A Kernel-based Discrimination Framework for Solving Hypothesis Testing Problems with Application to Speaker Verification

Yi-Hsiang Chao[1,2], Wei-Ho Tsai[3], Hsin-Min Wang[1], and Ruei-Chuan Chang[1,2]
[1] *Institute of Information Science, Academia Sinica, Taipei, Taiwan*
[2] *Dept. of Computer Science, National Chiao Tung University, Hsinchu, Taiwan*
[3] *Dept. of Electronic Engineering, National Taipei University of Technology, Taipei, Taiwan*
E-mail: {yschao, whm}@iis.sinica.edu.tw, whtsai@en.ntut.edu.tw, rc@cc.nctu.edu.tw

Abstract

Real-word applications often involve a binary hypothesis testing problem with one of the two hypotheses ill-defined and hard to be characterized precisely by a single measure. In this paper, we develop a framework that integrates multiple hypothesis testing measures into a unified decision basis, and apply kernel-based classification techniques, namely, Kernel Fisher Discriminant (KFD) and Support Vector Machine (SVM), to optimize the integration. Experiments conducted on speaker verification demonstrate the superiority of our approaches over the predominant approaches.

1. Introduction

In many practical applications, one may be faced with the problem of making a binary decision, such as "yes/no" or "accept/reject", with respect to an uncertain hypothesis which is known only through its observable consequences. Under a statistical framework, the problem is generally formulated as a test between a null hypothesis H_0 and an alternative hypothesis H_1 regarding some measurement $L(\cdot)$ on a given observation X:

$$\begin{aligned} H_0 &: L(X) \geq \theta \\ H_1 &: L(X) < \theta, \end{aligned} \tag{1}$$

where θ is the decision threshold. Depending on the applications, a number of measurements have been investigated, with the Likelihood Ratio (LR) measure in conjunction with parametric modeling being the most popular. Specifically, each of the hypotheses is represented by a set of probability-related parameters through a training process, and the probability of generating a given observation is then evaluated for each of the hypotheses' parameter sets.

However, in most applications, the alternative hypothesis is usually ill-defined and hard to be characterized precisely. One example is the problem of speaker verification, which aims to determine if a speaker is who he or she claims to be. Though the null hypothesis can be modeled straightforwardly using speech utterances from the speaker claimed by the test user, the alternative hypothesis does not involve any specific speaker, and thus lacks explicit data to model. Many approaches have thus been proposed in attempts to characterize the alternative hypothesis effectively and robustly, but none of them has been proven optimal. The pros and cons of the individual approaches motivate us to try to develop a framework that integrates multiple LR measures into a unified decision basis. To enable a reliable integration, this study formulates the hypothesis test as a problem of non-linear discrimination and applies kernel-based techniques, namely, Kernel Fisher Discriminant (KFD) [6] and Support Vector Machine (SVM) [7], to optimally separate the LR samples of the null hypothesis from those of the alternative hypothesis.

2. Hypothesis testing measures

From a speaker-verification point of view and without loss of generality, LR measure for a hypothesis testing problem comes in many choices. One simple approach [2] is to pool all speech data from a great amount of speakers, generally irrelevant to the clients, and train a single speaker-independent model Ω, named the world model. During a test, the possibility of an unknown utterance U being produced by the claimed speaker can be evaluated by

$$L_1(U) = \log p(U \mid \lambda) - \log p(U \mid \Omega), \tag{2}$$

where λ is the model trained using speech from the claimed speaker. Conceivably, the larger the value of $L_1(U)$, the more likely the utterance U is produced by the claimed speaker.

Instead of using a single model, an alternative way is to train a set of models $\{\lambda_1, \lambda_2, \ldots, \lambda_B\}$ using speech from several representative speakers, called cohort [3], which simulates the potential impostors. This gives the following possibilities in computing LR:

(i) picking the likelihood of the most competitive model [4], i.e.,

$$L_2(U) = \log p(U \mid \lambda) - \max_{1 \le i \le B} \log p(U \mid \lambda_i), \quad (3)$$

(ii) averaging the likelihoods of the B cohort models arithmetically [1], i.e.,

$$L_3(U) = \log p(U \mid \lambda) - \log\left\{\frac{1}{B}\sum_{i=1}^{B} p(U \mid \lambda_i)\right\}, \quad (4)$$

(iii) averaging the likelihoods of the B cohort models geometrically [5], i.e.,

$$L_4(U) = \log p(U \mid \lambda) - \frac{1}{B}\sum_{i=1}^{B} \log p(U \mid \lambda_i). \quad (5)$$

However, none of the LR measures above has been shown absolutely superior to the others. Usually, $L_1(U)$ tends to be weak in rejecting the impostors with voices similar to the client's, while $L_2(U)$ is prone to falsely rejecting a client speaker, and $L_3(U)$ and $L_4(U)$ are between these two extremes. The pros and cons of different LR measures motivate us to combine them into a unified framework by virtue of the complementary information that each LR can contribute.

