples $\mathrm{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{d \times N}$ and $\mathrm{X}' = [\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{N'}] \in \mathbb{R}^{d \times N'}$ (generally $N \neq N'$), the sequence kernel is defined as

$$\mathcal{K}(\mathrm{X}, \mathrm{X}') = \frac{1}{NN'} \sum_{i=1}^N \sum_{i'=1}^{N'} \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}'_{i'}\|^2}{2\sigma^2}\right).$$

Note that kernel logistic regression is a modeling assumption; thus the true class-conditional probability may not be exactly realized by the kernel logistic regression model. This implies that there exists some *model error*, i.e., even when the parameter is chosen optimally, there remains an approximation error. This setup is not of course preferable, but more or less there exists a model error in practice since it is not generally possible to prepare an exactly correct model. Traditional machine learning theories often assume that the model at hand is correct (i.e., no model error exists). However, this is not realistic and not useful in practice, so in this paper we explicitly take into account *model misspecification* within the covariate shift framework.

Kernel Logistic Regression [3, 8]: We employ maximum likelihood estimation for learning the parameter V. The negative regularized log-likelihood function $\mathcal{P}_\delta^{\log}(\mathrm{V}; \mathcal{Z})$ for the kernel logistic regression model is given by

$$\mathcal{P}_\delta^{\log}(\mathrm{V}; \mathcal{Z}) = -\sum_{i=1}^n \log P(y_i \,|\, \mathrm{X}_i; \mathrm{V}) + \frac{\delta}{2}\mathrm{tr}(\mathrm{VKV}^\top),$$

where $\frac{\delta}{2}\mathrm{tr}(\mathrm{VKV}^\top)$ is a regularizer introduced for avoiding overfitting and $\mathrm{K} = [\mathcal{K}(\mathrm{X}_i, \mathrm{X}_j)]_{i,j=1}^n$ is the kernel Gram matrix. $\mathcal{P}_\delta^{\log}(\mathrm{V}; \mathcal{Z})$ is a convex function with respect to V and therefore its unique minimizer can be obtained by, e.g., the Newton method.

Model Selection in KLR: KLR includes two tuning parameters—the Gaussian width σ and the regularization parameter δ. One of the popular approaches to model selection is *cross validation (CV)*.

Let us divide the training set $\mathcal{Z} = \{(\mathrm{X}_i, y_i)\}_{i=1}^n$ into k disjoint non-empty subsets $\{\mathcal{Z}_i\}_{i=1}^k$ of (approximately) the same size. Let $\widehat{y}(\mathrm{X}; \mathcal{Z}_j)$ be an estimate of a speaker of a test utterance sample X obtained from $\{\mathcal{Z}_i\}_{i \neq j}$ (i.e., without $\mathcal{Z}_j$). Then the k-fold CV (kCV) score is given by

$$\widehat{R}_{kCV}^{\mathcal{Z}} = \frac{1}{k} \sum_{j=1}^k \frac{1}{|\mathcal{Z}_j|} \sum_{(\boldsymbol{x}, y) \in \mathcal{Z}_j} I(y = \widehat{y}(\mathrm{X}; \mathcal{Z}_j)),$$

where $|\mathcal{Z}_j|$ denotes the number of samples in the subset $\mathcal{Z}_j$ and $I(\cdot)$ denotes the indicator function.

KLR, CV, and Covariate Shift: The use of KLR and CV could be theoretically justified when the training utterance features and the test utterance features independently follow the *same* probability distribution with density $p(\mathrm{X})$ and the class label y follows the *common* conditional probability distribution $p(y \,|\, \mathrm{X})$ in the training and test phases. Indeed, if these conditions are fulfilled, KLR is shown to be *consistent*, i.e., the learned parameter converges to the optimal value:

$$\lim_{n \to \infty} \widehat{\mathrm{V}} = \mathrm{V}^*,$$

where $\widehat{\mathrm{V}}$ is the parameter learned by KLR and V^* is the optimal parameter that minimizes the expected prediction error for test samples:

$$\mathrm{V}^* = \operatorname*{argmin}_{\mathrm{V}} \iint I(y = \widehat{y}(\mathrm{X}; \mathrm{V})) p(y \,|\, \mathrm{X}) p(\mathrm{X}) \mathrm{d}y \mathrm{d}\mathrm{X}.$$

