ESSV 2012

CONTINUOUS SPEECH RECOGNITION USING CORRELATION FEATURES AND STRUCTURED SVM PROBABILITY OUTPUT

Andreas Beschorner[1,2], Dietrich Klakow[1]
[1]*Department of Spoken Language Systems, Saarland University, Germany*
[2]*Precitec Vision, Neftenbach, Switzerland*
{*andreas.beschorner, dietrich.klakow*}*@lsv.uni-saarland.de; a.beschorner@precitec.ch*

Abstract: One potential area for improvement in continuous speech recognition is the modelling of phoneme transitions (not transition probabilties) arising from the non-stationarity of speech: refined models can then be used to compute probability distributions which can serve as emission probabilities for HMM-based speech recognition systems. In this paper we present our approach to improving phoneme transition modelling. Building on our previous work, we employ a phoneme partition approach (SME: start, middle, and end states) to build a structure of support vector (SV) classifiers as our main discriminative method. For the phoneme classification step, cross correlation features based on MFCC-vectors are computed and classified within the SME structure. Additionally, we make use of a special reproducing kernel build upon the correlation features, thus offering a direct integration into the SV classifiers. This paper discusses the computation of the afore-mentioned probability outputs as well as initial results using these outputs as emission probabilities in HMMs representing phonemes, applied within a standard speech recognition system.

1 Introduction

Continuous speech recognition can be broken down into a number of connected problems, among them phoneme classification, modellation of both the transitions within and between speech sounds such as phonemes and of the correspondences between the acoustic signal and elements of natural language. Concepts like HMMs have long been in use for speech recognition and phoneme classification; in recent years, systems have been influenced for instance by generative models like GMMs [15], maximum a posteriori adaptive sequence estimation [7], discriminative methods [19] and along with the latter the theory of reproducing kernels (RKs) and reproducing kernel Hilbert spaces (RKHSs). In the context of RKs, sequence kernels [6] have been developed, aiming at capturing the non-static nature of processes such as speech or even modelling HMMs (see [17], pp. 430–436). The same holds for previous hybrid concepts combining SV classifiers and HMM methods ([10], [11], [2]). Such progresses as well as the success of often sophisticated RKs capturing the specific nature of certai tasks or data structures motivate our approach.
A disadvantage of many current classification models is, despite the characteristic of frequently used MFCC-features, the failure to account for the transitional aspects of the nature of speech signals; human speech is not a sequence of stable phonemes but rather a constant process of moving from one speech sound to another. In contrast to other methods modeling the non-stationarity such as LDA, our concept will break the restriction of linearity, making use of in general non linear reproducing kernel methods. Again, individual phonemes are greatly

influenced by their phonetic context, and one goal of our work is to model intra-phoneme transitions. Toward the first end, we partition phoneme samples to induce a tristate **S**tart-, **M**iddle- and **E**nd (SME) model of phoneme transitions. This was introduced in [5] along with a a method for integrating cross-correlation features, developed in the SME context, into a reproducing kernel, thereby reducing feature-computation and connecting similarity/distance (used for classification) directly to the used features. The model is implemented with an array of pairwise binary support vector (SV) classifiers, following the afore-mentioned SME structure. Toward the second end, we will integrate this into a common HMM based speech recognition system, interpreting the SVMs output in a probabilistic context.

The paper is organized as follows. Subsequent to a brief review of reproducing kernels and support vector machines, section 3 summarizes the ideas presented in [5]: *SME*-structuring of SV classifiers, based on a partition of phoneme samples in the spirit of tristates. The paper also introduces a method for integrating certain cross correlation features, developed and proposed in this context, into the reproducing kernel. A short discussion of SVM-based probability outputs and their usage in multiclass classification tasks follows in section 4. Section 5 gives details on the integration of the posterior probabilties and their integrations into the Millenium ([20]) speech recognition system. Finally, part 6 describes our experiments.