Given N different LR measures $L_i(U)$, $i = 1, 2, \ldots, N$, we first normalize each of them to a value between 0 and 1 via a sigmoid function $S(v) = 1/[1+\exp(-av)]$, where a is a scalar. Let $\Lambda_i(U) = S(L_i(U))$. We define a combined LR measure $f(U)$ by,

$$\begin{aligned} f(U) &= w_1\Lambda_1(U) + \ldots + w_N\Lambda_N(U) + b \\ &= \mathbf{w}^T\mathbf{x} + b \\ &= \Psi(\mathbf{x}) \begin{cases} \ge \theta & \text{accept} \\ < \theta & \text{reject}, \end{cases} \end{aligned} \quad (6)$$

where $\mathbf{x} = [\Lambda_1(U), \Lambda_2(U), \ldots, \Lambda_N(U)]^T$ is an $N \times 1$ vector in the space R^N, $\mathbf{w} = [w_1, w_2, \ldots, w_N]^T$ is an $N \times 1$ weight vector, and b is a bias. $\Psi(\mathbf{x})$ in Eq. (6) forms a so-called linear discriminant classifier. This classifier translates the goal of solving a hypothesis testing problem into the optimization of $\mathbf{w}$ and b, such that the utterances from the clients and impostors can be separated. To realize this classifier, three distinct data sets are needed, one for generating each client's model, another for generating the world model or cohort models, and the other for optimizing $\mathbf{w}$ and b.

3. Kernel-based discrimination

Intuitively, $\Psi(\mathbf{x})$ in Eq. (6) could be solved via Fisher's Linear Discriminant (FLD) [6]. However, such a method is built upon the assumption that the observed data from different classes is linearly separable, which is obviously not adequate in most practical cases with nonlinearly separable data. To solve this problem more effectively, we propose using a kernel-based nonlinear discriminant classifier. It is hoped that the data from different classes, which is not linearly separable in the original input space R^N, can be separated linearly in a certain higher dimensional (maybe infinite) feature space F via a nonlinear mapping Φ. Let $\Phi(\mathbf{x})$ denote a vector obtained by mapping $\mathbf{x}$ from R^N to F. The objective based on Eq. (6) can be re-defined as,

$$\Psi(\mathbf{x}) = \mathbf{w}^T\Phi(\mathbf{x}) + b \ , \quad (7)$$

which constitutes a linear discriminant classifier in F.

In practice, it is usually difficult to know what kind of mapping is applicable, and therefore the computation of $\Phi(\mathbf{x})$ can be infeasible. To overcome this difficulty, a promising way is to characterize the relationship between data in F, instead of computing $\Phi(\mathbf{x})$ directly. This is done by introducing a kernel function $k(\mathbf{x},\mathbf{y}) = \langle\Phi(\mathbf{x}),\Phi(\mathbf{y})\rangle$, which is the inner product of two vectors $\Phi(\mathbf{x})$ and $\Phi(\mathbf{y})$ in F. The kernel function $k(\cdot)$ must be symmetric positive and conform to Mercer's condition [7]. Existing techniques, such as KFD or SVM, can be applied to carry out Eq. (7).

3.1. Kernel Fisher Discriminant (KFD)

The goal of KFD is to locate $\mathbf{w}$ in the feature space F that maximizes the between-class scatter while minimizes the within-class scatter. Since the solution of $\mathbf{w}$ must lie in the span of all mapped training data samples $\Phi(\mathbf{x}_j)$ in F [6], it can be expressed as,

$$\mathbf{w} = \sum_{j=1}^{l} \alpha_j \Phi(\mathbf{x}_j), \quad (8)$$

where l is the number of training data samples. Letting $\boldsymbol{\alpha}^T = [\alpha_1, \ldots, \alpha_l]_{1\times l}$, the goal is therefore changed from finding $\mathbf{w}$ to finding $\boldsymbol{\alpha}$. Accordingly, Eq. (7) can be equivalent to

$$\Psi(\mathbf{x}) = \sum_{j=1}^{l} \alpha_j k(\mathbf{x}_j, \mathbf{x}) + b. \quad (9)$$

Analogous to FLD, $\boldsymbol{\alpha}$ can be solved using a generalized eigen-decomposition algorithm [6]. The

bias b is actually the decision threshold θ in Eq. (6), which can be determined through the trade-off between false acceptance and false rejection.

