SUPPORT VECTOR MACHINES FOR AUTOMATIC DATA CLEANUP

Aravind Ganapathiraju and Joseph Picone

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

ABSTRACT

Accurate training data plays a very important role in training effective acoustic models for speech recognition. In conversational speech, in several cases, the transcribed data has a significant word error rate which leads to bad acoustic models. In this paper we explore a method to automatically identify such mislabelled data in the context of a hybrid Support Vector Machine/hidden Markov model (HMM) system, thereby building accurate acoustic models. The effectiveness of this method is proven on both synthetic and real speech data. A hybrid system for OGI alphadigits using this methodology gives a significant improvement in performance over a comparable baseline HMM system.

1. INTRODUCTION

A measure of the confidence of a speech recognizer's output has been used in many areas under the general topic of *confidence measures* [1]. Confidence measures can be used to reject hypotheses which are likely to be in error. Neural network-based systems have traditionally used word and phone posterior probability-based confidence measures to post-process recognition output. Limited success has been reported using such techniques [2].

Confidence measures can significantly impact the acoustic model training problem in speech recognition. The availability of accurately transcribed speech corpora, especially for conversational speech, has been a problem for several years now. For example, Switchboard, a commonly used corpus to calibrate conversational speech recognition systems has a transcription word error rate of ~8 %[1]. Acoustic model estimates can degrade significantly when estimated using such mislabelled data. Systems that are robust to transcription errors are extremely useful. An ability to reliably detect mislabelled data automatically is also very important.

A Support Vector Machine (SVM) is a new machine learning technique that is based on principles of discrimination [3]. This paradigm has gained prominence in recent years with the development of efficient training algorithms [4,5]. Hybrid speech recognition systems combining hidden Markov models (HMM) and SVMs are a promising area of research [6,7].

In this paper we explore the capability of SVMs to identify mislabelled data (or outliers) during the training process. Though this is not done via a true confidence measure, it is a first step towards automatically handling mislabelled data. SVMs are shown in the paper to be inherently suited to this task. A hybrid system trained using data cleaned using the above capability is described. This system has been evaluated on the OGI Alphadigits Corpus [8].

2. SUPPORT VECTOR MACHINES

The power of SVMs lies in their ability to transform data to a high dimensional space where the data can be separated using a linear hyperplane [3]. The optimization process for SVM learning begins with the definition of a functional that needs to be optimized in terms of the parameters of a hyperplane. The functional is defined such that it guarantees good classification (if not perfect classification) on the training data and also maximizes the margin (e.g. the distance between H1 and H2 in Figure 1). The points that lie on the hyperplane satisfy,

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = 0 \quad (1)$$

where $\boldsymbol{w}$ is the normal to the hyperplane and b is the

1. Note that a new version of the Switchboard transcriptions are available at *http://www.isip.msstate.edu/projects/switchboard/* that have a word error rate of less than 1%.

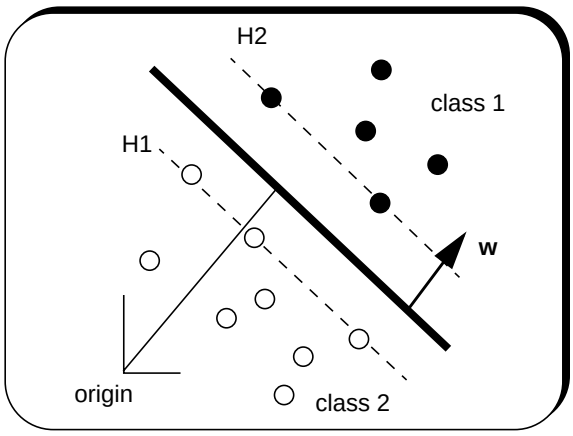

Figure 1. 2-class hyperplane classifier example

bias of the hyperplane from the origin. Let the N training examples be represented as tuples $\{\boldsymbol{x}_i, y_i\}, i = 1, \ldots, N$ where $y = \pm 1$ are the class labels. They satisfy the following constraints,

$$y_i(\boldsymbol{x}_i \cdot \boldsymbol{w} + b) - 1 \geq 0 \qquad \forall i \tag{2}$$

The distance between the margins can be shown to be $2/\|\boldsymbol{w}\|$ [3]. The goal of the optimization process should be to maximize the margin. Posing this as a quadratic optimization problem has several advantages and the functional can be compactly written as,

$$L_P = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{N} \alpha_i y_i(\boldsymbol{x}_i \cdot \boldsymbol{w} + b) + \sum_{i=1}^{N} \alpha_i \tag{3}$$

where the α_i 's are Lagrange multipliers.

3. PRACTICAL SVM ESTIMATION AND OUTLIER IDENTIFICATION

Quadratic optimization is typically a very computationally expensive process. When the number of training examples is in the thousands, efficient training algorithms need to be designed. When the number of training examples is N, the number of dot products that need to be evaluated is of the order of N^2 which can get restrictive when N is large. To make the optimization process computationally feasible, small chunks of data are processed at a time. The following section describes a SVM estimation procedure called *Chunking* which allows efficient SVM estimation even when the number of training examples is large [4].