$\widehat{y}(\mathrm{X}; \mathrm{V})$ is an estimate of speaker of an utterance feature X for parameter V. Also, when $p(\mathrm{X})$ and $p(y \,|\, \mathrm{X})$ are common in the training and test phases, kCV is (almost) *unbiased*:

$$\mathrm{E}_{\mathcal{Z}}\left[\widehat{R}_{kCV}^{\mathcal{Z}} - R^{\mathcal{Z}}\right] \approx 0,$$

where $\mathrm{E}_{\mathcal{Z}}$ is the expectation over the training set $\mathcal{Z}$ and $R^{\mathcal{Z}}$ is the expected prediction error defined by

$$R^{\mathcal{Z}} = \iint I(y = \widehat{y}(\mathrm{X}; \mathcal{Z}))p(y \,|\, \mathrm{X})p(\mathrm{X})\mathrm{d}y\mathrm{d}\mathrm{X}.$$

However, in practical speaker identification, speech features are not stationary due to utterance variation, the recording environment change, and speaker feeling. Thus, the training and test feature distributions are not necessarily the same. Then the above good theoretical properties are no longer true.

If the training and test feature distributions share nothing in common, we may not be able to learn anything about the test distribution from the training samples. In this paper, we explicitly deal with such changing environment via the *covariate shift* model [4]—the input distributions change between the training and test phases, $p_{tr}(\mathrm{X}) \neq p_{te}(\mathrm{X})$, but the conditional distribution $p(y \,|\, \mathrm{X})$ remains unchanged.

3. IMPORTANCE WEIGHTING TECHNIQUES FOR COVARIATE SHIFT ADAPTATION

In this section, we show how to cope with covariate shift.

Importance Sampling: In the absence of covariate shift, the expectation over test samples can be computed by the expectation over training samples since they are drawn from the same distribution. However, under covariate shift, the difference of input distributions should be explicitly taken into account. A basic technique for compensating for the distribution change is *importance sampling*, i.e., the expectation over training samples is weighted according to their importance in the test distribution. Indeed, based on the importance weight

$$w(\mathrm{X}) = \frac{p_{te}(\mathrm{X})}{p_{tr}(\mathrm{X})},$$

the expectation of some function $F(\mathrm{X})$ over the probability density $p_{te}(\mathrm{X})$ can be computed by

$$\mathrm{E}_{p_{te}(\mathrm{X})}[F(\mathrm{X})] = \mathrm{E}_{p_{tr}(\mathrm{X})}[F(\mathrm{X})w(\mathrm{X})].$$

Importance Weighted Kernel Logistic Regression: If the importance sampling technique is applied in KLR, we have the following *importance weighted KLR (IWKLR)*:

$$\widetilde{\mathcal{P}}_{\delta}^{\log}(\mathrm{V}; \mathcal{Z}) = -\sum_{i=1}^{n} w(\mathrm{X}_i) \log P(y_i \,|\, \mathrm{X}_i; \mathrm{V}) + \frac{\delta}{2}\mathrm{tr}(\mathrm{V}\mathrm{K}\mathrm{V}^{\top}).$$

IWKLR is consistent even under covariate shift [4]:

$$\lim_{n \to \infty} \widetilde{\mathrm{V}} = \mathrm{V}^*,$$

where $\widetilde{\mathrm{V}}$ is the parameter learned by IWKLR and V^* is the optimal parameter given by

$$\mathrm{V}^* = \operatorname*{argmin}_{\mathrm{V}} \iint I(y = \widehat{y}(\mathrm{X}; \mathrm{V}))p(y \,|\, \mathrm{X})p_{te}(\mathrm{X})\mathrm{d}y\mathrm{d}\mathrm{X}.$$

Note that $\widetilde{\mathcal{P}}_{\delta}^{\log}(\mathrm{V}; \mathcal{Z})$ is still convex and thus the global solution can be obtained by the Newton method.

