AUTOMATIC AUDIO TAGGING USING COVARIATE SHIFT ADAPTATION

[1]*Gordon Wichern, Makoto Yamada,* [1]*Harvey Thornburg,* [2]*Masashi Sugiyama, and* [1]*Andreas Spanias*

[1]SenSIP Center, School of Arts, Media, and Engineering, Arizona State University
[2]Department of Computer Science, Tokyo Institute of Technology

e-mail: Gordon.Wichern@asu.edu makoto.05@engalumni.colostate.edu sugi@cs.titech.ac.jp

ABSTRACT

Automatically annotating or tagging unlabeled audio files has several applications, such as database organization and recommender systems. We are interested in the case where the system is trained using clean high-quality audio files, but most of the files that need to be automatically tagged during the test phase are heavily compressed and noisy, perhaps because they were captured on a mobile device. In this situation we assume the audio files follow a covariate shift model in the acoustic feature space, i.e., the feature distributions are different in the training and test phases, but the conditional distribution of labels given features remains unchanged. Our method uses a specially designed audio similarity measure as input to a set of weighted logistic regressors, which attempt to alleviate the influence of covariate shift. Results on a freely available database of sound files contributed and labeled by non-expert users, demonstrate effective automatic tagging performance.

Index Terms— Importance, KLIEP, Acoustic signal analysis, Database query processing

1. INTRODUCTION

A fundamental challenge limiting widespread acceptance of large, freely available internet audio archives is the difficulty in organizing and accessing them. Traditionally, text-based retrieval approaches are often employed, with the disadvantage that it is impossible to search for unlabeled sound files. To overcome this there has been much recent interest in retrieving unlabeled audio from text queries and the related problem of *auto-tagging*, i.e., the ability to automatically describe and label a sound clip based on its audio content.

An additional challenge in auto-tagging general audio (non-music or speech) archives where the sound files are contributed by a user-community is a large variation in the types of files that are contributed. For example, an original collection might contain exclusively CD quality (44.1kHz, 16 bit) or better sound files recorded on expensive equipment which is used for training the auto-tagging system. When new sound files are contributed, that are possibly highly compressed, e.g., low bit-rate MP3s and recorded with different equipment these files could be very different in an audio feature space and difficult for the system to automatically annotate. This situation is likely to arise when users contribute audio files captured on their mobile phones to a previously trained automatic tagging system.

Recent approaches to auto-tagging have achieved much success including an ontological framework for connecting words and sounds [1], a generative model for annotation and retrieval of music and sound effects [2], tagging unlabeled sounds with the labels belonging to their nearest neighbor in an acoustic feature space [3], and a large scale comparative study of text-based audio retrieval [4]. These approaches assume that training and test data follow the same distribution, which might not be the case when clean, uncompressed audio is used for training, while low-quality compressed audio possibly captured via a mobile phone is used for testing. To overcome these possible shifts in the audio feature space we propose a semi-supervised learning framework, where training and test data are related by a covariate shift model, which has been successfully applied to improve session variation in speaker identification in [5, 6]. In covariate shift the input distributions are different in the training and test phases but the conditional distribution of labels (tags) remains unchanged.

In the present work we approximate trends in low-level audio feature trajectories (does each feature stay constant, go up, down, or vary in more complex ways), and use the distance between these low-level feature trends as a kernel function in a kernel logistic regression classification scheme. We then apply importance weights [7] to overcome possible shifts in the audio feature space between the training and test sets, where weights are estimated using the *Kullback-Leibler importance estimation procedure* [8]. We test the performance of our algorithm using data from the *Freesound* project [9], a database of freely available sound recordings uploaded by users of the site. We train classifiers using only uncompressed high bit rate audio files, while our testing data contains only low bit-rate compressed files. Results demonstrate how the proposed method improves retrieval performance.

This material is based upon work supported by the National Science Foundation under Grant No. 0504647, MEXT Grant-in-Aid for Young Scientists (A), 20680007, SCAT, and AOARD.

2. MEASURING AUDIO SIMILARITY

Methods for ranking sounds in terms of perceptual similarity, typically begin with the problem of acoustic feature extraction. We compute our features using 40ms Hamming windowed frames hopped every 20ms, and our chosen six-dimensional feature set is described in [10]. This feature set consists of *RMS level*, Bark-weighted *spectral centroid*, *spectral sparsity* (the ratio of ℓ^∞ and ℓ^1 norms calculated over the short-time Fourier Transform (STFT) magnitude spectrum), *transient index* (the ℓ^2 norm of the difference of Mel frequency cepstral coefficients (MFCC's) between consecutive frames), *harmonicity* (a probabilistic measure of whether or not the STFT spectrum for a given frame exhibits a harmonic frequency structure), and *temporal sparsity* (the ratio of ℓ^∞ and ℓ^1 norms calculated over all short-term RMS levels computed in a one second interval). In addition to its relatively low dimensionality this feature set is also specifically tailored to natural and environmental sounds while not being specifically adapted to a particular class of sounds (e.g., speech), which is incredibly important in diverse user-contributed audio databases.

