Beyond Density Functions: Direct Robust and Discriminative Acoustic Classification with Kernel Functions

Andreas Wendemuth

Institut f. Elektronik, Signalverarbeitung und Kommunikationstechnik, Otto-von-Guericke Universität, 39106 Magdeburg, Deutschland, Email: wendemu@iesk.et.uni-magdeburg.de

Abstract

In classical, HMM-based speech recognition, the probability of the production of a segment of acoustic data given a hidden Markov state of a linguistic unit (e.g. word, phoneme) is modelled via probability density functions (pdf), mostly mixtures of multinomial Gaussians (GMM). Replacing GMM production by direct classification of the states opens a new approach which can be made robust, discriminative and generalizing. This is achieved with so called kernel methods, namely with Support Vector Machines and Kernel Fisher Methods, embedded in an HMM framework. Kernel Methods require few training data and can be adjusted to the degree of data mismatch. Results are presented on established data sets which show the potential of the method.

Speech Recognition using Hidden Markov Models

The general task of Automatic Speech Recognition (ASR) is to deduce an unknown sequence of words (text) from its observed acoustical realization, an utterance. We must thus “reverse” the process of speech production.

Predominantly, Hidden Markov Models (HMMs) are used in ASR. A HMM is a stochastic finite state automaton (SFSA) built from a finite set of possible states $Q = \{q_1, \ldots, q_K\}$. Each of these states is associated with a specific probability distribution. A specific HMM M_i is, then, represented by a SFSA comprised of L_i states $S_i = \{s_1, \ldots, s_l, \ldots, s_{L_i}\}$ with each $s_l \in Q$, arranged according to a certain, most often predefined, topology.

Thus, HMMs can be used to model a sequence of feature-vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$ as a piecewise stationary process where each stationary segment is associated with a specific hidden (not directly observable) linguistic HMM state, typically word labels or phoneme states. This approach models the temporal structure of speech as well as the locally stationary generation of the speech signal from the internal states. Theory and methodology of HMMs are described in a number of sources, e.g. [5]. The fundamental equation describing this process is Bayes' rule, applied to speech recognition:

$$P(M|\mathbf{X}, \mathbf{\Theta}) = \frac{p(\mathbf{X}|M, \mathbf{\Theta})P(M|\mathbf{\Theta})}{p(\mathbf{X}|\mathbf{\Theta})} \quad (1)$$

in which $\mathbf{\Theta}$ is the parameter set and $P(M|\mathbf{X}, \mathbf{\Theta})$ is the posterior probability of the hypothesized HMM M given a seqence $\mathbf{X}$ of feature-vectors. Since this probability cannot be computed directly, it is usually split according to (1) into the acoustic model $p(\mathbf{X}|M, \mathbf{\Theta})$ and a prior $P(M|\mathbf{\Theta})$ representing the language model. The (full) acoustic likelihood is computed by expanding it into all possible state paths in M that can generate $\mathbf{X}$, usually approximated by the best possible path in the ‘Viterbi”-approximation. When decoding an observation $\mathbf{X}$, we have to find the model M_j which maximizes $P(M|\mathbf{X}, \mathbf{\Theta})$:

$$j \quad = \quad \underset{\forall i}{\mathrm{argmax}}\, p(\mathbf{X}|M, \mathbf{\Theta})P(M|\mathbf{\Theta})\ . \quad (2)$$

The acoustic model $p(\mathbf{X}|M, \mathbf{\Theta})$ is usually realized using GMMs. However, these models suffer from limited discrimination and generalization ability, calling for alternative descriptions with better properties. Artificial Neural Networks [2] have been used in the past. GMMs model the production probability of an acoustic observance to be created from internal state q_i. Bayes' rule then gives the probability of classification into a particular acoustic unit. In a more straightforward way, one can aim at designing appropriate classifiers which model $P(q_i|x, \Theta)$ directly for each q_i. One can probabilistically interpret $P(q_i|x, \Theta)$, e.g. use Bayes' rule (1) in reverse direction, to arrive again at production probabilities, as required if one wants to remain within the HMM terminology. Using such classifiers has the advantage that a number of desired effects can be modelled directly: the classifiers can be trained to achieve optimal generalization on given (even small) data sets, and they can be made discriminative in a true sense, i.e. being fine-tuned at sensitive class boundaries whereas not focussing on regions where class assignments are uncritical.

Kernel Methods

In order to reach the desired effects described in the last section, a number of so-called kernel-methods have been developed and extensively used in other fields of pattern recognition over the last decade. We give account here of the use of two prominent types of methods: Kernel Fisher Discriminants (KFD) and Support-Vector-Machines (SVM).