2 Reproducing Kernels and SVMs

2.1 Reproducing kernels

The concept of reproducing kernels is based on the fact that any Hilbert space $\mathscr{H}$ on a set X of complex-valued, bounded functionals endowed with an inner product $\langle\cdot,\cdot\rangle$ admits a mapping $k : X \times X \to \mathbb{C}$ such that for all $\mathbf{z} \in X$:

$$(1)\, k(\cdot,\mathbf{z}) \in \mathscr{H} \quad (2)\, \forall f \in \mathscr{H} : f(\mathbf{z}) = \langle f, k(\cdot,\mathbf{z})\rangle\,.$$

k is called a *reproducing kernel* and is unique within $\mathscr{H}$. It is easily verified ([3], [16]) that RKs defined as such are positive definite (pd). Conversely, for every pd mapping $k : X \times X \to \mathbb{C}$ there exists exactly one $\mathscr{H} \subset \mathbb{C}$ wherein k is a RK. Property (2) is called the *reproducing property* as the kernel reproduces the evaluation of the functional $f \in \mathscr{H}$ using the Hilbert space's inner product. Given such a k, the factorization lemma ([1]) furthermore implies the existence of a function $\Phi : X \to \mathscr{H}$ such that $k(\mathbf{x},\mathbf{z}) = \langle\Phi(\mathbf{x}),\Phi(\mathbf{z})\rangle$. Interpreting Φ as a feature mapping, reproducing kernels thus can replace potentially costly computations of Φ or inner products. Well known examples are the linear kernel $k_l(\mathbf{x},\mathbf{z}) = \mathbf{x}^T\mathbf{z}$, the polynomial kernel $k_{p^d}(\mathbf{x},\mathbf{z}) = (\mathbf{x}^T\mathbf{z}+r)^d$ and the exponential kernel $k_e(\mathbf{x},\mathbf{z}) = \exp(-\gamma||\mathbf{x}-\mathbf{z}||^2)$. Kernel functions are not generally restricted to numerical representations. Areas such as bioinformatics or part-of-speech-tagging make extensive use of kernels defined on strings or on more complex data structures like trees.

2.2 Support Vector Machines (SVMs)

Given a linear separable two-class dataset, SVMs compute a hyperplane $\mathbf{w}$ separating the two classes. The hyperplane is optimal in the sense that, amongst all hyperplanes separating the data, it has minimal margin, that is the distance from $\mathbf{w}$ to any (training) sample. Let $\mathscr{H}$ be an N-dimensional Hilbert space, M be the number of samples. Writing $\mathbf{x} = (x_1 \cdots x_N)$ for $\mathbf{x} \in X$ and $n = 1,\ldots,N$, let $\mathbf{w} \in \mathscr{H}$ and $\mathbf{x}_1,\ldots,\mathbf{x}_M$ be vectors in X. With $b \in \mathbb{R}$ being the bias or offset, $\{\langle\mathbf{w},\mathbf{x}\rangle + b = 0\,|\,\mathbf{x} \in \mathscr{H}\}$ is a subspace and hyperplane in $\mathscr{H}$ with normal vector $\mathbf{w}$. The

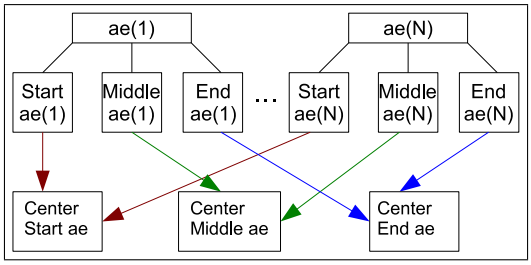

Figure 1 - Partitioning and clustering of N samples of class *ae*.

dot product equals the length of the projection of either component onto the direction of the remaining one.
The orientation of the hyperplane, $d(\mathbf{x}|\mathbf{w}) = \operatorname{sgn}(\langle \mathbf{x}, \mathbf{w} \rangle + b)$, serves as a decision criterion: For target labels $y_m \in \{\pm 1\}$, where $m = 1, \ldots, M$, the products $y_m \cdot d(\mathbf{x}|\mathbf{w})$ classify samples $\mathbf{x}$ into either class 1 or -1. The optimization problem of finding the hyperplane is subject to one constraint for each training sample: $y_m \cdot d(\mathbf{x}|\mathbf{w}) \geq 1, m = 1, \ldots, M$. To achieve better generalization, it has been proposed ([4] and, later [8]) to relax the constraints by introducing slack variables $\zeta_m \geq 1, m = 1, \ldots, M$. Using Langrangian multipliers $\alpha_m, m = 1, \ldots, M$ to optimize under those constraints, the final dual form of the optimization problem permits the substitution of the objective function's inner product by a RK k. The new decision function now is $d(\mathbf{x}|\mathbf{w}) = \sum_{m=1}^{M} \alpha_m y_m k(\mathbf{w}, \mathbf{x}_m) + b = 0$. As the dimension of the RKHS depends on the RK, data can be linearly separable in the RKHS even if this is not the case in the original space. For multiclass SVM cases, one-vs-one or one-vs-all strategies are commonly used, see [18] or [16] for details.

