INFINITE STRUCTURED SUPPORT VECTOR MACHINES FOR SPEECH RECOGNITION

J. Yang, R. C. van Dalen, S.-X. Zhang and M. J. F. Gales

Department of Engineering, University of Cambridge, Cambridge, UK
{jy308,rcv25,sxz20,mjfg}@eng.cam.ac.uk

ABSTRACT

Discriminative models, like support vector machines (SVMs), have been successfully applied to speech recognition and improved performance. A Bayesian non-parametric version of the SVM, the infinite SVM, improves on the SVM by allowing more flexible decision boundaries. However, like SVMs, the infinite SVMs model each class separately, which restricts them to classifying one word at a time. A generalisation of the SVM is the structured SVM, whose classes can be sequences of words that share parameters. This paper studies a combination of Bayesian non-parametrics and structured models. One specific instance called infinite structured SVM is discussed in detail, which brings the advantages of the infinite SVM to continuous speech recognition.

Index Terms— Bayesian non-parametrics, Dirichlet process, mixture of experts, infinite structured SVM

1. INTRODUCTION

Discriminative models, like support vector machines (SVMs), have been successfully applied to speech recognition. By introducing features in a generative feature space [1] extracted using HMMs, state-of-art speaker adaptation and noise robustness techniques [2] can be used to generate the features.

An SVM by itself uses a linear decision boundary, which may be inappropriate for the feature space. The standard approach is to apply a kernel function. An alternative that does not require choosing a kernel is to use a mixture-of-experts model [3, 4], which employs different classifiers for different regions of space. Normally the optimal number of experts is unknown. In order to sidestep the problem of setting the number of experts, a Bayesian non-parametric framework can be used. By using the infinite SVM (iSVM) [5, 6], better performance is achieved compared with using the SVM, because the iSVM allows SVMs to focus on regions of feature space. Since the interpolation weights for the SVM outputs depends on the location of the data in the feature space, the ensemble decision effectively uses a non-linear decision boundary.

The SVM and iSVM are unstructured models in that they model each class separately. They have therefore been applied to digit recognition in an acoustic code-breaking [7] setting: first continuous speech is segmented into segments, and then each segment from an HMM recognition pass is rescored independently [6]. In continuous speech recognition (CSR), however, each class is a sentence, and the number of possible classes is unlimited. For example, the possible number of classes for a 6-digit length utterance is 10^6. However, these classes have structure: they share words or phones. The structured SVM (SSVM) [8] was introduced to classify data with structured labels. In [9], the SSVM was successfully used in medium to large vocabulary CSR tasks. In order to apply the mixture-of-experts framework to large vocabulary CSR, the structure must be incorporated into the model.

This paper discusses Bayesian non-parametrics for structured SVMs and introduces the infinite structured SVM (iSSVM) in particular. Rather than using a kernel function in the SSVM [10], which requires a kernel to be chosen, the iSSVM deploys multiple SSVMs to yield a non-linear boundary. The generative feature space already implies a sequence kernel [1]. Though an additional kernel could be used by each expert in the iSSVM, the kernel trick is not considered here.

This paper is organised as follows. The mixture of experts and its infinite version are discussed in Section 2. The SSVM is detailed in Section 3, and the iSSVM is introduced in Section 4. Classification and corresponding issues are discussed in Section 5. Finally, the experimental results and conclusions are given in Section 6.

2. MIXTURE OF EXPERTS

The mixture of experts combines the posteriors from multiple experts focusing on different regions of the feature space, producing a model that can perform more complicated classifications than a single expert can. The framework of the mixture of experts with M experts is illustrated in Fig. 1. The gating network, e.g. a GMM with parameters Θ, uses the input to determine the mixture weights \( P(z = m|x, Θ) \) for each expert \( m \). \( z \) is a random variable, the indicator variable, which denotes which expert the input \( x \) is associated with.

The overall class posteriors are computed with

\[
P(w|x, Θ, H) = \sum_{z \in Z} P(z|x, Θ) P(w|x, η_z)
\]  

\[1\]
In equation (4), $p(O_i|\lambda_{\omega_i})$ is the likelihood of the HMM corresponding to label $\omega_i$ given the segment $O_i$.

