SUPPORT VECTOR MACHINES FOR SPEECH RECOGNITION

Aravind Ganapathiraju, Jonathan Hamaker, Joseph Picone

Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University, Mississippi State, Mississippi 39762
{ganapath, hamaker, picone}@isip.msstate.edu

ABSTRACT

A Support Vector Machine (SVM) is a promising machine learning technique that has generated a lot of interest in the pattern recognition community in recent years. The greatest asset of an SVM is its ability to construct nonlinear decision regions in a discriminative fashion. This paper describes an application of SVMs to two speech data classification experiments: 11 vowels spoken in isolation and 16 phones extracted from spontaneous telephone speech. The best performance achieved on the spontaneous speech classification task is a 51% error rate using an RBF kernel. This is comparable to frame-level classification achieved by other nonlinear modeling techniques such as artificial neural networks (ANN).

1. INTRODUCTION

Hidden Markov Models (HMM) have been the core of most speech recognition systems for over a decade. Most HMM systems use Maximum Likelihood (ML) approaches for training. There are two major problems with this framework. First, it is assumed that the model (topology and density functions) actually reflects the structure of the data, though learning the structure from the data would be a better idea. Second, increasing the representative power of the model is not the best criterion for achieving optimum speech recognition performance [1]. Discriminative approaches such as maximum *a posteriori* probability (MAP) estimation have proven to produce lower error rates [1,5] on specific applications. Unfortunately, such techniques tend to be very application specific.

The incremental model optimization approach in an ML framework simplifies the training process though losing discriminative information in the process. This is due to the fact that all rival state sequences (corresponding to other models) are not considered during optimization of parameters for a given model. However training by optimization over the entire parameter space gives better discriminative power to the models since the models now also learn patterns that need to be discriminated. ANNs are good at this type of learning [3] since the training involves a joint-optimization process.

Recently a different class of learning machines called Support Vector Machines (SVM) have attained prominence due to their inherent discriminative learning and generalization capabilities [2,6]. Advances in training techniques have facilitated their application to tasks involving large data sets [9].

In this work we present preliminary efforts to apply SVMs to phoneme classification as a first step towards integration into a complete speech recognition system. We demonstrate the efficacy of this classification scheme using two types of data: the Deterding Vowel data [3], and a subset of the Switchboard [10] corpus consisting of 16 phones extracted from continuous speech. Classification results described below are extremely encouraging given the rather simple approach we have used to integrating the classifier.

2. SUPPORT VECTOR MACHINES

The underlying concept behind an SVM is structural risk minimization [2]. A learning machine is chosen that minimizes the upper bound on the risk (or test error), which is a good measure of the generalizability of the machine. This is estimated as the ratio of misclassified vectors over the total number of training vectors when using a “leave-one-out” method [6]. It can be shown that this is equal to the ratio of expected number of support vectors to the total number of training vectors.

The power of SVMs lies in transforming data to a high dimensional space and constructing a linear binary classifier in this high dimensional space. Construction of a hyperplane in a feature space requires transformation of the n-dimensional input vector $\boldsymbol{x}$ into an $\boldsymbol{N}$-dimensional feature vector, i.e.

$$\Phi: \Re^n \rightarrow \Re^N \ . \tag{1}$$

An N-dimensional linear separator $\boldsymbol{w}$ and a bias b are then constructed for the set of transformed vectors. Classification of an unknown vector $\boldsymbol{x}$ is done by first transforming the vector to the feature space and then computing

$$\mathrm{sgn}(\boldsymbol{w} \cdot \phi(\boldsymbol{x}) + b) \ . \tag{2}$$

The vector $\boldsymbol{w}$ can be written as a linear combination of a small set of vectors in the feature space. This can be mathematically expressed as

$$\boldsymbol{w} = \sum \alpha_i \cdot y_i \cdot \phi(\boldsymbol{x_i}) \ , \tag{3}$$

where the summation is over all vectors in the training set whose corresponding α's are non-zero. These vectors are called *support vectors* [2]. Combining Equations 2 and 3, the classifier becomes

$$\mathrm{sgn}\left(\sum_{SVs} y_i \alpha_i \phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{x_i}) + b \right) . \tag{4}$$

The above equation is prohibitively expensive to implement directly in the feature space, since it would involve a dot product computation in a very high dimensional space. However if we could define a function in the input space which equals the dot product in the feature space, the overall complexity of the process can be drastically reduced since we significantly lower the dimensionality. The existence of such a function is guaranteed by Mercer's conditions [2,6]. Such functions are referred to as a kernel in the SVM approach. Thus Equation 4 can be equivalently written as

$$\mathrm{sgn}\left(\sum_{SVs} y_i \alpha_i K(\boldsymbol{x}, \boldsymbol{x_i}) + b \right) \ , \tag{5}$$

where K is the kernel. Some widely used kernels are:

Polynomial: $K(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u} \cdot \boldsymbol{v} + 1)^d$

2-layer NN: $K(\boldsymbol{u}, \boldsymbol{v}) = Sigmoid[s(\boldsymbol{u} \cdot \boldsymbol{v}) + c]$

RBF: $K(\boldsymbol{u}, \boldsymbol{v}) = \exp\{-\Upsilon|\boldsymbol{u} - \boldsymbol{v}|^2\}$

Since the decision region is dependent on the data set, by using prior knowledge of the data and the characteristics of various kernels, we can achieve better performance. For example, if a data set is known to need closed decision regions, it is better to use an RBF kernel rather than a linear or a low order polynomial kernel.

3. EXPERIMENTS AND RESULTS

Since SVMs have proven to be effective on classical pattern recognition problems, a logical progression was to apply it to classification of phonetic segments in speech. The Deterding vowel data [3], which consists of 11 vowels from British English spoken by 15 speakers in an h*d context, was chosen for this purpose. This task, though widely used to benchmark nonlinear classification algorithms, is not of immediate interest in continuous speech recognition because of a lack of variation in the phonetic context.

Hence, we decided to do frame-level classification experiments on a reduced phone set of the Switchboard Corpus [10]. In this work we used an SVM toolkit, *SVMlight* that is available as shareware [9]. It uses a variant of a training algorithm referred to as *Chunking* developed by Osuna et. al. [4]. This SVM package can be applied to large datasets and is capable of handling classification tasks with tens of thousands of support vectors.

Since SVMs are inherently binary classifiers, application to a multi-class problem involves using a mixture-of-experts paradigm [6]. We use a simple strategy to handle this that involves representing a given test vector classified as a positive example with the decision boundary that is furthest from the sample for the right classification. In case of data classified as a negative examples, the decision boundary with the shortest distance is chosen as the correct classification. Computationally this is achieved by

$$\arg max_{f\,=\,1,\ldots,16} \left(\sum_{SVs} \alpha_i^f y_i \cdot K(\boldsymbol{z}, \boldsymbol{z_i}) + b^f \right) . \tag{6}$$

for positive classifications and,

$$\arg min_{f = 1....16}\left(\sum_{SVs} \alpha_i^f y_i \cdot K(\boldsymbol{z}, \boldsymbol{z_i}) + b^f \right) \qquad (7)$$

for negative classifications.

3.1. Vowel Classification

In our first pilot experiment, we applied SVMs to a publicly available vowel classification task [3]. In this evaluation, the speech data was collected at a 10 kHz sampling rate and low pass filtered at 4.7 kHz. The signal was then transformed to 10 log-area parameters, giving a 10 dimensional input space. A window duration of 50 msec. was used for generating the features. The training set consisted of 528 frames from eight speakers and the test set consisted of 462 frames from the remaining seven speakers. Table 1 shows the performance of two type of kernels on the test data. Performance using both the kernels is better than most nonlinear classification schemes [3]. The best performance reported on this data set is, however, 29% error using a speaker adaptation scheme called Separable Mixture Models [8]. Neural network classifiers (Gaussian Node Network) produce a misclassification rate of 45% [3].

3.2. Switchboard Phone Classification

In our next experiment, 16 phones were extracted from selected utterances in the Switchboard Corpus. The phones were chosen to represent vowels, the fricatives 's' and 'f' and the liquids 'l' and 'r'. The segmentation was based on a 44 phone context-independent system. Feature vectors were generated by computing 12 mel-scaled cepstra along with energy. A frame duration of 10 msec. and a window duration of 25 msec. was used for data generation.

order/gamma/ hidden-units	RBF	Polynomial	Gaussian Node Network
2/0.025/22	32	**42**	46
3/0.05/88	**31**	44	47
4/0.1/528	32	45	**45**

Table 1: Misclassification rate of SVMs using an RBF kernel and a Polynomial kernel on vowel data

To avoid dealing with problems associated with the optimization process involved in training the database, we clustered the data for each phone into 200 clusters using 5000 exemplars. A simple K-MEANS algorithm with a mean-squared error distance measure was used for the clustering process. However, to avoid clusters representing features with large average values, we normalized the features to a [-1, +1] range before the clusters were generated. The test set was chosen from the normalized data to represent a speaker independent portion. It consisted of 100 exemplars per phone to a total of 1600 test vectors. Table 2 shows the classification results obtained by using RBF and polynomial kernels. Note that for the RBF network, going beyond a *gamma* value of 10 is not viable since the distance from the SVM margin has a low variance and hence cannot be effectively used for likelihood computations.

