Infinite Support Vector Machines in Speech Recognition

Jingzhou Yang, Rogier C. van Dalen and Mark Gales

Department of Engineering, University of Cambridge
Trumpington Street, Cambridge, CB2 1PZ, UK
{ jy308, rcv25, mjfg }@eng.cam.ac.uk

Abstract

Generative feature spaces provide an elegant way to apply discriminative models in speech recognition, and system performance has been improved by adapting this framework. However, the classes in the feature space may be not linearly separable. Applying a linear classifier then limits performance. Instead of a single classifier, this paper applies a mixture of experts. This model trains different classifiers as experts focusing on different regions of the feature space. However, the number of experts is not known in advance. This problem can be bypassed by employing a Bayesian non-parametric model. In this paper, a specific mixture of experts based on the Dirichlet process, namely the infinite support vector machine, is studied. Experiments conducted on the noise-corrupted continuous digit task AURORA 2 show the advantages of this Bayesian non-parametric approach.

Index Terms: generative feature space, Bayesian non-parametric, Dirichlet process, mixture of experts, infinite support vector machines

1. Introduction

In previous work [1, 2, 3, 4], a variety of discriminative models based on generative feature spaces were studied and experiments showed improvement compared with well-trained hidden Markov models (HMMs) [5]. The main advantage of using generative feature spaces, which are built on generative models, is that model-based adaptation and compensation can be implemented on the generative models in the process of evaluating the feature spaces. This can enhance the noise robustness of the recognition system. This paper focuses on a certain type of Bayesian non-parametric model [6] named the infinite support vector machine (iSVM) [7] in speech recognition based on generative feature spaces derived from vector Taylor series (VTS) [8] compensated HMMs.

Speech recognition can be considered as a problem of classifying sequential audio data (e.g. vectors of MFCCs). Normally, the sequential data vary in length, but classifiers such as SVMs can only handle data with fixed dimension. In order to bridge this gap, a generative model (e.g. an HMM) can be applied on the sequential data to derive features with fixed dimension [9]. In previous work [2, 3], one linear SVM classifier was adopted on the feature space may be not linearly separable. Rather than applying a single linear classifier on the feature space, it is more reasonable to utilise a mixture of experts [10] which makes an ensemble decision by all experts with various weights determined by the region of space. Figure 1 illustrates an single classifier and a mixture of experts on the feature space.

Although the kernel trick could be applied on the SVM to yield a non-linear decision boundary, it might be problematical to choose the type of kernel, and the number of support vectors might be large which leads to inefficiency in testing. Thus, the kernel trick is not considered in this paper, even if the kernel trick could be applied on the mixture of experts as well.

In terms of the mixture of experts, it is hard to set the number of experts to fit with the training data under a parametric framework. In contrast, the model complexity in a Bayesian non-parametric model is treated as one set of model parameters, and the posterior distribution of the complexity can be inferred. Then, the model complexity can be integrated out when making predictions, namely the model is averaged over all possible complexities. Thus, a Bayesian non-parametric model becomes a better choice in tackling the problem of selecting model complexity. In [11], a Bayesian non-parametric model named sticky hierarchical Dirichlet process hidden Markov model (HDP-HMM) [12] was introduced to infer the unknown number of speakers in a speech diarisation task, and state-of-art performance was achieved. In this paper, one example of Dirichlet process (DP) mixture of experts named iSVM is implemented to resolve the problem of the unknown number of experts.

This paper is organised as follows. Section 2 discuss various features derived from generative models. The mixture-of-experts model is discussed in section 3, and its non-parametric counterpart called DP mixture of experts is introduced in section 4. Section 5 details a specific example of the DP mixture of experts model, the iSVM. Finally, the experimental results and conclusions are presented in section 6 and 7.

2. Features

In speech recognition, the speech utterances normally vary in length. In order to handle the length variation of the observations, generative models can be utilised to map the sequential data (on the input space) with various length to feature vectors.

Figure 1: The single classifier and mixture of experts on the feature space.
\( x_i = \varphi(O_i, \lambda) \)

\( x_1 = \text{five}, x_2 = \text{eight}, x_3 = \text{seven}, x_4 = \text{six} \)

Figure 2: The process of generating the feature vectors.