Once the feature extraction process is complete, we represent a given sound file as $\mathrm{X} = \{\mathrm{G}, \lambda\}$, where $\mathrm{G} = [\mathbf{g}_1^\top, \ldots, \mathbf{g}_d^\top] \in \mathbb{R}^{d\times N}$ is the feature matrix with d the number of audio features, N the number of frames, and $^\top$ denotes the transpose. Utilizing the procedure in [11] we denote by λ a hidden Markov model (HMM) automatically created for every sound file by fitting constant, linear, and parabolic least squares (LS) polynomials to each feature trajectory. The first step in our similarity procedure is then to obtain the log-likelihood values $L(X_i, X_j) = \log p(X_i|\lambda_j)$ by computing the likelihood of the ith observation trajectory (feature matrix) using the jth HMM. Details on the estimation of λ and computation of the likelihood using a HMM is described in detail in [11]. Following [12] we compute the similarity between sounds X_i and X_j as

$$B(X_i, X_j) = \frac{1}{N_i}[L(X_i, X_i) - L(X_i, X_j)] + \frac{1}{N_j}[L(X_j, X_j) - L(X_j, X_i)]. \quad (1)$$

Although the semi-metric in (1) does not satisfy the triangle inequality, its properties are: (a) symmetry $B(X_i, X_j) = B(X_j, X_i)$, (b) non-negativity $B(X_i, X_j) \geq 0$, and (c) distinguishability $B(X_i, X_j) = 0$ iff $i = j$.

As a final step in our audio similarity procedure, we use *local scaling* [13] to create an affinity or Gram matrix:

$$K(X_i, X_j) = \exp(-B^2(X_i, X_j)/\sigma_i\sigma_j), \quad (2)$$

where $\sigma_i = B(X_i, X_{i_M})$ and i_M is the Mth nearest neighbor of sound i ($M = 7$ in this work) and σ_j is defined similarly. Local scaling offers more flexibility in cases where acoustic feature sets exhibit multi-scale behavior [14].

3. FORMULATION OF AUTOMATIC TAGGING PROBLEM

We consider a vocabulary $\mathcal{T}$ containing $|\mathcal{T}|$ possible tags, whose elements are denoted by $t_l \in \mathcal{T}$. For each sound file $\mathrm{X}_i \in \mathcal{X}$, we use the binary vector $\mathbf{y}_i$ of dimension $|\mathcal{T}|$ as a label vector. The elements of the label vector are $y_{i,l} = 1$ if tag t_l is relevant to the sound indexed by X_i and $y_{i,l} = 0$ otherwise. For training we are given a set of n tagged audio clips $\mathcal{Z} = \{(\mathrm{X}_i, \mathbf{y}_i)\}_{i=1}^n$. In an effective auto-tagging system relevant tags should have a higher score than irrelevant ones, i.e.,

$$F(t_l, \mathrm{X}_i) > F(t_j, \mathrm{X}_i) \quad y_{i,l} = 1, \;\; y_{i,j} = 0,$$

where $F(t, X) \in \mathbb{R}$ is a scoring function. As our scoring function we use the approximate class-posterior probability

$$F(t_l, \mathrm{X}) = p(y_{i,l} = 1 \,|\, \mathrm{X}; \mathrm{V}^l) = \frac{\exp f^l_{\boldsymbol{v}_1}(\mathrm{X})}{\exp f_{\boldsymbol{v}^l_0}(\mathrm{X}) + \exp f_{\boldsymbol{v}^l_1}(\mathrm{X})},$$

where $\mathrm{V}^l = [\boldsymbol{v}^l_0, \boldsymbol{v}^l_1]^\top \in \mathbb{R}^{2\times n}$ is a parameter to be estimated for each word, while $f_{\boldsymbol{v}^l_1}$ and $f_{\boldsymbol{v}^l_0}$ are discriminant functions for tag l corresponding to relevant and irrelevant, respectively. 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 $f_{\boldsymbol{v}^l_k}$:

$$f_{\boldsymbol{v}^l_k}(\mathrm{X}) = \sum_{i=1}^n v^l_{k,i}\mathcal{K}(\mathrm{X}, \mathrm{X}_i) \quad k = 0, 1$$

where $\boldsymbol{v}^l_k = (v_{k,1}, \ldots, v_{k,n})^\top \in \mathbb{R}^n$ are parameters corresponding to tag l and $\mathcal{K}(\mathrm{X}, \mathrm{X}')$ is the affinity function from (2) used here as a kernel function.

