A NOVEL DATA DESCRIPTION KERNEL BASED ON ONE-CLASS SVM FOR SPEAKER VERIFICATION

Yufeng Shen and Yingchun Yang[*]

College of Computer Science and Technology,
Zhejiang University, Hangzhou, P.R. China, 310027

ABSTRACT

In this paper we develop a novel Data Description kernel based on One-Class SVM (OCSVM-DD kernel) used for text-independent SVM speaker verification. The basic idea of the new kernel is to combine the data description model OCSVM with SVM discriminant classifier. Utterances are firstly mapped to the normal vector of the separating hyperplane in OCSVM model. Then a SVM classifier with linear kernel is applied on those mapped vectors. Experiments results on NIST 2001 SRE database show that the performance of our new kernel is superior to Generalized Linear Discriminative Sequence (GLDS) kernel and comparative with UBM-MAP-GMM method.

Index Terms—Speaker verification, SVM, Kernel One-Class SVM

1. INTRODUCTION

Support Vector Machine (SVM) [1] has been widely used in Speaker Verification fields for its excellent classifying ability and generalizing capacity. The performance of SVM is comparable with those state-of-the-art classifiers such as GMMs [2], while requiring relatively less training data.

Initial Speaker recognition works using SVMs by Schmidt and Gish [3], Wan and Campbell [4] employed frame-level classification: train and test are performed on the frame level and the scores of each frame are combined to obtain the overall score of an utterance. This method has two main disadvantages: one is that the amount of frame data is too large for efficient computation; the other is that the sequence information contained in the utterance is lost when each frame is treated individually.

Due to those drawbacks of frame-level classification, utterance-based kernel methods are now the mainstream methods in SVM speaker verification fields. The basic idea of utterance-based kernel method is to map a whole utterance to a single vector in feature space and do SVM classification on those mapped vectors. Some mapping methods are straightforward such as Generalized Linear Discriminant Sequence (GLDS) kernels by Campbell [5], where mapping is done using simple polynomial expansion Some other methods rely on using data description models to map utterances such as Fisher kernels by Jaakkola and Haussler [6], Probabilistic Distance Kernels by P. Moreno and P.P.Ho [7] and Pair HMM kernels by Durbin [8].

Though GLDS kernels' mapping method through polynomial expansion is simple and cheap in computation, it actually does little in modeling of the utterance and does not extract enough feature information from utterances. On the other hand, mapping methods using data description models can benefit a lot from their data characterizing ability. Based on these observations, we develop a new kernel whose construction of feature space is similar to GLDS kernels' method while the characterizing abilities of the mapped vectors are improved by a new data description model: One-Class SVM (OCSVM) [9].

OCSVM is a variation of standard SVM which deals with the situation where only one class of example data can be obtained. The objective of OCSVM is to find a hyperplane to separate the only positive examples from the origin with maximum margin. We choose OCSVM as the data description model in kernel construction for its strong data descriptive ability. So the new kernel is called One-Class SVM based Data Description (OCSVM-DD) kernel.

This paper is organized as follows: section 2 provides some background knowledge; section 3 gives the detailed description of OCSVM-DD kernels; experimental evaluation and results are presented in section 4; finally, section 5 is the conclusion.

2. BACKGROUND KNOWLEDGE

2.1. GLDS kernels

For a sequence of observations $x_1^n : x_1, x_2, ..., x_n$ the mapping $x_1^n \rightarrow \bar{b}$ is defined as

[*] Corresponding author

$$x_1^n \to \frac{1}{n}\sum_{i=1}^{n} b(x_i) \qquad (2.3)$$

where b(x) is an expansion of the input space into a vector of scalar functions. Usually the b(x) is chosen to be the vector of polynomial basis terms of the input vector x.

Given two sequences of speech feature vectors, x_1^n and y_1^m, the GLDS kernel is defined as

$$K_{GLDS}(x_1^n, y_1^m) = \overline{b}_x^t \overline{R}^{-1} \overline{b}_y \qquad (2.4)$$

where matrix $\overline{R}$ is trained from the speech data of both speakers and imposters and in essence is used to normalize the mapped vectors. $\overline{b}_x^t$ and $\overline{b}_y$.

2.2. Data description model and discriminant classifier

For a discriminant classifier to achieve good performance, the pre-requisite is that the extracted feature vectors can convey enough information of example data. The central idea of utterance-based method in speaker verification tasks is to map the whole utterance to a single vector as the input of discriminant classifier. So a good mapping should be able to extract useful information contained in utterances and encode them into the mapped single vector.

