Probabilistic Speaker Identification with dual Penalized Logistic Regression Machine

Tomoko Matsui and Kunio Tanabe

The Institute of Statistical Mathematics
Tokyo, Japan
{t matsui, tanabe}@ism.ac.jp

Abstract

This paper investigates a probabilistic speaker identification method based on the dual Penalized Logistic Regression Machines (dPLRMs). The machines employ kernel functions which map an acoustic feature space to a higher dimensional space as is the case with the Support Vector Machines (SVMs). Nonlinearity in discriminating each speaker is implicitly handled in this space. While SVMs maximize the margin between two classes of data, dPLRMs maximize a penalized likelihood of a logistic regression model for multi-class discrimination. dPLRMs provide a probability estimate of each identification decision. We show that the performance of dPLRMs is competitive with that of SVMs through text-independent speaker identification experiments in which speech data recorded by 10 male speakers in four sessions are analyzed.

1. Introduction

Speaker recognition techniques are widely applied not only to access controls of information service systems but also to such problems as speaker detection problems in speech dialogue and speaker indexing problems with large audio archives. The demand has been increasing for techniques with higher-accuracy.

The conventional method for text-independent speaker recognition is based on the Gaussian mixture model (GMM) [1]. In this approach, utterance variation is well captured by a mixture of a well-chosen number of Gaussian distributions. However, the maximum likelihood (ML) estimates of the mixture distribution with the expectation maximization algorithm tend to be unreliable especially when the number of training data is relatively small. In such a case additional fine-tuning process of the estimated mixture is needed for gaining higher discriminative power [2].

SVMs [3] have been successfully applied to various pattern recognition problems: handwritten digit recognition, face detection, text categorization and speaker recognition [4-8]. Recently, Tanabe [9,10] proposed dPLRMs based on the penalized logistic regression model with a specific penalty term for bringing about induction-generalization capacity of the machine. The model yields a certain duality which leads intrinsically to the kernel regressors as is the case with quadratic programming models in SVMs. Hence comes the name with “dual” in dPLRMs. We applied dPLRMs to speaker identification problems and showed that the dPLRM-based method performed better than the conventional GMM-based method in our text-independent experiments with 10 male speakers [11]. We list several distinctions between dPLRMs and SVMs in Table 1.

In this paper, we compare dPLRMs with SVMs in text-independent speaker recognition experiments. In section 2, we briefly sketch dPLRMs. In section 3, we introduce our speaker identification procedure. In section 4, we show the speaker identification experimental results. In section 5, we demonstrate its significance as a probabilistic identifier.

| 1. dPLRM | originally designed for multi-class discrimination and can handle multi-class situations at once. |
| SVM | designed for two-class discrimination and need some special techniques for multi-class discrimination such as a pairwise coupling method. |
| 2. dPLRM | provides a probabilistic estimate. |
| SVM | gives a confidence index (‘margin’). |
| 3. dPLRM | forms discrimination boundaries depending on the whole set of the training sample vectors. |
| SVM | forms a discrimination boundary in the middle of the margin consisted of support vectors. |
| 4. dPLRM | needs to solve a simple matrix nonlinear equation. |
| SVM | needs to solve a series of inequality constrained maximization problems for multi-class situations. |

Table 1. Distinctions between dPLRMs and SVMs

2. Dual penalized logistic regression machine

Let \( \mathbf{x}_j \) is a column vector of size \( n \) and \( c_j \) takes a value in the finite set \( \{1, 2, \ldots, K\} \) of classes. The learning machine dPLRM feeds a finite number of training data \( \{(\mathbf{x}_j, c_j)\}_{j=1}^N \), and then produces a conditional multinomial distribution \( \text{M}(p^*(\mathbf{x})) \) of \( c \) given \( \mathbf{x} \in \mathbb{R}^d \), where \( p^*(\mathbf{x}) \) is a predictive probability vector whose \( k \)-th element \( p_k^*(\mathbf{x}) \) indicates the probability of \( c \) taking the value \( k \).