Importance Weighted Cross Validation: IWKLR includes the Gaussian width σ and the regularization parameter δ as tuning parameters. Here, we introduce *important weighted cross validation (IWCV)* [6] for model selection: the k-fold IWCV (kIWCV) score is given by

$$\widetilde{R}_{kIWCV}^{\mathcal{Z}} = \frac{1}{k}\sum_{j=1}^{k} \frac{1}{|\mathcal{Z}_j|} \sum_{(\mathrm{X},y) \in \mathcal{Z}_j} w(\mathrm{X}) I(y = \widehat{y}(\mathrm{X}; \mathcal{Z}_j)).$$

Even under covariate shift, kIWCV is almost unbiased [6]:

$$\mathrm{E}_{\mathcal{Z}}\left[\widetilde{R}_{kIWCV}^{\mathcal{Z}} - R^{\mathcal{Z}}\right] \approx 0,$$

where $R^{\mathcal{Z}}$ is the expected prediction error defined by

$$R^{\mathcal{Z}} = \iint I(y = \widehat{y}(\mathrm{X}; \mathcal{Z}))p(y \,|\, \mathrm{X})p_{te}(\mathrm{X})\mathrm{d}y\mathrm{d}\mathrm{X}.$$

Importance Weight Estimation: As shown above, the importance weight $w(\mathrm{X})$ plays a central role in covariate shift adaptation. However, the importance weight is usually unknown, so it needs to be estimated from samples. Here, we assume that in addition to the training input samples $\mathcal{X}^{tr} = \{\mathrm{X}_i^{tr}\}_{i=1}^{n_{tr}}$ drawn independently from $p_{tr}(\mathrm{X})$, we are given unlabeled test samples $\mathcal{X}^{te} = \{\mathrm{X}_i^{te}\}_{i=1}^{n_{te}}$ drawn independently from $p_{te}(\mathrm{X})$ (i.e., the semi-supervised setup).

Under the semi-supervised setup, the importance weight may be simply estimated by estimating $p_{tr}(\mathrm{X})$ and $p_{te}(\mathrm{X})$ from training and test samples and then taking their ratio. However, density estimation is known to be a hard problem and taking the ratio of estimated quantities tends to magnify the estimation error. Thus such a two-shot process may not be reliable in practice. Below, we introduce a method called the *Kullback Leibler Importance Estimation Procedure (KLIEP)* [5], which allows us to directly learn the importance weight function without going through density estimation.

Let us model the importance function $w(\mathrm{X})$ by the following linear model:

$$\widehat{w}(\mathrm{X}) = \sum_{l=1}^{b} \alpha_l \varphi(\mathrm{X}, \mathrm{C}_l),$$

where $\{\alpha_l\}_{l=1}^{b}$ are parameters to be learned from data samples, $\{\mathrm{C}_l\}_{l=1}^{b}$ are template points randomly chosen from the test input set $\{\mathrm{X}_i^{te}\}_{i=1}^{n_{te}}$, and $\varphi(\mathrm{X}, \mathrm{X}')$ is a basis function chosen as

$$\varphi(\mathrm{X}, \mathrm{X}') = \frac{1}{NN'}\sum_{i=1}^{N}\sum_{i'=1}^{N'} \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2}{2\tau^2}\right).$$

We determine the coefficient $\{\alpha_l\}_{l=1}^b$ by maximum likelihood estimation, which is formulated as

$$\max_{\{\alpha_l\}_{l=1}^b} \left[\sum_{i=1}^{n_{te}} \log \left(\sum_{l=1}^{b} \alpha_l \varphi(\mathrm{X}_i^{te}, \mathrm{C}_l) \right) \right]$$
$$\text{s.t.} \sum_{i=1}^{n_{tr}} \sum_{l=1}^{b} \alpha_l \varphi(\mathrm{X}_i^{tr}, \mathrm{C}_l) = n_{tr} \text{ and } \alpha_1, \ldots, \alpha_b \geq 0.$$

This optimization problem is convex and thus the global solution may be obtained by simply performing gradient ascent and feasibility satisfaction iteratively. Note that the solution $\{\widehat{\alpha}_l\}_{l=1}^b$ tends to be sparse, which contributes to reducing the computational cost in the test phase.

4. EXPERIMENTS

In this section, we report the results of speaker identification in the light of covariate shift adaptation.

Training and test samples were collected from 10 male speakers. The pronounced sentences are common to all speakers, but the test sentences are different from those for training. Moreover, the utterance samples for training were recorded in 1990/12, while the utterance samples for testing were recorded in 1991/3, 1991/6, and 1991/9, respectively. Since the recording time is different between training and test utterance samples, the session dependent variation is expected to be included. So this would be a challenging task. We used three sentences for training and five sentences for the test, where the average duration of the sentences is about 4[s].