(on the feature space) with fixed dimension. Assume the speech observations are \( O = \{O_1, \ldots, O_I\} \), where \( \{O_1, \ldots, O_I\} \) are one possible segmented data of the observations \( O \) (e.g. the 1-best hypothesis of the lattice), and each segment \( O_i \) specifies a word/phone/sub-phone. One possible utterance is illustrated in Figure 2. Given a segment \( O_i \), the log-likelihood feature space can be described as follows [4]:

\[
\varphi' (O; \lambda) = \frac{1}{T_i} \begin{bmatrix} \log (p(O_i | \lambda_{\tilde{w}_1})) \\ \vdots \\ \log (p(O_i | \lambda_{\tilde{w}_n})) \end{bmatrix}
\]

(1)

where \( \{\tilde{w}_1, \ldots, \tilde{w}_n\} \) are all the unique classes, or vocabularies, \( T_i \) is the number of frames in the observation \( O_i \), and \( p(O_i | \lambda_{\tilde{w}_i}) \) is the likelihood of the generative model parameters corresponding to class \( \tilde{w}_i \) given \( O_i \). Figure 2 illustrates the process of deriving features from an utterance. For simplicity, \( \varphi (O; \lambda) \) is written as \( x \).

A more general form of generative feature space is the derivative of the feature space, which not only includes the log-likelihood of the parameters but also incorporates the derivatives of the log-likelihood with respect to the parameters of the generative models. If only the first order of the derivatives is considered, the feature space can be described as follows:

\[
\varphi' (O; \lambda) = \frac{1}{T_i} \begin{bmatrix} \nabla_{\lambda_{\tilde{w}_1}} \log (p(O_i | \lambda_{\tilde{w}_1})) \\ \vdots \\ \nabla_{\lambda_{\tilde{w}_n}} \log (p(O_i | \lambda_{\tilde{w}_n})) \end{bmatrix}
\]

(2)

According to previous work [3, 4], the derivatives with respect to the mean have better discrimination than the derivatives with respect to other parameters. The derivative with respect to the mean of the component \( m \) can be described as follows:

\[
\nabla_{\mu_m} \log (p(O_i | \lambda_{\tilde{w}_i})) = S^{-1}(\alpha_i - \mu_m)
\]

(3)

where \( \gamma_m(t) \) is the posterior probability of component \( m \) generating \( \tilde{w}_i \).

In order to adapt the generated features to the target noise condition, state-of-art model-based adaptation and compensation technology can be implemented on generative models in the process of deriving the features. VTS compensation [8] is adopted in our work. The parameters of convolutional noise and additive noise are estimated on each utterance to maximise the likelihood of the HMM system.

3. Mixture of experts

Rather than making prediction on the whole feature space from a single model, another method is choosing different models to make prediction according to different features, and the choice of the model is input-dependent. When the choice of the model is a hard decision, this framework is known as a decision tree. When the choice of each model is given a probability depending on the input, this is known as a mixture of experts [13]:

\[
P(w|x, \theta_a, \eta_a) = \sum_{m=1}^{M} P(z = m|x, \theta_a)P(w|x, \eta_m)
\]

(4)

where \( x \) is the variable in the feature space \( \varphi(O, \lambda) \) given in section 2, \( w \) is the target variable, \( M \) is the number of experts. \( P(z = m|x, \theta_a) \) is the gating network which assigns probabilities to different experts according to the input \( x \), and \( P(w|x, \eta_m) \) is the \( m^{th} \) expert, which is a discriminative model. \( \theta_a = \{\theta_1, \ldots, \theta_M\} \) are the parameters of the gating network, and \( \eta_m = \{\eta_1, \ldots, \eta_M\} \) are the parameters of the \( M \) experts. If the gating network is given from the component posteriors of the mixture model and the number of experts is given \( M \), the graphical model of the mixture of experts is illustrated in Figure 3, and the corresponding generative process of this model can be described as follows:

\[
\pi \sim \text{Dirichlet}(\alpha) \quad z_i \sim \text{Categorical}(\pi)
\]

(5)

\[
x_i \sim p(x|\theta_{z_i}) \quad w_i \sim P(w|x_i, \eta_{z_i})
\]

(6)

where \( \pi = \{\pi_1, \ldots, \pi_M\} \) is a discrete distribution which is drawn from a symmetric Dirichlet distribution with concentration parameter \( \alpha \), and \( z_i \) is the indicator variable that denotes the \( i^{th} \) datum is associated with which expert. Categorical(\( \pi \)) is the categorical distribution which is the generalisation of the Bernoulli distribution with \( M \) possible outcomes. \( \theta_{z_i} \) are the parameters of the component indicated by \( z_i \), and \( \eta_{z_i} \) are the parameters of the expert indicated by \( z_i \).