The Fisher Discriminant (FD) is an approach for two-class discrimination problems. Consider a training set $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$ belonging to an input space $\mathcal{X}$ and consisting of M samples which are split into two classes. Let the classes be labelled with ±1 defining a corresponding label vector $\mathbf{y} \in \{-1, 1\}^M$. The number of samples labelled with ±1 is $M_{\pm}$, the class means are $\mathbf{m}_{\pm}$. Successful classification of the samples can be achieved by aiming at $y_p = \mathbf{w}\mathbf{x_p}$. Projection on $\mathbf{w}$ must separate the class means and at the same time minimise the variances within the classes. Thus, one has to maximise

$$R(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \tag{3}$$

with $\mathbf{S}_B$ and $\mathbf{S}_W$ denoting the unnormalized between-class and within-class covariance (scatter) matrices By differentiating (3) one can see that $\mathbf{w}$ is the leading eigenvector of the generalised eigenvalue problem $\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w}$. However, the application of the FD for practical problems suffers from the restriction to linear discrimination in the input space. Sufficiently rich discrimination directions are aimed at by using a nonlinear mapping Φ applied to the data $\mathbf{X}$. To avoid an explicit mapping the so called *kernel trick* is used [3], where the space $\mathcal{F}$ is induced by certain inner products (kernels)

$$k(\mathbf{x}, \mathbf{x}') = (\Phi(\mathbf{x})\Phi(\mathbf{x}')), \tag{4}$$

among k being Gaussian (or radial basis), polynomial and sigmoid kernels. The resulting KFD does not require knowledge of Φ. Both FD and KFD are equivalent to a least square regression to the labels [3]. Thus instead of solving the generalised eigenvalue problem imposed by the KFD one can obtain an equivalent direction $\mathbf{w}$ by solving

$$\min_{\mathbf{w},b} E(\mathbf{w}) = \sum_{i=1}^{M} \|\Phi^T(\mathbf{x}_i)\mathbf{w} + b - y_i\|^2 \tag{5}$$

with b denoting a bias term. By defining $\Phi \rightarrow (\Phi^T, 1)^T$ and $\mathbf{w} \rightarrow (\mathbf{w}^T, b)^T$ and by exploiting the fact that $\mathbf{w}$ can be expressed as an expansion $\mathbf{w} = \sum_{i=1}^{M} \alpha_i \mathbf{\Phi}(\mathbf{x}_i)$, eq. (5) can be written in dual form as

$$\min_{\boldsymbol{\alpha}} E(\boldsymbol{\alpha}) = \|\mathbf{K}\boldsymbol{\alpha} - \mathbf{y}\|^2 \tag{6}$$

with $[\mathbf{K}]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ denoting the symmetric $M \times M$ kernel matrix and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)^T$. The least square solution $\hat{\boldsymbol{\alpha}}$ of (6) is given by the pseudoinverse

$$\hat{\boldsymbol{\alpha}} = (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\mathbf{y}. \tag{7}$$

Note that the dimensionality of $\mathbf{K}$ is the number of training samples, hence the computation of $\mathbf{K}$ bears practical limitations in terms of memory and computational cost for large datasets.

The KFD is a linear kernel method for arriving at the class labels y_i. SVMs instead aim at computing the labels directly, i.e. $y_i = \text{sign}(\mathbf{w}\mathbf{\Phi}(\mathbf{x}_i))$, with maximum margin of the two classes. Using the same kernel methods as above, one arrives at equations for the *constrained* quadratic optimization problem (8), (9) implemented by the SVM and the resulting decision function (10). Note that the α_i are within bounds C, which is a theoretical requirement of the SVM to produce optimal generalization.

$$\begin{aligned} \vec{\alpha}^0 &= \arg\max_{\vec{\alpha}} W(\vec{\alpha}) && (8) \\ W(\vec{\alpha}) &= \sum_i \alpha_i - \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j K(\vec{x}_i, \vec{x}_j) && (9) \\ & \sum_i \alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C && \\ h(\vec{x}) &= \text{sign}\left[\sum_i \alpha_i^0 y_i K(\vec{x}_i, \vec{x})\right] && (10) \end{aligned}$$

Experiments and Conclusion

Results are given for first experiments the details of which are published elsewhere. Throughout all experiments Gaussian kernels with variance σ and monophone recognizers were used. For comparison, results with GMMs are given. We trained the KFD-classifier on a subset of the WSJ consisting of 20 sentences. The test was performed on a second disjoint subset with 10 sentences [1]. The SVM was used for rescoring N-best lists and lattices of TIMIT recognition results [4]. Results on evaluation test sets are shown in table 1.

WER	GMM	KFD $\sigma = 16$	KFD $\sigma = 17$	SVM N-best	SVM Lat.
WSJ	58.7%	51.8%	42.0%		
TIMIT	43.3%			42.1%	39.9%

Table 1: Results on the test set for KFD and SVM.

We have shown that it is possible to use the probabilistically interpreted outputs of the KFD and SVM as an acoustic model for a HMM-based speech recognizer. They serve to alleviate a number fo drawbacks of GMMs, in particular they are useful for sparse and noisy data.

References

[1] Edin Andelic et al.: Iterative Implementation of the Kernel Fisher Discriminant for Speech Recognition, Proc. SPECOM, St. Petersburg 2004, pp. 99–103.

[2] Hervé Bourlard and Nelson Morgan: Hybrid HMM/ANN Systems for Speech Recognition: Overview and New Research Directions, in: C. L. Giles and M. Gori: Adaptive Processing of Sequences and Data Structures, Springer, Heidelberg (1998), 389–417

[3] Sebastian Mika: Kernel Fisher Discriminants, Ph.D Thesis, Technische Universität Berlin, 2002

[4] André Stuhlsatz et al.: Classification of speech recognition hypotheses with Support Vector Machines, In: A. Wendemuth (ed.): Proc. Speech Processing Workshop Speech-DAGM, Magdeburg 2003, pp. 65-72. Published by University of Magdeburg. ISBN 3-929757-59-1.

[5] Andreas Wendemuth: Grundlagen der stochastischen Sprachverarbeitung, Oldenbourg, München 2004. ISBN: 3-486-57610-0