Data description model is a good tool to implement such mapping: well-constructed descriptive model can accurately characterize the utterance features and well-selected model parameters can be used as the feature vector to represent the model.

Classic descriptive models such as GMMs and HMMs have been used in kernel construction [6] [7] [8]. Both GMMs and HMMs are probabilistic models. In the next section, we will construct our kernel using a descriptive but non- probabilistic model: One-Class SVM.

3. ONE-CLASS SVM BASED DATA DESCRIPTION KERNEL

3.1. Review on One-Class SVM

The conception of OCSVM [9] is to separate the only positive examples from the origin with maximum margin. We can view OCSVM as a descriptive model for it actually estimates the distribution of positive examples in the high-dimension space through kernel mapping.

We first introduce terminology and notation conventions. We consider training data

$$x_1, x_2, \ldots\ldots\ x_\ell \in \chi \qquad (3.1)$$

Where $\ell \in \mathrm{N}$ is the number of observations. Let Φ be a feature map $\chi \to F$, then by evaluating some simple kernel functions we can compute the inner product of the image of Φ in the feature space F

$$k(x, y) = (\Phi(x) \cdot \Phi(y)) \qquad (3.2)$$

OCSVM's objective of finding the optimal hyperplane can be formulated in a quadratic program (QP) problem

$$\min_{\omega \in F, \xi \in R^\ell, \rho \in R} \frac{1}{2}\|\omega\|^2 + \frac{1}{\nu\ell}\sum_i \xi_i - \rho \qquad (3.3)$$

Subject to

$$(\omega \cdot \Phi(x_i)) \ge \rho - \xi_i, \xi_i \ge 0 \qquad (3.4)$$

where ω is the normal vector of that separating hyperplane and parameter υ controls the trade-off between ω and slack variables ξ.

After solving this QP problem, the final decision function is

$$f(x) = \mathrm{sgn}(\sum_i \alpha_i k(x_i, x) - \rho) \qquad (3.5)$$

where all patterns x_i in equation (3.5) are support vectors.

The feature of OCSVM is that the framework of a two-class classifier is re-constructed to do the job of one-class data description. And the data characterizing ability of OCSVM is comparative with classic probabilistic models such as GMMs and HMMs.

3.2. Conception of the OCSVM-DD kernel

When substituting equation (3.2) into equation (3.5)

$$f(x) = \mathrm{sgn}(\sum_i \alpha_i k(x_i, x) - \rho) \qquad (3.5)$$

$$= \mathrm{sgn}(\sum_i \alpha_i \Phi(x_i) \cdot \Phi(x) - \rho) \qquad (3.6)$$

$$= \mathrm{sgn}((\omega \cdot \Phi(x)) - \rho) \qquad (3.7)$$

Where $\omega = \sum_i \alpha_i \Phi(x_i)$ is the normal vector of the separating hyperplane in OCSVM.

Viewed in another way, the inner product $\omega \cdot \Phi(x)$ can be thought as the similarity between the testing point x and the already trained model. Constant ρ is the threshold. So the normal vector ω is actually a weight vector, reflecting $\Phi(x)$'s each dimension's contribution to the total similarity $\omega \cdot \Phi(x)$. Or we can say that ω well characterizes the OCSVM model.

With this observation, we have good reason to believe that normal vector ω well represents the whole utterance. So comes the idea of our new kernel: mapping the utterance to the normal vector ω and then use ω as the input of SVM classifier.

From the definition $\omega = \sum_i \alpha_i \Phi(x_i)$ we can see that to compute ω the concrete form of Φ must be known first. Usually SVM performs the mapping Φ implicitly through simple kernel function and it is hard to get the concrete expression of Φ. Some special polynomial kernels are exceptions and we will use a kind of specific polynomial kernel functions to accomplish the mapping.