4. Dirichlet process mixture of experts

As a parametric model, the number of experts \( M \) needs to be fixed in advance for the mixture of experts. In order to bypass this problem of choosing model complexity, a Bayesian non-parametric version of the mixture-of-experts model, namely DP mixture of experts, is used here. In [14], the non-parametric model called infinite Gaussian mixture model (iGMM) is derived by setting the number of components to infinity in the Gaussian mixture model (GMM). Similarly, in this paper, the corresponding non-parametric counterpart of the mixture-of-experts model is derived, when the number of experts goes to infinity \( M \to \infty \) in the mixture-of-experts model.

Since \( \pi \) is a discrete probability and \( z_i \sim \text{Categorical}(\pi) \), then \( P(z_1, \ldots, z_i) \) has a multinomial distribution. By marginalising out \( \pi \), the following result can be derived [14]:

\[
P(z_1, \ldots, z_i|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(i + \alpha)} \prod_{m=1}^{M} \frac{\Gamma(N_m + \alpha/M)}{\Gamma(\alpha/M)}
\]

(7)
where \( N_m \) is the number of data points associated with the \( m^{th} \) expert. The conditional probability of the indicator \( z_{i+1} \) given the other \( i \) indicators can be described as follows:

\[
P(z_{i+1} = m | z_1, \ldots, z_i, \alpha) = \frac{P(z_1, \ldots, z_{i+1} | \alpha)}{P(z_1, \ldots, z_i | \alpha)} = \frac{N_m + \alpha / M}{i + \alpha}
\]

When the number of experts goes to infinity \( M \to \infty \), the experts’ weights \( \alpha \) are given by the Dirichlet process (DP) [15], and the conditional probability of the indicator variable \( z_{i+1} \) given all the previous \( i \) indicators can be described as follows:

\[
P(z_{i+1} = m | z_1, \ldots, z_i, \alpha) = \begin{cases} \frac{N_m}{i + \alpha}, & \text{where } m \text{ is an existing expert} \\ \frac{\alpha}{i + \alpha}, & \text{where } m \text{ is a new expert} \end{cases}
\]

According to equation (9), the probability of the indicator indicating an existing component is proportional to the number of data associated with that component, and the probability of assigning to a new component is proportional to \( \alpha \). The process of assigning the data to the components according to equation (9) is also known as the Chinese Restaurant Process (CRP) [16]. The CRP provides a mechanism to draw from the distribution given by a DP without specifying that distribution.

When \( M \to \infty \), the DP mixture-of-experts model could be derived, then the corresponding generative process of this model can be described as follows:

\[
z \sim \text{CRP}(\alpha) \quad \theta_m \sim G_1, \eta_m \sim G_2, \forall m \in z \quad (10)
\]

\[
x_i \sim p(x | \theta_{z_i}) \quad w_i \sim P(w | x_i, \eta_{z_i}) \quad (11)
\]

where \( z = \{z_1, \ldots, z_N\} \) are all the indicators, which are drawn by the Chinese restaurant process CRP(\( \alpha \)) with parameter \( \alpha \). The parameters of the gating network \( \theta_m \) and the parameters of the experts \( \eta_m \) are given by the base distributions \( G_1 \) and \( G_2 \) respectively. The corresponding graphical model is illustrated in Figure 4.

This non-parametric model used for classification can be approximated by the samples from the posterior distribution of the model parameters:

\[
P(w | x, \mathcal{D}) = \int P(w | x, \Theta) P(\Theta | \mathcal{D}) d\Theta 
\approx \frac{1}{K} \sum_{k=1}^{K} P(w | x, \mathcal{G}^{(k)})
\]

\[
= \frac{1}{K} \sum_{k=1}^{K} \sum_{m=1}^{M_k} P(z = m | x, \Theta^{(k)}) P(w | x, \eta^{(k)}_m, z = m)
\]

\[
(12)
\]