For large-margin training of the log-linear model, the margin is defined as the log-posterior ratio between the reference $W_n$ and the most competing hypothesis $W$. There are $N$ training instances. By introducing the prior $P(\eta)$, the training criterion to be minimised can be described as:

$$
\sum_{n=1}^{N} \left[ \max_{W_i \neq W_n; \rho_n} \left\{ \mathcal{L}(W, W_n) - \log \left( \frac{P(W_n, \rho_n|O_n, \eta, \lambda)}{P(W_i, \rho|O_n, \eta, \lambda)} \right) \right\} \right] + \log P(\eta)
$$

where $\mathcal{L}(W, W_n)$ is the loss between the hypothesis $W$ and the reference $W_n$. The set of possible $W$ and $\rho$ are obtained from a lattice. When the prior $P(\eta)$ is given a Gaussian distribution $P(\eta) = \mathcal{N}(\eta; \mu_\eta, \Sigma_\eta)$ with mean $\mu_\eta$ and scaled identity covariance matrix $\Sigma_\eta = C I$, and substituting equation (2) into equation (5), the large-margin training criterion can be further described as follows [9]:

$$
\frac{1}{2} \|\eta - \mu_\eta\|^2 + C \sum_{n=1}^{N} \left[ \max_{W_i \neq W_n; \rho_n} \left\{ \eta^T \Phi(O_n, W; \lambda, \rho) \right\} - \mathcal{L}(W, W_n) \right] - \eta^T \Phi(O_n, W_n; \lambda, \rho_n) + \mathcal{L}(W, W_n)
$$

### 4. INFINITE STRUCTURED SVM

The Dirichlet process mixture of experts is given in equation (1) with infinite-sized indicator set $Z$. In order to apply this type of model to continuous speech recognition, structure needs to be incorporated in the model. The direct way to do this is to incorporate the structure into the experts. If each expert is an SSV, described in equation (2), then the DP mixture of experts given in equation (1) becomes:

$$
P(W, \rho|O, \lambda, \Theta, H) = \sum_{z \in Z} P(z|O, \lambda, \Theta) P(W, \rho|O, \lambda, \eta_z)
$$

(7)

Here, the indicator variable $z$ corresponds to the whole utterance $O$, and the utterance indicator $z$ is a scalar. The resulting model consists of infinite number of SSVs, and is called the infinite structured SVM (iSSVM). An alternative is to model each word as a mixture of SSVs. The utterance indicator $z$ then becomes a vector, and the resulting model can be called a structured infinite SVM (SiSVM). However, this paper focuses on the infinite structured SVM, as given in (7) with scalar indicator $z$.

Suppose the training data are $D = \{O_1, \ldots, O_N; W_{1}, \ldots, W_{N}; \rho_1, \ldots, \rho_N\}$, then classification of the iSSVM can be described as follows:

$$
P(W, \rho|O, \lambda, D) = \int P(W, \rho|O, \lambda, A)p(A|D) dA
$$

(8)

where $A = \{\Theta, H\}$ are all the parameters of the iSSVM. Since the integral in equation (8) is intractable to compute, a...
Monte Carlo method can be applied to approximate this integral. Classification can then be described as:

\[
P(W, \rho|O, \lambda, D) \approx \frac{1}{K} \sum_{k=1}^{K} P(W, \rho|O, \lambda, A^{(k)})
\]

\[
= \frac{1}{K} \sum_{k=1}^{K} \sum_{m=1}^{M_k} P(z = m|O, \lambda, \Theta^{(k)}) P(W, \rho|O, \lambda, \eta^{(k)}_{m})
\]

where \(A^{(k)}\) are sampled from the model posterior distribution \(p(A|D)\). Here, \(K\) samples are used to approximate this intractable integral. Since \(A\) is the whole parameter set of the iSSVM, the joint posterior distribution \(p(A|D)\) does not have a closed form. Thus, Gibbs sampling [14] is used to obtain samples from this joint posterior distribution. In sampling, the auxiliary variables \(z = \{z_1, \ldots, z_N\}\) (which are the indicator variables for the training data) are introduced. The samples \(A^{(k)}\) are obtained by sampling from \(p(A, z|\lambda, D)\), yielding \(\{A^{(k)}, z^{(k)}\}\). \(A^{(k)}\) can be considered as being sampled from \(p(A|\lambda, D)\) [15]. \(M_k\) is the number of unique values of the sampled indicators \(z^{(k)}\).