We define Φ_d to map $x \in R^N$ to the vector $\Phi_d(x)$ whose entries are all possible dth degree ordered products of the entries of x . Then the corresponding kernel computing the dot product of vectors mapped by Φ_d is

$$k(x,y) = \Phi_d(x) \cdot \Phi_d(y) = (x \cdot y)^d \qquad (3.8)$$

The proof is straightforward:

$$\begin{aligned}\Phi_d(x) \cdot \Phi_d(y) &= \sum_{j_1=1}^{N} \cdots \sum_{j_d=1}^{N} [x]_{j_1} \cdot \ldots \cdot [x]_{j_d} \cdot [y]_{j_1} \cdot \ldots \cdot [y]_{j_d} \\ &= \sum_{j_1=1}^{N} [x]_{j_1} \cdot [y]_{j_1} \cdots \sum_{j_d=1}^{N} [x]_{j_d} \cdot [y]_{j_d} \\ &= (\sum_{j=1}^{N} [x]_j \cdot [y]_j)^d = (x \cdot y)^d\end{aligned}$$

So if the kernel function is chosen to be the form $k(x,y) = (x \cdot y)^d$ in OCSVM, then the map Φ_d has the explicit expression and $\omega = \sum_i \alpha_i \Phi(x_i)$ can be computed explicitly.

The definition of OCSVM-DD kernel is given by:

$$k(A,B) = \omega_A \cdot \omega_B \qquad (3.9)$$

where A and B represent two utterances and ω_A and ω_B are the normal vector ω in A and B's OCSVM models.

The mapped space of OCSVM-DD kernel is similar to GLDS kernels' in that both are explicitly constructed through polynomial expansion. The difference is that for GLDS kernels, once all the frames are mapped to feature vectors, they are simply summed and averaged (see equation 2.3); while for OCSVM-DD kernel, a descriptive model OCSVM is constructed on those mapped frames and a representative vector (normal vector ω) is chosen to be the feature vector. We will see how this difference can affect the performance of SVM classifier in the next experiments section.

4. EXPERIMENTS

4.1. Database and front-end processing

Experiments are performed on the NIST2001 SRE database according to the rules of one-speaker detection evaluation described in evaluation plan [10]. In the database there are 174 target speakers of which 74 are male and 100 are female. For the training, each speaker has a speech lasting about 1~2 minutes. For the testing, there are about 2200 test segments and each is evaluated against 11 hypothesized speakers of the same sex as the segment speaker.

In the front-end processing, a 16-dimensional mel-cepstral vector is extracted from the speech signal every 16ms using a 32ms window. Delta-cepstral coefficients are then computed and appended to the cepstral vector to form a 32-dimensional feature vector. Lastly, to make the features more robust to different channel and noise effects, we also map the raw features to the standard normal distribution, using feature warping described in [11].

4.2. OCSVM-DD kernel based system

OCSVM is implemented using LIBSVM [12]. Both degree 2 and 3 polynomial kernel functions are tried and we set the penalty parameter C = 1 (the one resulting in the best performance according to experiences). In the classification of SVM, we use linear polynomial kernel and set C = 1.

In practical implementation one optimization about the mapping function Φ_d can be done: the dimension of feature space is p^d after mapping Φ_d , where p is the dimension of the original input space. Since the mapping Φ_d is ordered, there are many redundant components in the mapped vector $\Phi_d(x)$ for that many components of $\Phi_d(x)$ are the product of the same entries of x with different orders. An unordered version of $\Phi_d(x)$ in the computing of ω can reduce the dimension of feature space by a factor of about $\frac{1}{d!}$.

After the computation of ω on all utterances, normalization is preferred to control the variability between different ω of different speakers. In our experiments a simple normalization $\omega \rightarrow \frac{\omega - \mu}{\sigma}$ is used， where μ is the mean of all utterances and σ is the stand deviation computed separately along each dimension on all utterances.

4.3. Reference systems

4.3.1 GLDS kernel based SVM system

The first reference system is a SVM system with GLDS kernel. Comparison between GLDS kernel and OCSVM-DD kernel can show how modeling of input data in the mapping process can affect the performance of classifier. In experiments, we try GLDS kernels with both degree 2 and degree 3 polynomial expansion. Matrix $\overline{R}$ in equation (2.4) is trained using DEVTEST database and diagonal matrix is used.

4.3.2 UBM-MAP-GMM system

The other reference system is UBM-MAP-GMM [13] based. UBM-MAP-GMM represents the highest level technology in speaker verification field. Comparison with this state-of-the-art system can test the validation of our new kernels. In our experiments, 2048 components Gaussian Mixture Models (GMM) with diagonal covariance matrices are used. The male and female background models are trained respectively using the DEVTEST database and then each target speaker's model is derived from the corresponding background model according to a MAP criterion [14].