3.1. Chunking and Working Sets

Chunking is based on the idea of dividing the optimization problem into sub-problems whose solution can be found efficiently. This method divides the training data into chunks and optimizes the functional for each chunk. Osuna proves that the chunking algorithm does in fact give the same solution as a global optimization process but takes much less operating memory and time [4].

The Chunking algorithm can be specified in three simple steps. Suppose we define the working set as B and the non-working set (whose multipliers do not change while solving the sub-problems) as N.

1. Choose $|B|$ training points from the data set at random.
2. Solve the optimization problem defined by the set B.
3. For some example $j \in N$, which violates the optimality constraints, replace λ_i, $i \in B$, with λ_j and the solve the new sub-problem.

The above algorithm is guaranteed to strictly improve the objective function based on the observations made by Osuna. The convex quadratic form of the objective function also guarantees that the algorithm will reach the global optimum solution within a finite number of iterations. The key is to choose the working set such that the algorithm converges to the final solution rapidly using methods such as the steepest feasible descent [5]. Figure 2 depicts the chunking algorithm.

3.2. Bounded Support Vectors

Apart from choosing a good working set, the optimization process can be made efficient by identifying support vectors, whose multipliers are at the upper bound, C, early in the training process.

For noisy data, (for example, classes with significant overlap possibly due to mislabeling), there are often several support vectors with their multipliers at the upper bound C. When an example has its multiplier consistently at the upper bound across iterations, it is a good indication that the example is either an

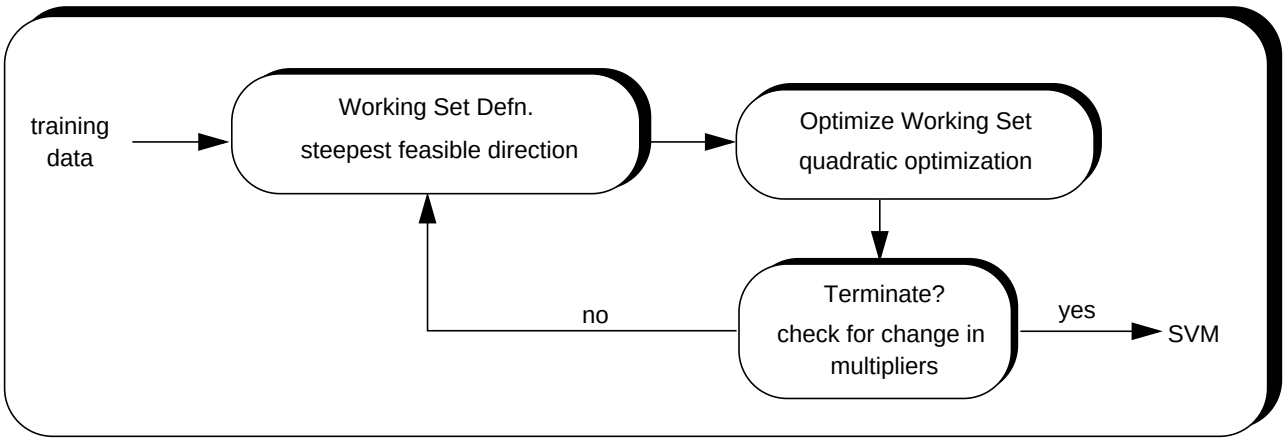

Figure 2. An overview of the SVM estimation process. The working set definition is where mislabelled data is identified and removed from the estimation process.

outlier, an area of overlap between features, or is mislabelled data.

Removing these bounded support vectors from the optimization problem helps reduce the size of later iterations. In the work presented in this paper, the bounded support vectors are removed from the optimization process altogether and are considered as mislabelled data, especially when C is large. This is the premise under which the following experiments are reported. Note that when the bounded support vectors are not removed form the optimization process they end up as support vectors for the final solution. In the case of noisy data this could lead to classifiers modeling inaccurate decision surfaces as reported in the following section.

4. EXPERIMENTS

In this work we have studied the efficiency of identifying mislabelled data using the proposed technique on real speech data.

4.1. Synthetic Data Example

Figure 3 shows the dataset used to perform initial experiments to confirm the data cleanup abilitiy of the proposed method. The data was generated by choosing two mel-cepstral coefficients for the phones “s” and “f” in the OGI alphadigit corpus. When the bounded support vectors are not eliminated from the optimization process, they end up as support vectors which results is a skewed desicion surface. The proposed method is able to identify 87% of the mislabeled data points.

4.2. Hybrid SVM/HMM System

A hybrid SVM/HMM system has been trained on the OGI alphadigits corpus using the proposed method to identify mislabelled data [6]. In this system SVMs are trained using alignments provided by the baseline HMM system. Since the baseline system has a significant WER, training vectors generated in this manner are often mislabelled and the need for data cleanup becomes critical.