\(\Theta\) is the parameter set of the gating network which here is a DP mixture model. The conditional posterior distribution of the parameter set can be described as follows:

\[
p(\Theta|H^{(k)}, z^{(k)}, \lambda, D) = p(\Theta|\phi(O_n, \lambda))_{n=1}^{N}, z^{(k)}
\]

where \(\phi(O_n, \lambda)\) is the feature space for the utterance \(O_n\), which maps the observation \(O_n\) to a space with fixed dimension. The feature is the log-likelihood feature of the whole utterance. Here, the normalised features based on segments are used: \(\phi(O_n, \lambda) = (1/T_n) \sum \phi(O_i, \lambda)\) where \(T_n\) is number of frames in utterance \(O_n\), and \(\phi(O_i, \lambda)\) is the log-likelihood feature described in equation (4). Given the features \(\{\phi(O_n, \lambda)\}_{n=1}^{N}\) and corresponding indicators \(z^{(k)}\), \(\Theta^{(k)}\) can be sampled through the methods described in [11, 12].

In terms of the parameters of the experts \(H\), each expert is a log-linear model with large margin training, so the parameter of the \(m^{th}\) expert \(\eta_m\) is obtained through equation (6) with the data associated with expert \(m\). If there are few observations associated with an expert, generalisation can become a problem. Thus, each expert uses an informative prior. Similar to the method used in [6], the mean of the prior \(\mu_m\) is obtained from the SSVM trained on the whole training set. By introducing this mean, the iSSVM should recover the performance of the SSVM, if \(C\) is small enough (i.e. if the variance is small enough). Better performance could be achieved by gradually increasing \(C\). The 1-slack cutting plane algorithm [16] is used to train the SSVM. The constraint set that this algorithm uses for training the current SSVM parameter \(\eta^{(k)}_m\) can be cached and propagate to the next iteration of obtaining \(\eta^{(k+1)}_m\). This caching method can make the training more efficient, especially when applying the iSSVM to large vocabulary CSR.

The indicator variable \(z_n\) is sampled according to the following posterior distribution:

\[
P(z_n = m|A^{(k)}, z_{-n}, \lambda, D) \propto \frac{P(z_n = m|z_{-n}, \lambda) p(\phi(O_n, \lambda|\Theta^{(k)})) P(W_n, \rho_n|O_n, \lambda, \eta^{(k)}_m)}{\sum_{m=1}^{M_k} P(z_n = m|z_{-n}, \lambda) p(\phi(O_n, \lambda|\Theta^{(k)})) P(W_n, \rho_n|O_n, \lambda, \eta^{(k)}_m)}
\]

where \(z_{-n}\) denotes all the indicators except \(z_n\). The first term \(P(z_n = m|z_{-n}, \lambda)\) is given by the Chinese Restaurant Process (CRP) with concentration parameter \(\alpha\) [17]. The term \(P(W_n, \rho_n|O_n, \lambda, \eta^{(k)}_m)\) is the posterior distribution given by the log-linear model described in equation (2), and \(p(\phi(O_n, \lambda|\Theta^{(k)}))\) is the component likelihood. When \(z_n\) indicates an existing expert, it is straightforward to calculate the conditional posterior distribution of \(z_n\). When \(z_n\) denotes a new expert, following the method introduced in [11], in calculating the likelihood \(p(\phi(O_n, \lambda|\Theta))\), the parameter \(\Theta\) is sampled from its prior distribution as an auxiliary parameter, then the likelihood can be easily obtained. In order to make the newly generated expert have good generalisation, in calculating the third term, the parameter for the expert \(\eta\) is given as the the mean of its prior, namely the optimised parameter of the SSVM trained on the whole training set.