3 Linear mapping kernel and *SME*-structure

In [5], based on new correlation features and the concept of tristate phoneme representation, initial steps were taken towards continuous speech recognition using SVMs as a main classifier. We furthermore aimed at adjusting the kernel to the features while leaving the SV optimization problem as simple as possible. Both the correlation features and the specific classifier structure proposed capture transitional aspects (information between speech frames). A RK that integrates the correlation mapping of the features from the MFCC vectors computed from the phoneme samples is derived. The SV optimization itself remains untouched and does not suffer from additional complexity.

We start by computing standard MFCC features of phoneme samples, with the number of features per training sample dependent on its length. They are grouped classwise and partitioned into the above mentioned sections **S**tart-, **M**iddle- and **E**nd. Subsequently, representative *center vectors* are computed. Dropping class information, they are denoted by c_s, c_m and c_e. Our features are build from those center vectors by correlation with training data, see figure 2 for details. Figure 1 illustrates the process for phoneme *ae*. Training samples themselves are split into up to seven *SME*-based states: *SSS, SSM, SMM, MMM, MME, MEE* and *EEE*. They represent the time position in a phoneme sample. Figure 3 illustrates which 3-vector sequences contribute to which training set.
For the integration of the correlation mapping into the RK, consider a linear mapping $T : \mathscr{H} \to \mathscr{H}_T$ (not necessarily cross correlation) between a finite dimensional Hilbert space $\mathscr{H}$ and a Hilbert space $\mathscr{H}_T$ and its graph $G(T) = \{(\mathbf{x}, T\mathbf{x}) \,|\, \mathbf{x} \in \mathscr{H}\}$, in [5] we showed that for

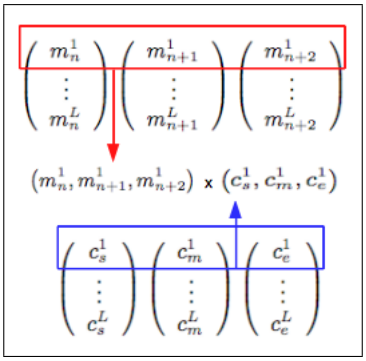

Figure 2 - Componentwise computation of the cross correlation features. The $\times$-symbol denotes the correlation operation. L is the size of the feature vectors, n the frame index of training data $\mathbf{m}$.

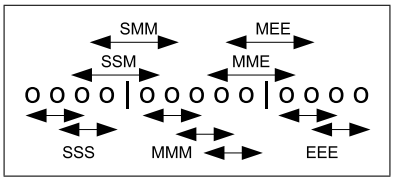

Figure 3 - Phoneme sample comprised of 13 MFCC-vectors/ frames and its *SME*-partitions. Two vector sequences will contribute to the phoneme specific classes *SSS* and *EEE*, three to *MMM* and one to *SSM, SMM, MME* and *MEE*.

any $p,q \in \mathscr{H}$ the following equation holds:

$$\langle (p,Tp),(q,Tq)\rangle_{G(T)} = \langle p,(w_{\mathscr{H}} I + w_{\mathscr{H}_T} T^*T)q\rangle_{\mathscr{H}}\,, \tag{1}$$

is positive definite for weights $w_{\mathscr{H}}, w_{\mathscr{H}_T} \in \mathbb{R}_+$ and thus defines a RK k_T. Important is that the new Kernel and the Hilbert space $\mathscr{H}$ are defined on the same set, thus allowing its use in kernel combinations such as addition or composition. In our work we apply the letter operation by composing the transform specific with the rbf kernel: $k_e \circ k_T$.
As an example for our setup where T realizes the correlation mapping of MFCC-vectors and center vectors c_s, c_m and c_e as described above, consider a point of time at which a sequence of three MFCC vectors covers the current phoneme position (=state) *SSM*, respectively. For one component $1,\ldots,L$ within equation (1) we then get, omitting the component's index,

$$T^*T_{ssm} = \begin{pmatrix} 2c_s^2+c_m^2 & 2c_s+c_sc_m & c_sc_m \\ 2c_s+c_sc_m & 2c_s^2+c_m^2 & 2c_s+c_sc_m \\ c_sc_m & 2c_s+c_sc_m & 2c_s^2+c_m^2 \end{pmatrix}.$$