The baseline HMM system was trained on 39-dimensional feature vectors comprised of 12 cepstral coefficients, energy, delta and acceleration coefficients. The training set for the HMM system had 50,000 sentences averaging 6 words a sentence while the SVM classifiers were trained using only 9000 sentences. The test set was an open-loop speaker independent set with 1000 sentences. The system performs at 11.6% WER which is better than the baseline cross-word triphone HMM system with 8 Gaussian mixture components per state which gives 12.7% WER on this dataset. The hybrid system performs at 12.1% when mislabelled data is not culled from the training set.

5. CONCLUSIONS

In this work we have explored a unique method of cleaning up mislabelled speech data implicitly as part of the SVM estimation process. The identification of mislabelled data helps build more

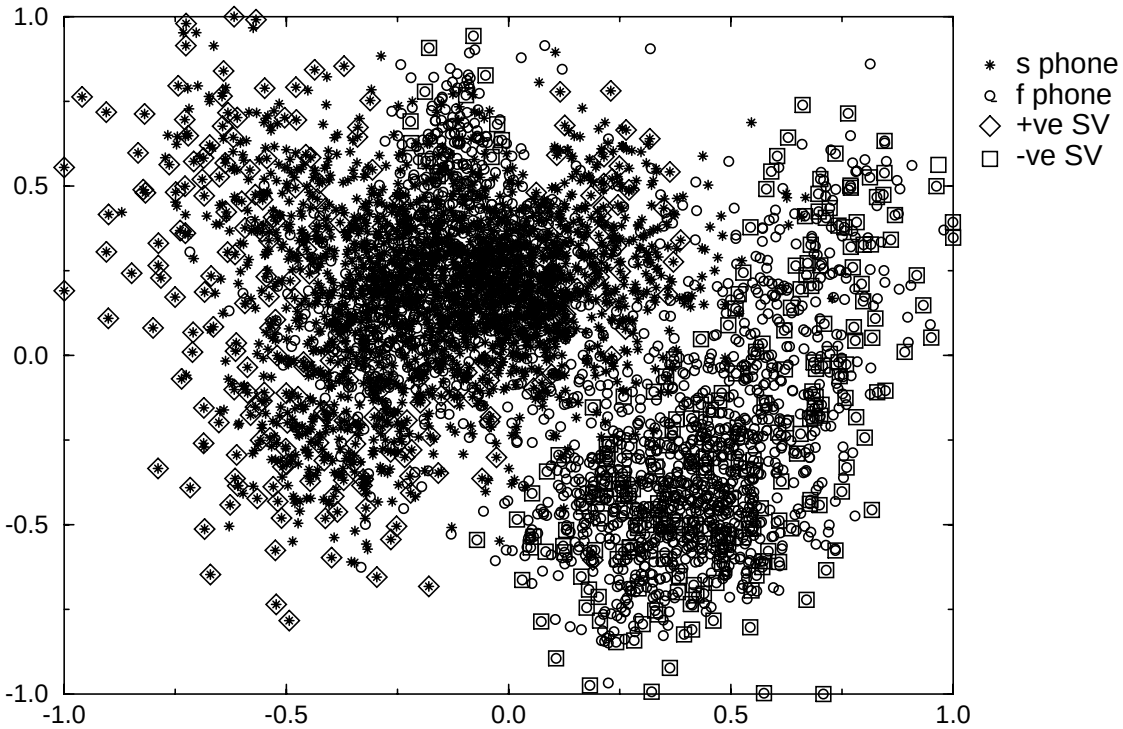

Figure 3. Two-dimensional example with significant percent of mislabeled data points.

accurate acoustic models and also speeds up the SVM estimation process. This method has been proven to be effective for a small vocabulary task, OGI Alphadigits, where a hybrid SVM/HMM system trained using the proposed method performs 10% relatively better than the baseline HMM system.

Though this method does not bear direct resemblance to confidence measures often used in speech recognition systems, we are in the process of developing a mathematical formalism to quantify the mislabelled data thereby making the proposed method compatible with other methods of measuring confidence in classifier output.

6. REFERENCES

1. G. Williams and S. Renals, “Confidence Measures for Hybrid HMM/ANN Speech Recognition,” *Proc. Eurospeech ‘97*, pp. 1955-1958, Rhodes, Greece, September 1997.

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

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

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. T. Joachims, SVMLight: Support Vector Machine, *http://www-ai.informatik.uni-dortmund.de/FORSCHUNG/VERFAHREN/SVM_LIGHT/svm_light.eng.html*, University of Dortmund, November 1999.

6. A. Ganapathiraju, et. al., "Hybrid SVM/HMM Architectures for Speech Recognition," *Department of Defense Hub 5 Workshop*, College Park, Maryland, USA, May 2000.

7. A.Ganapathiraju, et. al., “Support Vector Machines for Speech Recognition,” *Proc. of the ICSLP*, pp. 2923-2926, Sydney, Australia, November 1998.

8. R. Cole et al, “Alphadigit Corpus,” *http://www.cse.ogi.edu/CSLU/corpora/alphadigit*, Oregon Graduate Institute, 1997.