For mathematical convenience, we code the class data \( c_j \) by \( j \)-th unit column vector \( \mathbf{e}_j = (0, \ldots, 0, 1)^t \) of size \( K \) and define an \( K \times N \) constant matrix \( \mathbf{Y} \) by

\[
\mathbf{Y} = [\mathbf{y}_1; \ldots; \mathbf{y}_N] = [\mathbf{e}_{c_1}; \ldots; \mathbf{e}_{c_N}]
\]

(1)

whose \( j \)-th column vector \( \mathbf{y}_j = \mathbf{e}_{c_j} \) indicates the class to which the data \( \mathbf{x}_j \) is attached.
While SVM determines a single valued dichotomous discriminative function
\[ f(x) = v^T k(x) \] (2)
where \( v \) is a row vector of size \( N \), we introduce a multi-valued polychotomous function
\[ F(x) = V^T k(x) \] (3)
mapping \( R^N \) into \( R^K \), where \( V \) is an \( K \times N \) parameter matrix which is to be estimated by using the training data set \( \{(x_j, c_j)\}_{j=1,...,N} \). \( k(x) \) is a map from \( R^N \) into \( R^N \) defined by
\[ k(x) = (k(x_1, x), ..., k(x_N, x))^T \] (4)
and \( K(x, x') \) is a certain positive definite kernel function. Then we define a model for multinomial probabilistic predictor \( p(x) \) by
\[ p(x) = \hat{p}(F(x)) = (\hat{p}_1(F(x)), ..., \hat{p}_K(F(x)))^T \] (5)
where \( \hat{p}_k(F(x)) = \frac{\exp(F_k(x))}{\sum_{j=1}^K \exp(F_j(x))} \) is the logistic transform.

Under this model assumption, the negative-log-likelihood function \( L(V) \) for \( p(x) \) is given by
\[ L(V) = -\sum_{j=1}^N \log(p_{c_j}(x_j)) = -\sum_{j=1}^N \log(\hat{p}_{c_j}(V^T k(x_j))) \] (6)
which is a convex function (see [9,10]). This objective function \( L(V) \) is of discriminative nature, and that if the kernel function is appropriately chosen, the map \( F(x) \) can represent a wide variety of functions so that the resulting predictive probability \( p(x) \) can be expected to be close to the reality. A predictive vector \( \hat{p}(x) \) can be obtained by putting \( \hat{p}(x) = \hat{p}(V^T k(x)) \) where \( V^* \) is the ML estimate which minimize the function \( L(V) \) with respect to \( V \).

However, over-learning problems could occur with \( V^* \) with the limited number of training data. In order to deal with the problems, the penalty term is introduced and the negative-log-penalized-likelihood
\[ PL(V) = L(V) + \frac{\delta}{2} \| D^{1/2} V K^2 D^{-1/2} \|_F \] (7)
is minimized to estimate \( V \) where \( \| \cdot \|_F \) is the Frobenius norm. The penalty term is intended to reduce the effective freedom of the variable \( V \). The matrix \( D \) is an \( K \times K \) positive definite matrix. A frequent choice of \( D \) is given by
\[ D = \frac{1}{N} Y Y^T \] (8)
which equilibrates a possible imbalance of classes in the training data. The matrix \( K \) is the \( N \times N \) constant matrix, given by
\[ K = [K(x_j, x_j)]_{j=1,...,N} \] (9)
The \( \delta \) is a regularization parameter and can be determined by the empirical Bayes method.

Due to the introduction of the specific quadratic penalty in (7), the minimizer \( V^* \) of \( PL(V) \) is a solution of the neat matrix equation,
\[ \nabla PL = (P(V) - Y + \delta V)K = O_{K,N}. \] (10)
where \( P(V) \) is an \( N \times N \) matrix whose \( j \)-th column vector is the probability vector \( p(x_j) = \hat{p}(V^T k(x_j)) \). The matrix \( Y \) is given in (1).

The minimizer \( V^* \), which gives the probabilistic predictor \( \hat{p}(x) = \hat{p}(V^T k(x)) \), is iteratively computed by the following algorithm.

Algorithm: Starting with an arbitrary \( K \times N \) matrix \( V^0 \), we generate a sequence \( \{V^t\} \) of matrices by
\[ V^{t+1} = V^t - \alpha_t \Delta V^t, \quad i = 0, ..., \infty \] (11)
where \( \Delta V^t \) is the solution of the linear matrix equation,
\[ \sum_{j=1}^N ([p(x_j) - p(x_j) p(x_j)^T]) \Delta V^t (k(x_j) k(x_j)^T) \]
\[ + \delta \Delta V^t K = (P(V^t) - Y + \delta V^t) K \] (12)
The detailed algorithm for estimation is shown in [9-12]. Note that we only need to solve an unconstrained optimization of a strictly convex function \( PL(V) \) or equivalently, to solve the simple matrix nonlinear equation (10).