4.4. Results

System	EER (%)	Min DCF
GLDS Kernel (d=2)	14.2	0.068
GLDS Kernel (d=3)	11.4	0.061
OCSVM-DD Kernel (d=2)	14.0	0.063
OCSVM-DD Kernel (d=3)	9.6	0.049
UBM-MAP-GMM	10.5	0.044

Table 1. the experiment results comparing OCSVM-DD kernel with GLDS kernels and UBM-MAP-GMM, using the criterion of Equal Error Rate (EER) and minimal DCF.

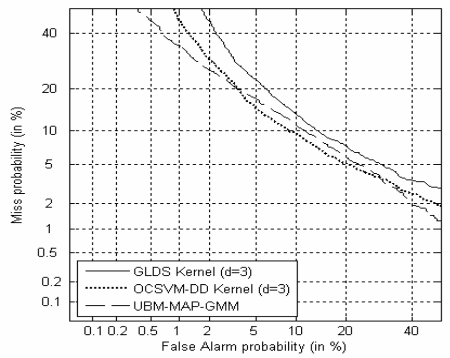

Figure 2. DET plots showing the comparison of OCSVM-DD kernel with GLDS kernel and UBM-MAP-GMM system.

Experiment Results are showed in Table 1 and the Detection Error Tradeoff curves are presented in Figure 2. The metric is Equal Error Rate and Detection Cost Function [10].

From the results we can see that OCSVM-DD kernel is superior to GLDS kernel in terms of both EER and min DCF, verifying that modeling of the utterance using OCSVM is better than the method of simple polynomial expansion used in GLDS kernel. Although the UBM-MAP-GMM system has lower Min DCF, our new kernel has better results in the EER. So the performance of OCSVM-DD kernel based system is comparative with UBM-MAP-GMM system as a whole

5. CONCLUSION

In this paper we present a new OCSVM-DD kernel applied in SVM speaker verification system. By exploiting the good modeling ability of OCSVM, our new kernel outperforms the widely used GLDS kernels and achieves comparative experiment results with UBM-MAP-GMM system. One main drawback of our new method is that it takes a long time to train an OCSVM for each utterance. So for the future work, we will focus on decreasing the time complexity of OCSVM training while improving, at least retaining, the performance of our new kernel.

Acknowledgments. This work is supported by National Science Fund for Distinguished Young Scholars 60525202, Program for New Century Excellent Talents in University NCET-04-0545 and Key Program of Natural Science Foundation of China 60533040, Zhejiang Provincial Natural Science Foundation (Y106705), National 863 Plans (2006AA01Z136).

6. REFERENCES

[1] V. N. Vapnik, Statistical Learning Theory. New York: Wiley, 1998.

[2] G. Doddington, M. Przybocki, A. Martin, and D. Reynolds, "The NIST speaker recognition evaluation-Overview, methodology, systems, results, perspective" Specch Common, vol. 31. no. 2-3,pp, 225-254, 2000

[3] M. Schmidt and H. Gish, "Speaker identification via support vector classifiers," in Proc. ICASSP, vol.1, 1996, pp.105-108

[4] V. Wan and W. M. Campbell, "Support vector machines for speaker verification and identification," in Proc, Neural Networks for Signal Processing X, 2000, pp. 775-784

[5] W. M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in Proc. ICASSP, 2002.

[6] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminant classifiers", in Advances in Neural Information Processing Systems 11, M. S. Kearns, S. A. Solla, and D. A. Cohn, Eds, MIT Press,1999.

[7] PJ Moreno, and PP Ho. A New SVM Approach to Speaker Identification and Verification Using Probabilistic Distance Kernels. in Eurospeech. 2003. Geneva, Switzerland.

[8] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, Biological Sequence Analysis, Cambridge University Press, 1998.

[9] B. Scholkopf, J. C. Platt, J. T. Shawe, A. J. Smola, R. C. Williamson, "Estimating the support of a high-dimensional Distribution", Technical Report MSR-TR-99-87, Microsoft Research

[10] "The NIST Year 2001 Speaker Recognition Evaluation Plan", http://www.nist.gov/speech/tests/spk/2001/

[11] J.Pelecanos and S.Sridharan, "Feature warping for robust speaker verification", Proc. Speaker Odyssey 2001 conference, June 2001.

[12] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[13] Frederic Bimbot "A Tutorial on Text-Independent Speaker Verification", EURASIP Journal on Applied Signal Processing 2004:4,430-451

[14] Gauvain, J. L. and Lee, C.-H., Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, IEEE Trans. Speech Audio Process. 2 (1994), 291-298