3.2. Support Vector Machine (SVM)

Alternatively, Eq. (7) can be designed with SVM, in analogy to a fusion classifier proposed in [8][9]. The goal of SVM is to seek a separating hyperplane in the feature space F that maximizes the margin between classes. Following [7], $\mathbf{w}$ is expressed as,

$$\mathbf{w} = \sum_{j=1}^{l} y_j \alpha_j \Phi(\mathbf{x}_j), \tag{10}$$

which yields

$$\Psi(\mathbf{x}) = \sum_{j=1}^{l} y_j \alpha_j k(\mathbf{x}_j, \mathbf{x}) + b, \tag{11}$$

where each training sample $\mathbf{x}_j$ belongs to one of the two classes identified by the label $y_j \in \{-1,1\}$, j=1, 2,…, l. The coefficients α_j and b can be solved using the quadratic programming techniques [11]. Note that α_j is non-zero for a few support vectors, and is zero otherwise. A number of kernel functions exist, with the dot product kernel function, i.e., $k(\mathbf{x}, \mathbf{x}_j) = \mathbf{x}_j^T \mathbf{x}$, being the simplest, and the Radial Basis Function (RBF) kernel function, i.e., $k(\mathbf{x}, \mathbf{x}_j) = \exp(-\|\mathbf{x} - \mathbf{x}_j\|^2 / 2\sigma^2)$, being the most popular, where σ is a tunable parameter. The SVM with a dot product kernel function is known as Linear SVM.

4. Experiments

4.1. Experimental setup

The proposed methods were examined via speaker-verification experiments conducted on speech data extracted from the XM2VTSDB multi-modal database [12]. In accordance with "Configuration II" described in [12], the database was divided into three subsets: "Training", "Evaluation", and "Test". In our experiments, "Training" was used to build the individual client's model, while "Evaluation" was used to optimize $\mathbf{w}$ and b. Then, the performance of speaker verification was evaluated on "Test". As shown in Table 1, a total of 293 speakers[1] in the database were divided into 199 clients, 25 "evaluation impostors", and 69 "test impostors". Each speaker involved 4 recording sessions taken at approximately one-month intervals, and each recording session consisted of 2 shots. In a shot, every speaker was prompted to utter 3 sentences "0 1 2 3 4 5 6 7 8 9", "5 0 6 9 2 8 1 3 7 4", and "Joe took father's green shoe bench out".

[1] We discarded 2 speakers (ID numbers 313 and 342) because of partial data corruption.

We used 12 (3×2×2) utterances/speaker from sessions 1 and 2 to train the individual client's model. For each client, the other 198 clients' utterances from sessions 1 and 2 were used to generate the world model or cohort models. Then, we used 6 utterances/client from session 3, along with 24 (3×4×2) utterances/evaluation-impostor to optimize $\mathbf{w}$ and b. In the performance evaluation, we tested 6 utterances/client in session 4 and 24 utterances/test-impostor, which gave 1,194 (6×199) client trials and 329,544 (24×69×199) impostor trials. Each utterance, sampled at 32 kHz, was converted into a stream of 24-order feature vectors, each consisting of 12 Mel-scale cepstral coefficients and their first time derivatives, by a 32-ms Hamming-windowed frame with 10-ms shifts.

Table 1. Configuration of the speech database.

Session	Shot	199 clients	25 impostors	69 impostors
1	1	Training	Evaluation	Test
	2			
2	1			
	2			
3	1	Evaluation		
	2			
4	1	Test		
	2			

4.2. Experimental results

The LR measures, $L_1(U)$, $L_2(U)$, $L_3(U)$, and $L_4(U)$, served as our baseline systems for performance comparison, in which each speaker model was represented by a Gaussian Mixture Model (GMM) [1] with 64 mixture components, while the world model was a GMM with 256 mixture components. We implemented a combined-LR system via FLD, Linear SVM, SVM, and KFD, respectively, where SVM and KFD used an RBF kernel function with σ=0.1. 1,194 (6×199) client examples and 119,400 (24×25×199) impostor examples from "Evaluation" were used to optimize $\mathbf{w}$ and b. However, recognizing the fact that a kernel-based method can be intractable when a huge amount of training examples involves, we downsized the number of impostor examples from 119,400 to 2,250 using a uniform random selection.

Fig. 1 shows the results of speaker verification conducted on "Evaluation" with DET curves [10], obtained equivalently by adjusting the decision threshold, i.e., b or θ. Though this experiment was an inside test for our proposed framework, it can be observed that SVM and KFD perform better than FLD and Linear SVM. To verify the superiority of the combined-LR systems over the baseline systems, experiments were next conducted on "Test". The

results in DET curves are depicted in Fig. 2, where we focused on the performance improvements of SVM and KFD with respect to the baseline systems. It is clear that both the combined-LR systems, SVM and KFD, outperform the baseline systems. Further analysis of the results via the equal error rate (EER) showed that a 13.2% relative improvement was achieved by KFD (EER = 4.6%), compared to 5.3% of $L_3(U)$.

