ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp.7 -12

Recognition of Consonant-Vowel Utterances using Support Vector Machines

C.Chandra Sekhar, K.Takeda and F.Itakura
CIAIR, Itakura Laboratory, Dept. of Information Electronics,
Nagoya University, Nagoya-4648603, Japan.
email: chandra@itakura.nuee.nagoya-u.ac.jp

Abstract. In conventional approaches for multi-class pattern recognition using Support Vector Machines (SVMs), each class is discriminated against all the other classes to build an SVM for that class. We propose a *close-class-set* discrimination method suitable for large class set pattern recognition problems. An SVM is built for each of the 145 Consonant-Vowel (CV) classes by discriminating that class against only a small number (about 15) of classes close to it phonetically. The method leads to about 17% reduction in the average number of support vectors per class with a decrease of only 4.4% in the recognition accuracy.

1. Introduction

Support vector machines (SVMs) have been shown to give a good generalization performance in solving pattern recognition problems [1]. The main idea of a support vector machine for pattern classification is to construct a hyperplane as the decision surface in such a way that the margin of separation between positive and negative examples of a class is maximized [2]. In the conventional *one-against-the-rest* method for multi-class pattern recognition using SVMs [3], one SVM is built for each of the M classes. The SVM for a class is built by using the training examples of that class as positive examples and the training examples of the remaining $(M-1)$ classes as negative examples. When the number of classes is large, each class may be close to only a subset of classes. A *close-class-set* may be identified for each class based on a description of the classes. It is expected that the data points belonging to a particular class and the classes in its *close-class-set* fall in adjacent regions in the pattern space. The data points of the other classes are expected to be far from the region of the class under consideration, and may not be important in construction of the decision surface for that class. In such a case, the decision surface may be constructed by discriminating a class against the classes in its *close-class-set* only. We demonstrate the effectiveness of the *close-class-set* discrimination method for recognition of isolated utterances belonging to a large number of Consonant-Vowel (CV) classes.

In Section 2, we explain a method for identification of *close-class-sets* for CV classes. In Section 3, we discuss how a decision surface is constructed by a support vector machine. In Section 4, we present the studies on recognition of CV utterances.

2. Close-Class-Set Identification for CV classes

The list of 145 Consonant-Vowel (CV) classes corresponding to combinations of 29 consonants and 5 vowels in Indian languages is given in Table 1.

Table 1: List of Consonant-Vowel (CV) classes.

Description of Consonant			Vowel				
Phonetic Category	Place of Articulation	Manner of Articulation	/a/	/i/	/u/	/e/	/o/
Stops	Velar	UV_UA	ka	ki	ku	ke	ko
		UV_A	kha	khi	khu	khe	kho
		V_UA	ga	gi	gu	ge	go
		V_A	gha	ghi	ghu	ghe	gho
	Alveolar	UV_UA	ṭa	ṭi	ṭu	ṭe	ṭo
		UV_A	ṭha	ṭhi	ṭhu	ṭhe	ṭho
		V_UA	ḍa	ḍi	ḍu	ḍe	ḍo
		V_A	ḍha	ḍhi	ḍhu	ḍhe	ḍho
	Dental	UV_UA	ta	ti	tu	te	to
		UV_A	tha	thi	thu	the	tho
		V_UA	da	di	du	de	do
		V_A	dha	dhi	dhu	dhe	dho
	Bilabial	UV_UA	pa	pi	pu	pe	po
		UV_A	pha	phi	phu	phe	pho
		V_UA	ba	bi	bu	be	bo
		V_A	bha	bhi	bhu	bhe	bho
Affricates	Palatal	UV_UA	ca	ci	cu	ce	co
		UV_A	cha	chi	chu	che	cho
		V_UA	ja	ji	ju	je	jo
		V_A	jha	jhi	jhu	jhe	jho
Nasals	Dental		na	ni	nu	ne	no
	Bilabial		ma	mi	mu	me	mo
Semivowels	Palatal	Glide	ya	yi	yu	ye	yo
	Alveolar	Trill	ra	ri	ru	re	ro
	Dental	Lateral	la	li	lu	le	lo
	Dental	Glide	va	vi	vu	ve	vo
Fricatives	Dental	Unvoiced	sa	si	su	se	so
	Alveolar	Unvoiced	ṣa	ṣi	ṣu	ṣe	ṣo
	Glottal	Unvoiced	ha	hi	hu	he	ho