3.3. ANN-Based Phone Classification

ANNs have been used for estimating phone probabilities in hybrid-HMM systems for over a decade now. Their performance has been better than classical Gaussian classifiers on frame-level speech data. Morgan et.al. [1] report performance of a multi-layered perceptron (MLP) and a Gaussian classifier on speech data extracted from a speaker independent German continuous speech database, *SPICOS*. Thirty mel-scaled cepstral features were used. A phone set of 50 phones was classified. A frame misclassification rate of 35% was achieved using a continuous Gaussian classifier. On the same test set, an MLP with a context of 21 frames achieved a misclassification rate of 32%. With a context of 9 frames the performance was at 41%. At first glance, these numbers seem much better than the performance of the SVM systems. However, it is worth noting some important differences/ simplifications in the SVM system which include lack of context and smaller feature dimension. Larger acoustic context and larger feature dimension have been known to provide vast performance improvements on speech data.

3.4. Observations

An analysis of our preliminary experiments demonstrates some interesting trends. On the vowel classification data, though there were 11 different classifiers trained, 36% of the support vectors were

Order/Gamma	RBF	Polynomial
2/0.1	68	68
3/1	64	66
5/10	**51**	68

Table 2: Misclassification rate of SVMs using an RBF and a Polynomial kernel on Switchboard phone data

shared by at least 2 classifiers. This closely parallels our notion of tied Gaussians in traditional HMM systems. On the other hand, in classifiers trained for Switchboard, the amount of sharing was much less — only 8%. The lack of sharing of support vectors makes one believe that the underlying structure of the data suggests the need for more phonetic classes than the 16 we selected. The topic of an expanded phone set for Switchboard, or even sub-phonetic elements, has been a topic of great debate [7].

4. CONCLUSIONS AND FUTURE WORK

SVMs have enjoyed widespread acceptance over the past few years as an efficient nonlinear learning machine for static pattern classification problems. Development of various efficient optimization techniques for training of these classifiers has played a major role in this recent success. In this work we have demonstrated encouraging performance of SVMs on two types of speech data.

The performance of SVMs on vowel data is better than most nonlinear classification techniques such as ANNs, k-NNs, and Gaussian classifiers. On the Switchboard task, the lowest misclassification rate achieved was 51% which compares well with performances obtained by the classical Gaussian classifiers and MLPs.

The mixture-of-experts scheme described here is easily integrated with likelihood computations that are an integral part of a speech recognition system. Future plans include replacing the Gaussians in HMM states with SVMs. In order to convert distance measures into probabilities, we plan on generating a regression model for combining distances so that the sum of the distances across all the classifiers sum to unity, similar to a probability measure.

5. REFERENCES

1. H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition*, Kluwer Academic Publishers, Boston, MA, USA., 1994.

2. V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, NY, USA,1995.

3. A. J. Robinson, "Dynamic Error Propagation Networks," *Ph.D. Thesis*, Cambridge Univ. Eng. Dept., February 1989.

4. E. Osuna, et. al. "An Improved Training Algorithm for Support Vector Machines," *Proceedings of the IEEE NNSP'97*, pp. 24-26, Amelia Island, FL, USA, September 1997.

5. M. Hochberg, et. al., "Large Vocabulary Continuous Speech Recognition Using a Hybrid Connectionist/HMM system," *Proceedings of the ICSLP*, Yokohama, Japan, pp. 1499-1502, September 1994.

6. B. Schölkopf, *Support Vector Learning*., *Ph.D. Thesis*, R. Oldenbourg Verlag Publications, Munich, Germany, 1997.

7. S. Greenberg, "The Switchboard Transcription Project," *Technical Report of the 1996 LVCSR Summer Research Workshop*, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA., April 1997.

8. J. Tenenbaum, et. al., *Advances in Neural Information Processing Systems 9*, MIT Press, Cambridge, MA, USA., 1997.

9. T. Joachims, *Making Large-scale SVM Learning Practical*: *Advances in Kernel Methods - Support Vector Learning*, MIT Press, Cambridge, MA, USA., 1998.

10. J. Godfrey, E. Holliman and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development", *Proceedings of the IEEE ICASSP*, vol. 1, pp. 517-520, San Francisco, CA, USA, March 1992.