5. CLASSIFICATION

The equation used for classification has been given in equation (9). By substituting the log-linear model given in equation (2) into (9), it becomes:

\[
P(W, \rho|O, \lambda, D)
\]

\[
\approx \frac{1}{K} \sum_{k=1}^{K} \prod_{m=1}^{M_k} P(z_k = m|O, \lambda, \Theta^{(k)}) \frac{\exp \left( \eta^{(k)}_m^T \Phi(O, W; \lambda, \rho) \right)}{S_m^{k}}
\]

where all possible \((W, \rho)\) are obtained from a lattice, and \(S_m^{k}\) is the normalisation term:

\[
S_m^{k} = \sum_{W', \rho'} \exp \left( \eta^{(k)}_m^T \Phi(O, W'; \lambda, \rho') \right)
\]

In the SSVM, this term can be ignored, since no posterior needs to be calculated and the normalisation term stays the same for all possible labels. For the iSVM on the other hand, the posterior, given by the log-linear model, does need to be calculated. Thus, this term cannot be ignored, but it is trivial to calculate, since the possible number of labels are small for each segment. In the iSSVM, the calculation of this term \(S_m^{k}\) is nontrivial, since the possible number of labels are exponentially large for the utterance \(O\).

Given that the possible number of labels are extremely large, the summation in equation (13) is quite inefficient. The forward algorithm can be adopted to calculate this summation efficiently on the lattice. According to the definition of the joint feature space given in equation (3), and describing the parameter of the \(m^{th}\) SSVM as \(\eta^{(k)}_m = \ldots\)
Table 1. The results on Aurora 2 database

<table>
<thead>
<tr>
<th>System</th>
<th>Features</th>
<th>Test Set WER(%) testa</th>
<th>testb</th>
<th>testc</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>MFCC</td>
<td>9.83</td>
<td>9.11</td>
<td>9.53</td>
<td>9.48</td>
</tr>
<tr>
<td>iSVM</td>
<td>Log-Like</td>
<td>8.29</td>
<td>7.90</td>
<td>8.61</td>
<td>8.20</td>
</tr>
<tr>
<td>SSVM</td>
<td>Log-Like</td>
<td>8.25</td>
<td>7.87</td>
<td>8.53</td>
<td>8.15</td>
</tr>
<tr>
<td>iSSVM</td>
<td>Joint Feat</td>
<td>7.78</td>
<td>7.29</td>
<td>7.98</td>
<td>7.63</td>
</tr>
</tbody>
</table>

Fig. 2. The iSSVM performance on set A with different C

6. EXPERIMENTS AND CONCLUSIONS

The performance of the proposed iSSVM is evaluated on the Aurora 2 database [19]. The utterances in this database are continuous digit strings with vocabulary size 12 (one to nine, plus zero, oh and silence). The noise model for VTS compensation [20] is estimated on each utterance, and the performance of the VTS compensated HMM (which is used to obtain log-likelihood features for the discriminative models in the experiments) is listed in Table 1. The SVM, iSVM, SSVM and iSSVM are trained on a subset of the multi-style training data containing 3 noise conditions (N2, N3 and N4) and 3 SNRs (20dB, 15dB and 10dB). All 3 test databases, A, B and C with numbers of utterances 20020, 20020 and 10010 respectively, are used in the evaluation.

In the experiments, the log-likelihood features described in equation (4) are used by the SVM and iSVM, and the joint features described in equation (3) are used by the SSVM and iSSVM. All experiments are conducted with the number of samples \( K = 10 \). The results are listed in Table 1. On test set A and C, the iSSVM get around 3% relative improvement in all SNRs, but only a small improvement is achieved on test set B. The large margin training criterion described in equation (6) is adopted to train the experts (SSVM) of the iSSVM, and different experts share the same \( C \). The parameter \( C \) is tuned on test set A, with word error rates illustrated in Fig. 2. Since the prior mean \( \mu_w \) in equation (6) is given the optimised parameter of the SSVM trained on the whole training set, when \( C \) is small, the SSVM performance is recovered, and the optimised \( C \) can be found by increasing \( C \).

This paper studies the combination of Bayesian non-parametrics with structured models. Specifically, the infinite structured SVM is detailed, which is an extension of the iSVM described in previous paper [6]. Taking advantage of the infinite mixture of experts that are structured models, the iSSVM outperforms the iSVM and SSVM. As discussed in Section 4, the indicator variable of the iSSVM is a scalar, which means all the segments in an utterance share the same indicator. This might limit the flexibility of the gating network. In order to make better use of the data, a more granular (vector) indicator could be introduced. Thus, future work will study a structured model with a mixture of experts for each word, the structured infinite SVM (SiSVM).
7. REFERENCES