To estimate the parameters V^l of each word-level classifier we use maximum likelihood estimation. The negative regularized log-likelihood function $\mathcal{P}^{\log}_\delta(\mathrm{V}^l; \mathcal{Z})$ for the kernel logistic regression model is given by

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

where $\frac{\delta^l}{2}\mathrm{tr}(\Gamma^l \mathrm{V}^l \mathrm{K}(\mathrm{V}^l)^\top)$ is a regularizer introduced to avoid overfitting, $\mathrm{K} = [\mathcal{K}(\mathrm{X}_i, \mathrm{X}_j)]^n_{i,j=1}$ is the Gram matrix, and Γ^l is a 2×2 diagonal matrix whose nonzero elements contain the ratio of training samples with and without tag l, respectively. Since $\mathcal{P}^{\log}_{\delta^l}(\mathrm{V}^l; \mathcal{Z})$ is a convex function with respect to V^l its unique minimizer can be obtained by, e.g., the Newton method. The values of the regularization parameters δ^l are determined automatically using a three-fold cross validation procedure [5, 6].

4. IMPORTANCE WEIGHTING TECHNIQUES FOR COVARIATE SHIFT ADAPTATION

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.

Importance Sampling: 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

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

where $p_{te}(\mathrm{X})$ and $p_{tr}(\mathrm{X})$ are test and training input densities, the expectation of some function $Q(\mathrm{X})$ over the probability density $p_{te}(\mathrm{X})$ can be computed by

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

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

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

Note that $\widetilde{\mathcal{P}}^{\log}_{\delta^l}(\mathrm{V}^l; \mathcal{Z})$ is still convex and thus the global solution can be obtained, e.g., by the Newton method. Here, three-fold *importance weighted cross validation (IWCV)* [15] is used to determine the parameters δ_l.

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)* [8], 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:

$$\widehat{w}(\mathrm{X}) = \sum_{k=1}^{b} \alpha_k \varphi(\mathrm{X}, \mathrm{C}_k),$$

where $\{\alpha_k\}_{k=1}^{b}$ are parameters to be learned from data samples, $\{\mathrm{C}_k\}_{k=1}^{b}$ 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 the locally scaled kernel function (2). We determine the coefficient $\{\alpha_k\}_{k=1}^{b}$ by maximum likelihood estimation, which is formulated as

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

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 $\{\widehat{\alpha}_k\}_{k=1}^{b}$ tends to be sparse, which contributes to reducing the computational cost in the test phase.

5. EXPERIMENTS

In this section we test the performance of automatic audio tagging when a system is trained on a small number of clean uncompressed audio files and tested using low-quality compressed audio files. Our dataset consists of sound files from the *Freesound* project website [9]. The sound files were randomly selected from among all files on the site containing any of the 50 most used tags and between 3-60 seconds in length. We then used 193 randomly selected uncompressed audio files with a sampling rate greater than 44.1kHZ as our training set, and 1612 randomly selected audio files stored in a compressed format for testing. Our tag vocabulary consists of the $|\mathcal{T}| = 129$ tags that appeared on at least four sounds in the testing set and one sound in the training set. As a baseline system we estimate the score function parameters using KLR, and compare the automatic tagging performance to IWKLR with covariate shift.

Given an unlabeled sound from the testing set all 129 tags in the vocabulary are ranked in order of decreasing score. A tag is considered relevant if it was used by actual Freesound users. We then truncate the ranked list to the top L words and compute *recall* as the number of relevant tags ranked in the top L divided by the total number of relevant tags, and *precision* as the number of relevant tags ranked in the top L divided by L. It is trivial to maximize either precision or recall independently, but a truly effective automatic tagging system should achieve both objectives simultaneously. Figure 1 displays the precision-recall curves averaged over all 1612 sounds in the test set for both KLR and IWKLR where a single precision-recall pair is obtained at each position in the ranked list. From Figure 1 we see that IWKLR with covariate shift improved precision most dramatically at recall values between 0.2 and 0.4. This indicates that if we decide to automatically tag the test sound with all tags above a certain