The motivation for the correlation features is twofold: It measures similarity to a certain degree and modells class (or phoneme) specific transitions with representative center vector and works as a smoothing lowpass filter, lessening distortion and noise effects.

On the way to the greater goal of using discriminative classification within probability models like HMMs, we continue by following ([13]), using an improved version of Platt's approach ([14]), computing parameters of a sigmoid probability distribution by approximating the decision boundaries of the pairwise SV classifiers. Subsequently, we continue pursuing the track of [21]: Based on the sigmoid parameters and functions, prior probabilities for previously unseen

samples in a multiclass setting are estimated. The theoretical background is given in the next section.

4 Probability outputs for SVMs

4.1 The binary case

Let $i = 1,\ldots,N$ enumerate the trainingset, $y, y_i \in \{+1,-1\}$ be class labels and N_+, N_- denote the number of training samples of the respective classes. Furthermore, $x, x_i \in \mathbb{R}^n$ are test examples and $f(\cdot)$ the decision function of a binary SVM. Platt ([14]) approximates posterior probabilities based on $f(\cdot)$ by a sigmoid function

$$P_{a,b}(f(x)) := \frac{1}{1+\exp(af(x)+b)} \approx p(y=+1|x).$$

Writing $p_i := P_{a,b}(f(x_i))$, the parameters a, b are determined by

$$\begin{array}{ll} \text{minimize} \\ a,b \in \mathbb{R} \end{array} \quad -\sum_{i=1}^{N} [t_i \log(p_i) + (1-t_i)\log(1-p_i)] \tag{2}$$

and

$$t_i = \begin{cases} \frac{N_+ +1}{N_+ +2} & y_i = +1 \\ \frac{1}{N_- +2} & y_i = -1 \end{cases}, \quad i = 1,\ldots,N.$$

[13] propose rewriting the summand of problem (2) as

$$\begin{aligned} & [t_i \log(p_i) + (1-t_i)\log(1-p_i)] && (3) \\ = \; & (t_i - 1)(af_i + b) + \log\left(1+\exp(af_i+b)\right) && (4) \\ = \; & t_i(af_i+b) + \log\left(1+\exp(-af_i-b)\right) && (5) \end{aligned}$$

to address two major sources of numerical problems in Platt's algorithm: By substitution of $1-p_i$ (original version (3)) by either form (4) or (5), depending on the signature of af_i+b, they avoid numerical cancellation. The authors achieve sound stability and in line with this better classification results.

4.2 Multiclass classification tasks

In the case of multiple classes, in [21] various existing methods are discussed and two new proposed, the second of which we use in this work. Let $r_{i,j}$ be the pairwise class probabilities computed by evaluating the sigmoid functions using parameters a, b and $p_i = P(y=i|x), i = 1,\ldots,N$ be the sought estimates. Denoting the solution as a vector $\mathbf{p}$ of multi-class probability estimates, the authors prove that solving the optimization problem

$$\begin{array}{ll} \text{minimize} \\ \mathbf{p} \end{array} \quad \sum_{i=1}^{N} \sum_{j \neq i, j=1}^{N} (r_{j,i}p_i - r_{i,j}p_j)^2 \tag{6}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} p_i = 1$$

guarantees a unique solution. The classification decision function simply is

$$\text{argmax}_i(p_i). \tag{7}$$

The algorithm boils down to solving a system of linear equations using Gaussian elemination or, with small modifications to meet the preliminarity of positive definiteness, Cholesky factorization – for details, the reader is referred to the above mentioned paper.