where \( \Theta = \{\theta_m, \eta_m\} \) are all the parameters of the DP mixture of experts. For this non-parametric model, the posterior distribution of the parameters \( p(\Theta | \mathcal{D}) \) is extremely complicated. This makes the integral in equation (12) intractable. Thus, a Markov chain Monte Carlo (MCMC) [17] method is used to approximate this integral. \( \{\Theta^{(1)}, \ldots, \Theta^{(K)}\} \) are sampled from the posterior distribution of the parameters \( p(\Theta | \mathcal{D}) \). Since this distribution is complicated, it is impractical to draw samples from this full joint distribution directly. Therefore, the Gibbs sampling [18] is implemented here to draw samples from the conditional posterior distribution of each parameter given all others \( p(\theta_m | \Theta_{\neq m}, \mathcal{D}) \), rather than sample from the posterior distribution of the whole parameter set \( p(\Theta | \mathcal{D}) \), where \( \Theta_{\neq m} \) denotes all the parameters \( \Theta \) except \( \theta_m \).

5. Infinite support vector machine

The infinite SVM (iSVM) was introduced in [7], where the iSVM is based on the stick-breaking representation [21] of the DP. In this paper, the iSVM based on the DP from a CRP perspective is introduced. In the previous section, the DP mixture of experts model is discussed. When each expert is defined as a multi-class SVM [22] in the DP mixture-of-experts model and the gating network is given from the component posterior of the DP mixture model, the iSVM can be derived. Since the mixture of experts is a probabilistic model, in order to derive the iSVM, each expert needs to be interpreted in a probabilistic way.

According to [2], the multi-class SVM can be interpreted probabilistically as a log-linear model with large-margin training criterion. The log-linear model can be described as follows:

\[
P(w | x, \eta_m) = \frac{\exp \left( \eta_m^T \Phi(x, w) \right)}{\sum_{w} \exp \left( \eta_m^T \Phi(x, w) \right)}
\]

\[
(14)
\]

where \( \Phi(x, w) = \delta(w) \otimes x = [0^T, \ldots, x^T, \ldots, 0]^T \) is the joint feature space, \( \delta(w) = [\delta(w - \tilde{w}_1), \ldots, \delta(w - \tilde{w}_T)]^T \) denotes the position of the label \( w \) in the vocabulary, and \( \otimes \) is the tensor product. The large-margin training criterion of the log-linear model is defined as follows [2]:

\[
- \log p(\eta_m) + \sum_{i=1}^{N_m} \max_{w \neq w_i} \left\{ \mathcal{L}(w, w_i) - \log \left( \frac{P(w_i | x_i, \eta_m)}{P(w | x_i, \eta_m)} \right) \right\}
\]

\[
(15)
\]

where \( \mathcal{L}(w, w_i) \) is the loss function which measures the distance between the reference \( w_i \) and label \( w \), and \( [\cdot]_+ \) is the hinge-loss function. Assume the prior of \( \eta_m \) is given a Gaussian distribution \( p(\eta_m) = N(\mu_\eta, \Sigma_\eta) \) with mean \( \mu_\eta \).
and scaled identity covariance matrix $\Sigma_\eta = CI$. By substituting this Gaussian prior and equation (14) into equation (15), the large margin criterion can be described as minimizing the follows:

$$\frac{1}{2C} ||\eta_m - \mu_\eta||^2 + \sum_{n=1}^{N_m} \max_{w \neq w_1} \{\eta_m^T \Phi(x_n, w) + \mathcal{L}(w, w_1)\}$$

$$- \eta_m^T \Phi(x_n, w_1)$$

(16)

In equation (16), when the mean $\mu_\eta = 0$, this becomes the training criterion of the multi-class SVM. In the iSVM, if there are very few data associated with an expert ($N_m$ is small), the trained expert might lack generalisation. Thus, each expert deploys a non-zero mean $\mu_\eta$ which is obtained from the multi-class SVM trained on the whole training set. By introducing the non-zero mean, the iSVM should retrieve multi-class SVM performance, if $C$ is small enough. Better performance could be achieved by gradually increasing $C$.