5. Conclusions

This study has presented a framework for solving a hypothesis testing problem by combining multiple likelihood-ratio measures into a unified discrimination basis. The combination has been formulated as a non-linear classification problem and solved by using the kernel-based classifiers, namely, the kernel Fisher Discriminant and Support Vector Machine. Experiments conducted on a speaker-verification task showed a notably improvement of performance with such a combination. It should be noted that the proposed framework can be applied to handle other variety of data and hypothesis testing measures.

6. Acknowledgements

This project was funded by the National Science Council, Taiwan, under grant NSC94-2213-E-001-009.

7. References

[1] D. A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models", *Speech Communication*, vol.17, 1995, pp. 91-108.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models", *Digital Signal Processing*, vol. 10, 2000, pp. 19-41.

[3] A. E. Rosenberg, J. Delong, C. H. Lee, B. H. Juang and F. K. Soong, "The use of Cohort Normalized Scores for Speaker Verification", *Proc. ICSLP1992*.

[4] A. Higgins, L. Bahler, and J. Porter, "Speaker Verification using Randomized Phrase Prompting", *Digital Signal Processing*, vol. 1, no. 2, 1991, pp. 89-106.

[5] C. S. Liu, H. C. Wang, and C. H. Lee, "Speaker Verification using Normalized Log-Likelihood Score", *IEEE Trans. Speech and Audio Processing*, vol. 4, 1996, pp. 56-60.

[6] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K. R. Müller, "Fisher Discriminant Analysis with Kernels", *Neural Networks for Signal Processing IX*, 1999, pp. 41-48.

[7] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", *Data Mining and Knowledge Discovery*, vol.2, 1998, pp. 121-167.

[8] H. T. Cheng, Y. H. Chao, S. L. Yen, C. S. Chen, H. M. Wang, and Y. P. Hung, "An Efficient Approach to Multi-Modal Person Identity Verification by Fusing Face and Voice Information", *Proc. ICME2005*.

[9] S. Ben-Yacoub, "Multi-modal Data Fusion for Person Authentication using SVM", *Proc. AVBPA1999*.

[10] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance", *Proc. Eurospeech1997*.

[11] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.

[12] J. Luettin and G. Maître, *Evaluation Protocol for the Extended M2VTS Database (XM2VTSDB)*, IDIAP-COM 98-05, IDIAP, 1998.

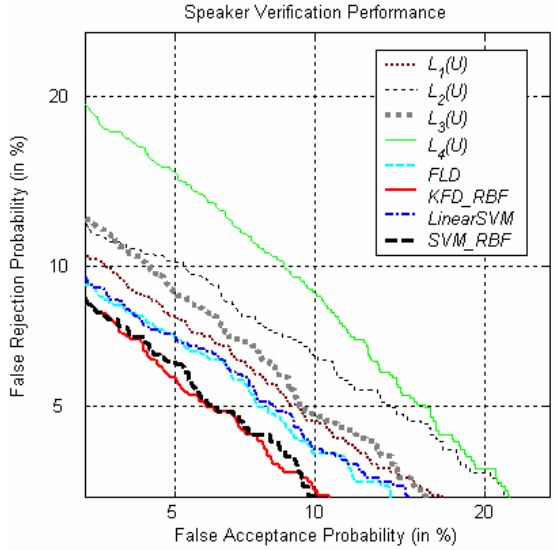

Figure 1. DET curves for "Evaluation".

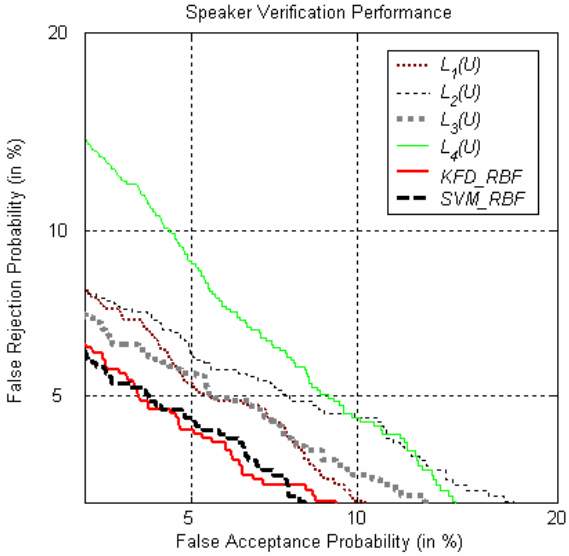

Figure 2. DET curves for "Test".