The input utterance is sampled at 16kHz with close-talking microphone. A feature vector consists of 26 components: 12 MFCCs, the normalized log energy, and their first derivatives. Feature vectors are derived at every 10[ms] over the Hamming-windowed speech segment of 25.6[ms]. We divide each utterance sequence into 300[ms] disjoint segments, each of which corresponds to a set of features of size $\mathrm{X}_i \in \mathbb{R}^{26\times 30}$. We compute the speaker identification rate at every 1.5[s] and judge the speaker ID at time t based on the average posterior probability $\frac{1}{5}\sum_{i=1}^{5} p(\mathrm{Y}_{t-i} \mid \mathrm{X}_{t-i}; \mathrm{V})$.

We compare KLR and IWKLR in terms of speaker identification for 1991/3, 1991/6, and 1991/9 [9]. For KLR training, we only use the 1990/12 dataset (inputs $\mathcal{X}^{tr}$ and their labels), where the Gaussian width σ and the regularization parameter δ are selected based on 5-fold CV. For IWKLR training, we use unlabeled samples $\mathcal{X}^{te1}$, $\mathcal{X}^{te2}$, and $\mathcal{X}^{te3}$ in addition to the training inputs $\mathcal{X}^{tr}$ and their labels. We first estimate the importance weight from training and test data pairs ($\mathcal{X}^{tr}$, $\mathcal{X}^{te1}$), ($\mathcal{X}^{tr}$, $\mathcal{X}^{te2}$), or ($\mathcal{X}^{tr}$, $\mathcal{X}^{te3}$) by KLIEP, and then use 5-fold IWCV to decide the Gaussian width σ and regularization parameter δ.

Table 1 summarizes the speaker identification rates, showing that IWKLR+IWCV outperforms KLR+CV for all sessions. This result implies that importance weighting is useful in coping with the influence of non-stationarity in practical speaker identification such as utterance variation, the recording environment change, and physical condition/emotion. Therefore, we conclude that IWKLR+IWCV is a novel promising approach to handling session dependent variation.

Table 1. Identification rates in percent. IWKLR+IWCV refers to IWKLR with σ and δ chosen by 5-fold IWCV, and KLR+CV refers to KLR with σ and δ chosen by 5-fold CV. Values of chosen σ and δ are described in the bracket.

Test date	IWKLR+IWCV	KLR+CV
1991/3	**86.8** (1.2, 0.0001)	86.1 (1.2, 0.0001)
1991/6	**83.9** (1.3, 0.0001)	82.0 (1.2, 0.0001)
1991/9	**92.0** (1.2, 0.0001)	91.6 (1.2, 0.0001)
Average	**87.6**	86.6

5. CONCLUSIONS

In this paper, we proposed a novel semi-supervised speaker identification method that can alleviate the influence of non-stationarity such as session dependent variation, the recording environment change, and physical condition/emotion. we conducted a text-independent speaker identification simulation and experimentally found that the covariate shift formulation is useful in dealing with session dependent variations.

6. REFERENCES

[1] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing* 10(1):19–41, 2000.

[2] J. Mariethoz and S. Bengio, "A kernel trick for sequences applied to text-independent speaker verification systems", *Pattern Recognition*, 40(8):2315–2324, 2007.

[3] T. Matsui and K. Tanabe, "Comparative Study of Speaker Identification Methods: dPLRM, SVM, and GMM", *IEICE Transactions on Information and Systems*, E89-D(3):1066–1073, 2006.

[4] H. Shimodaira. "Improving predictive inference under covariate shift by weighting the log-likelihood function", *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[5] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, "Direct importance estimation for covariate shift adaptation", *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008

[6] M. Sugiyama, M. Krauledat, K.-R. Müller, "Covariate shift adaptation by importance weighted cross validation", *Journal of Machine Learning Research*, 8:985–1005, 2007.

[7] L. Rabiner and B-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993

[8] K. Tanabe, "Penalized logistic regression machines: New methods for statistical prediction 1", Technical Report 143, Institute of Statistical Mathematics, 2001.

[9] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition", *Proceedings of ICASSP'93*, II:391–394, 1993.