5 From phonemes to utterances: Posterior probabilties and HMMs

We now move beyond straight phoneme classification to continuous speech recognition. The major contribution of this paper is the integration of our SME- and classifier net based structure into common systems for continuous speech recognition and the evaluation of the classification quality of the new architecture: The priors are transformed into posterior probabilities using Bayes' Rule to serve as phoneme emission probabilities in an HMM-based system.
Section 6 gives details on the setup for our experiments. For now it suffices to say that they are conducted on MFCC-vectors extracted from the TIMIT data set. Following [12], we reduced the amount of classes to 38 by merging similar ones. Dealing with continuous speech, several decisions have to be made about both the process of training and the setup of the final recognizer. While focusing on important characteristics of the *SME*-concept, we also want to keep things comparable to standard methods by adopting a certain amount of common consense. For our inaugural experiments we thus aim at being as close as possible to a standard tristate setup for individual phonemes.

In a first step, the up to seven (or even eight, including the three-frame *SME*-state) states, are reduced to three representative states, serving directly as the (usual) tristate representation of phonemes within HMMs. Due to space constraints, the details are not presented here; however, let us mention that given both the different distributions of phoneme lengths amongst the classes and the larger amount of the states *SSS, MMM, EEE* compared to *SSM, SMM, MME, MEE* by construction, both the number of samples of different states within the classes themselves and the number of samples of the same state for different phonemes are subject to great variance and are considered in the process.
For M classes the respective set of $3 \cdot M \cdot (M-1)/2$ pairwise SV-classifiers for the new set of states is computed, followed by the estimation of the parameters for the sigmoid probability functions on the validation set. Given an observation ω and classses $m = 1,\dots,M$, the multiclass prior probabilities $\tilde{\mathbf{p}} = p(\omega|m,s_m^i)$ for each of the $3 \cdot M$ decision values can now be approximated. Applying Bayes' Rule we compute the postertior probabilities for the speech decoder to get the probability for each class and state given the observation ω,

$$\mathbf{p}(m,s_m^i|\omega) = \left((p(m=1,s_{m=1}^1|\omega),\cdots p(m=M,s_{m=M}^1|\omega),\cdots\cdots,p(m=M,s_{m=M}^3|\omega)\right)^T. \quad (8)$$

For each observation, M such posterior probability vectors are computed. We apply a simple linear combination, which is potentially suboptimal w.r.t distribution considerations, but guarantees, that the resulting vector is a probability vector again. The coordinatewise negative log likelihood is finally used within the HMMs of the millenium speech recognition system.

6 Experiments and results

We conclude the paper with initial results. MFCC-vectors were extracted via HTK3.3 with a framesize of 25ms and an overlap of 10ms. Training of the SV classifiers was performed on extracted phonemes of all *si* and *sx* utterances of the TIMIT dataset, using a modified version of *svmlight* ([9]), using the RK derived in section 3, using equal weights $w_{\mathscr{H}} = w_{\mathscr{H}_T} = 0.5$. (see eq. (1)) within the transform specific kernel. For computing the probabilities, the training set is split using 70% for training of the SVMs and the remaining 30% to serve as data for the estimation of the parameters a,b of the sigmoid functions. *Leave-one-out* cross validation is applied to reduce the biasses originating from the training data. A first evaluation was performed on a subset of the TIMIT phonemes that offers both similar and distinct sounds. Table 1 shows the results that motivate advancing to continuous speech recognition.

	SVM 70	Prob 70	SVM 100	SVM 3frame MFCC	GMM 16 MFCC
aa	90.03	88.80	90.40	69.26	65.89
ae	84.17	85.17	86.08	59.79	49.81
ay	85.11	87.46	87.60	52.22	49.94
eh	79.43	77.59	82.69	44.22	37.46
ey	81.47	83.02	85.37	56.39	48.61
ih	74.19	74.85	78.38	37.82	30.00
ix	69.31	70.55	73.38	47.63	37.59
iy	93.92	94.79	95.68	77.44	71.50
n	95.87	96.31	96.28	88.63	82.42
s	95.52	94.11	95.66	88.43	70.35
z	52.97	63.44	63.35	42.50	59.93
avg.	82.00	83.28	84.95	60.39	54.94

Table 1 - **Recognition rates**. The first column represents *SME*-feature based SVM-classification results averaged over the results for the individual states. Column two gives an impression of the outcome using the multiclass posterior probabilities described in sections 4.2 and applying Bayes' Rule. Columns three and four depict the recognition rates presented in [5]: SVM-classification using standard MFCC vectors over three frames (for reasons of fair comparison) without Δ, $\Delta\Delta$ and classification rates using GMMs with 16 gaussian mixtures and diagonal covariance matrix estimated on ordinary 39-dimensional MFCC-vectors (13 bins plus Δ and $\Delta\Delta$).