3. dPLRM-based Speaker identification
The training data set \( \{(x_j, c_j)\}_{j=1,...,N} \) which covers all the speakers’ data is collected. The class data \( \{c_j\} \) is converted into matrix \( Y \). The key matrix \( V^* \) is estimated by dPLRM. Finally the predictor \( \hat{p}(x) \) is obtained.

For testing, the predictive probability \( \hat{p}(x') \) is calculated for each data \( x' \). Then we sum up the log-probability for each class over samples and choose the class which attains its maximum as the speaker who utters the testing data.

4. Experiments and results
The performance of our method is compared with that of the SVM-based method in text-independent speaker identification experiments.

4.1. Data and system description
The data for training and testing have been collected for 10 male speakers. Each speaker utters several sentences and words. Duration of utterance for each sentence is approximately four seconds, and for each word one second. Although the texts are common for all speakers, the sentences used for testing are different from those for training. The utterances of the same set of sentences and words was recorded for testing in three sessions (T1 to T3) over six
months and sampled at 16 kHz. A feature vector of 26 components, consisting of 12 mel-frequency cepstral coefficients plus normalized log energy and their first derivatives, is derived once every 10 ms over a 25.6 ms Hamming-windowed speech segment.

For training, the utterances of three sentences recorded three months earlier than Session T1 (12 second speech in total) are used for estimating $V^*$. The following polynomial function is used as the kernel function.

$$K(x, x') = (x^T x' + 1)^d$$  \hspace{1cm} (13)

The parameters $\alpha$ and $\delta$ in $d$PLRM are experimentally set to be 1.0 in (11) and 7.7e-5 in (7), respectively. In order to execute effective computation with 32-bit precision, the data is affinely transformed so that all the elements of feature vectors lie in the interval [-0.5, 0.5]. The estimation step (11) is terminated at the 10-th iteration.

Then, the utterances of the five sentences and the five words in Sessions T1, T2 and T3 are individually tested. The case number is 150 for each sentence and word data.

In the SVM-based method, the SVM$^{[60]}$ software is used [14] with the polynomial kernel function (13). For each speaker, a one-versus-rest classifier is trained and the speaker who attains the largest positive confidence index averaged over the test speech is regarded as the speaker who uttered the speech for testing.

### 4.2. Results

Tables 2 lists the numbers of sentences and words identified correctly for each session and the accuracy rates with the confidence intervals at a confidence level of 90% averaged over Sessions T1, T2 and T3. For both $d$PLRM- and SVM-based methods, $s = 5, 7, 9$ in (13) were tried. The confidence intervals are calculated based on the assumption that the accuracy rates follow the binomial distribution. Our method performed better than the SVM-based method for both sentence and word speech. $d$PLRMs can capture speaker characteristics with a small amount of data and effectively discriminate each speaker.

### 5. Discussion

$d$PLRMs give a frame-wise probability estimate * and SVMs a frame-wise confidence index. Fig. 1 shows the waveform of the word “MOICHIDO” spoken by the identified speaker, the frame-wise $d$PLRM probability estimates and SVM confidence indices for 10 candidate speakers. The graph in the bottom shows the frame-wise entropy.

$$H_i = - \sum_{k=1}^{K} p_k^*(x_i) \log p_k^*(x_i)$$  \hspace{1cm} (14)

calculated from $d$PLRM probability estimates. It can be used as a measure of discriminatory power of the frame-data.

For the identified speaker, the segments from 24 through 50 and from 91 through 109 frame numbers have higher probability estimates and lower entropy values for $d$PLRM, and can be regarded as key parts which substantially contain the discriminative information. SVM also shows the same tendency with higher positive confidence indices for the segments. With $d$PLRMs, the probability estimates are small among all speakers and the entropy values are large for the beginning frames and ending frames which does not contain speech information. Thus, we can use $d$PLRMs as a probabilistic identifier of speech segments containing the discriminative information.

### 6. Conclusions

This paper investigated a speaker identification method based on $d$PLRMs. We conducted the experiments with training speech in 12-second duration, and showed that the $d$PLRM-based method was competitive with the conventional SVM-based methods. It was also shown that $d$PLRMs could be used as a probabilistic identifier of speech segments having the discriminative information of speakers.

The evaluation of $d$PLRMs with a larger dataset is treated in a forthcoming paper. The probability estimates of $d$PLRMs could be used in speaker verification and end-point detection with modification of the probabilistic model (5) for each objective. Those are also left for our future study.

### 7. References


http://htk.eng.cam.ac.uk, the hidden Markov model toolkit (HTK).

http://www.cs.cornell.edu/People/tj/svm_light/, the support vector machine software, SVMlight.