Equation (16) is also known as the training criterion of the structural SVM [23, 24]. The main difference between the structural SVM and the multi-class SVM is the form of the joint feature space $\Phi(x_n, w_i)$ which is defined by the observation-label pair. In the structural SVM, the observation-label pair is composed of the whole sentence $\{x_1, \ldots, x_n\}$ and the corresponding labels $\{w_1, \ldots, w_n\}$, then the corresponding joint feature space is $\Phi(\{x_1, \ldots, x_n\}, \{w_1, \ldots, w_n\})$. In contrast, the segments are treated independently in the multi-class SVM, and the pair consists of one segment $x_i$ and the corresponding label $w_i$. The definition of the joint feature space is detailed in [24, 25]. Moreover, in the multi-class SVM, the loss function $\mathcal{L}(w, w_1)$ is defined as the $Kronecker delta$ $\delta(w, w_1)$. In the structural SVM, the loss is set to a more refined function, say the Levenshtein distance, rather than the zero loss.

The classification criterion of the iSVM is given in equation (12), and the number of samples $K$ is 10 here. The experimental results are listed in Table 1. All the discriminative models outperform the VTS compensated HMM baseline system. On the log-likelihood feature space and derivative feature space, the iSVM achieves better performance than the multi-class SVM. This gain is obtained by the fact that the iSVM explores the distribution of the training data and infers the number of experts, then applies different experts focus on different regions of the feature space to make an ensemble decision, rather than applying a single classifier on the whole feature space.

### 6. Experiments

In this paper, the experiments are conducted on the Aurora 2 database [26], which was designed to evaluate the performance of speech recognition algorithm in various noisy conditions. The utterances in this database are the continuous digit strings with vocabulary size 12 (one to nine, plus zero, oh and silence), and 8 real-world noise conditions have been added to the speech artificially over a variant of signal to noise ratio (SNR). The generative models (HMMs) are trained on the clean data with 8840 utterances recorded from 55 male and 55 female adults. The feature vectors used by the front-end HMMs consisted of 12 MFCCs appended with zeroth cepstrum, delta and delta-delta coefficients. The noise model for VTS compensation is estimated on each utterance. The performance of the VTS compensated HMM is listed in Table 1. The SVM and iSVM are trained on a subset of the multi-style training data containing 4 noise conditions (N2, N3 and N4) and 3 SNRs (20dB, 15dB and 10dB). All three test database, A, B and C, are used in the evaluation. In set A and B, each contains 4 noise conditions at 5 different SNRs (0dB to 20dB). There are 2 noise conditions at 5 different SNRs in set C.

In the experiments, log-likelihood feature vectors and derivative feature vectors are used, and these feature vectors are derived from the VTS compensated HMMs. To keep training with derivative feature vectors feasible, only the first element of the derivative with respect to each mean is used in this paper. All the experts (SVMs) of the iSVM share the same $C$, and the parameter $C$ is tuned on the test set A. Figure 5 illustrates the WER of the iSVM on different feature spaces with various $C$. Since the parameter of each expert is given a Gaussian prior with mean $\mu_\eta$, which is obtained from the multi-class SVM, the iSVM only achieves the baseline performance of the multi-class SVM when the $C$ is small. By introducing the non-zero mean $\mu_\eta$, the iSVM can at least achieve the performance of the multi-class SVM, and the optimised configuration can be obtained by gradually increasing $C$. Without the mean $\mu_\eta$, the iSVM could have poor performance, because not all the experts associate with enough data.

The classification criterion of the iSVM is given in equation (12), and the number of samples $K$ is 10 here. The experimental results are listed in Table 1. All the discriminative models outperform the VTS compensated HMM baseline system. On the log-likelihood feature space and derivative feature space, the iSVM achieves better performance than the multi-class SVM. This gain is obtained by the fact that the iSVM explores the distribution of the training data and infers the number of experts, then applies different experts focus on different regions of the feature space to make an ensemble decision, rather than applying a single classifier on the whole feature space.

### 7. Conclusions

In this paper, a DP mixture-of-experts model based on the Chinese restaurant process has been presented, and a specific example of this type of model, the iSVM, is studied. The iSVM not only infers the number of expert in the mixture-of-experts model, but also inherits the advantages of the mixture of experts that difference experts focus on different regions of the feature space in order to make better predictions. The experiments show the advantages of the iSVM comparing with the multi-class SVM. In this paper, the word segmentations are obtained from the 1 best hypothesis of the word lattice. Future work will study the way to optimise both the segmentations and $\eta$, and generalise the iSVM to large vocabulary ASR by incorporating the structural SVM within the iSVM.
8. References