In ([5]) we started with pure phoneme recognition where silence was not an (by HTK) extractable sound . As a consequence, all silence parts were skipped and the speech recognition system is trained without silence model. Currently, we are conducting initial experiments with the millenium speechrecognition system. Phoneme representation is reduced to monophones one in contrast to stat-of-the-art triphone models, which will be addressed in a second phase. Also at this point, accoustic/languag model weighting and beam optimization need to be considered. We expect profitable results from our hybrid approach.

Finally, big thanks to Friedrich Faubel for his vast help with the implementation of the Millenium-specific parts.

References

[1] J. Agler and K. McCarthy. *Pick Interpolation and Hilbert Functions Spaces*, volume Vol. 44 of *Graduate Studies in Mathematics*. American Mathematical Society, 2002.

[2] Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden markov support vector machines. *20th International Conference on Machine Learning (ICML)*, 2003.

[3] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, Vol. 68:337–404, 1950.

[4] K.P. Bennet and O.L. Magasarian. Robust linear programming discriminaion of two linearly inseparable sets. *Optimization Methods and Software*, Vol. 1:22 – 34, 1992.

[5] A. Beschorner and D. Klakow. Correlation features and a linear transform specific reproducing kernel. *13th International Conference on Text, Speech and Dialogue (TSD)*, 2010.

[6] W. M. Campbell. A sequence kernel and its application to speaker recognition. *Neural Information Processing Systems*, Vol. 14:1157 – 1163, 2001.

[7] Shantanu Chakrabartty and Gert Cauwenberghs. Forward-decoding kernel-based phone recognition. *Advances in Neural Information Processing Systems (NIPS)*, pages 1165–1172, 2002.

[8] C. Cortes and V. Vapnik. Support vector machines. *Machine Learning*, Vol. 20:273–297, 1995.

[9] T. Joachims. Making large-scale svm learning practical. *Advances in Kernel Methods - Support Vector Learning*, 1999.

[10] S.E. Krüger, M. Schafföner, M. Katz, E. Andelic, and A. Wendemuth. Speech recognition with support vector machines in a hybrid system. In *Proc. EuroSpeech, 2005*, pages 993–996, 2005.

[11] S.E. Krüger, M. Schafföner, M. Katz, E. Andelic, and A. Wendemuth. Mixture of support vector machines for hmm based speech recognition. In *In 18th International Conference on Pattern Recognition (ICPR)*, 2006.

[12] K-F. Lee and H-W. Hon. Speaker-independent phone recognition using hidden markiv models. *IEEE Transactions on Accoustic Speech and Signal Processing*, 37 No. 11, 1989.

[13] H-T. Lin, C-J. Lin, and R.C. Weng. A note on platt's paper on probabilistic outputs for support vector machines. *Technical Report, Department of Computer Science*, 2003.

[14] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 1999.

[15] E. Rodríguez et al. Speech/speaker recognition using a hmm/gmm hybrid model. *Lecture Notes in Computer Science*, Vol. 1206:227 – 234, 1997.

[16] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

[17] J. Shawe-Taylor and N. Christianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[18] A. Shigoe. *Support Vector Machines for Pattern Classification*. Advances in Pattern Recognition. Springer, 2005.

[19] N.D. Smith and M. J. F. Gales. Using svms and discriminative models for speech recognition. *IEEE International conference on accoustic speech and signal processing*, Vol. 1:77 – 80, 2002.

[20] Matthias Wölfel and John McDonough. *Distant Speech Recognition*. Wiley and Sons, Ltd., 1. edition, 2009.

[21] T-F. Wu, C-J. Lin, and R.C. Weng. Probability estimates for multiclass classification by pairwise coding. *Journal of Machine Learning Research*, Vol. 5, 2004.