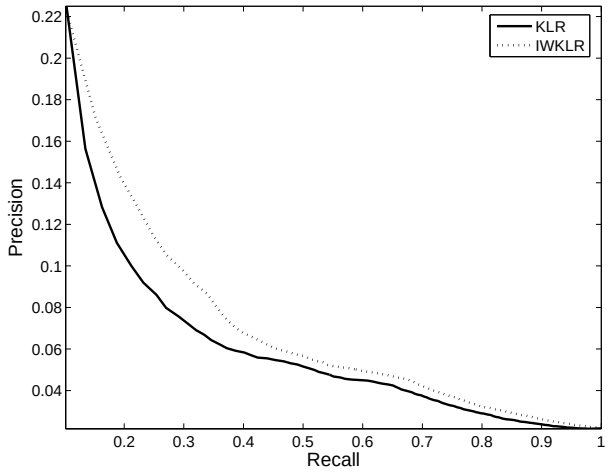

Fig. 1. Precision-recall curves for baseline KLR and IWKLR.

Table 1. Mean average precision (AP), mean area under the ROC curve (AROC), and total AROC for all 1612 compressed testing files.

	KLR	IWKLR
Mean AP	0.205	0.221
Mean AROC	0.729	0.735
Total AROC	0.643	0.667

position in the ranked list to include on average between 20-40% of relevant tags, fewer of the automatically applied tags will be irrelevant when using IWKLR as compared to KLR.

Table 1 numerically compares the performance of the baseline KLR automatic tagging system to the proposed system using IWKLR with covariate shift. Average precision (AP) is found by averaging the precision values at all points in the ranked list where a relevant tag is located. The area under the receiver operating characteristics curve (AROC) is found by integrating the ROC curve, which plots the true positive versus false positive rate for the ranked list of output tags. We compute the AP and AROC separately for each sound in the test set, and then average over all sounds in the test set to obtain the mean AP and mean AROC values. The total AROC values are obtained by integrating a ROC curve for all testing sounds simultaneously. In Table 1 IWKLR outperforms KLR for all metrics, which implies that importance weighting is useful in coping with the influence of recording environment change and lossy file compression in automatic audio tagging.

6. CONCLUSIONS

In this paper we demonstrated a semi-supervised approach to automatic tagging of general audio files under a covariate shift assumption. We evaluated the proposed approach on a testing set of 1612 audio files tagged by actual users of an internet audio archive. While initial results were promising, possible topics of future work include optimally choosing a small yet representative training set to keep the computational cost low during testing, while improving automatic tagging performance.

7. REFERENCES

[1] G. Wichern, H. Thornburg, and A. Spanias, "Unifying semantic and content-based appoaches for retrieval of environmental sounds," in *IEEE WASPAA*, New Paltz, NY, 2009.

[2] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, pp. 467–476, 2008.

[3] E. Martinez, O. Celma, M. Sordo, B. de Jong, and X. Serra, "Extending the folksonomies of freesound.org using content-based audio analysis," in *Sound and Music Computing Conference*, Porto, Portugal, 2009.

[4] G. Chechik, E. Le, M. Rehn, S. Bengio, and D. Lyon, "Large-scale content-based audio retrieval from text queries," in *ACM MIR*, Vancouver, BC, 2008.

[5] M. Yamada, M. Sugiyama, and T. Matsui, "Covariate shift adaptation for semi-supervised speaker identification," in *IEEE ICASSP*, Taipei, Taiwan, Apr. 19–24 2009, pp. 1661–1664.

[6] ——, "Semi-supervised speaker identification under covariate shift," *Signal Processing*, 2010, to appear.

[7] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of Statistical Planning and Inference*, vol. 90, no. 2, pp. 227–244, 2000.

[8] M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2008, pp. 1433–1440.

[9] "Freesound," `http://www.freesound.org`.

[10] G. Wichern, H. Thornburg, B. Mechtley, A. Fink, K. Tu, and A. Spanias, "Robust multi-feature segmentation and indexing for natural and environmental sounds," in *IEEE CBMI*, Bordeaux, Fr., 2007.

[11] G. Wichern, J. Xue, H. Thornburg, and A. Spanias, "Distortion-aware query by example for environmental sounds," in *IEEE WASPAA*, New Paltz, NY, 2007.

[12] B. H. Huang and L. R. Rabiner, "A probabilistic distance measure for hidden Markov models," *AT&T Tech. Journal*, vol. 64, pp. 1251–1270, 1985.

[13] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems*, Whistler, BC, 2004.

[14] J. Xue, G. Wichern, H. Thornburg, and A. S. Spanias, "Fast query by example of environmental sounds via robust and efficient cluster-based indexing," in *IEEE ICASSP"*, Las Vegas, NV, 2008.

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