The consonants belong to the following phonetic categories: Stops, Affricates, Nasals. Semivowels and Fricatives. Stop consonants are described using their Manner of Articulation (MOA) and Place of Articulation (POA). The four manners of articulation are: Unvoiced Unaspirated (UV_UA), Unvoiced Aspirated (UV_A), Voiced Unaspirated (V_UA) and Voiced Aspirated (V_A). The four places of articulation are: Velar, Alveolar, Dental and Bilabial.

The phonetic description is used to identify the *close-class-sets* for each CV class. The *close-class-sets* for the class /ga/ are given in Table 2. The MOA of the stop consonant /g/ is voiced unaspirated and the POA is velar. Each of the 56 classes in the *close-class-set*$_1$ has at least one feature (MOA, POA or vowel) in common with the class /ga/. All the other 88 classes do not have any feature in common with /ga/. The 18 classes in the *close-class-set*$_2$ may be considered closer to /ga/ than the remaining 126 classes. This set includes 10 Stop Consonant-Vowel classes each of which has two features in common with /ga/. The other 8 classes correspond to combinations of vowel /a/ with sonorant consonants (belonging to the other phonetic categories) that have acoustic similarities with the consonant /g/. In a similar way, the *close-class-sets* for each CV class can be identified. The sizes of the *close-class-sets* vary for different categories of consonants, and are given in Table 3.

Table 2: List of CV classes close to the class /ga/.

Close-class-set	No. of classes	Classes in the *close-class-set*
Close-class-set$_1$	56	ka kha gha ki khi ghi ku khu ghu ke khe ghe ko kho gho ḍa da ba ḍha dha bha ṭha tha pha ṭa ta pa gi gu ge go ḍi ḍu ḍe ḍo di du de do bi bu be bo ca cha ja jha na ma ya ra la va ha sa ṣa
Close-class-set$_2$	18	ka kha gha ḍa da ba gi gu ge go ja na ma ya ra la va ha

Table 3: Sizes of the *close-class-sets* for different categories of consonants.

Phonetic category	*Close-class-set*$_1$	*Close-class-set*$_2$
Stops	56	18
Affricates	44	14
Nasals	36	15
Semivowels	44	15
Fricatives	40	16

3. Construction of Decision Surface by SVM

In a support vector machine for pattern classification [2], the input pattern vector $\mathbf{x}$ is nonlinearly mapped into a feature space to obtain the corresponding feature vector $\mathbf{z}$. Then an optimal hyperplane is constructed in the feature space in such a way that the margin of separation between positive and negative examples of a class is maximized. The discriminant function of a hyperplane with a weight vector $\mathbf{w}$ and bias b is given by

$$\mathbf{w}^T\mathbf{z} + b \tag{1}$$

The optimum value of margin of separation is $\frac{2}{|\mathbf{w}_o|}$. Maximizing the margin of separation is equivalent to minimizing the Euclidean norm of $\mathbf{w}$.

For patterns that are not linearly separable in the feature space, some data points may fall inside the region of separation or on the wrong side of separation. The nonnegative slack variable β_i denoting a measure of the deviation of a training example $\mathbf{z}_i$ from the ideal condition of pattern separability is introduced in the cost function. The SVM learning problem is to find the values of $\mathbf{w}$ and b such that they satisfy the constraint:

$$d_i(\mathbf{w}^T\mathbf{z}_i + b) \geq 1 - \beta_i \quad for \quad i = 1, 2, ..., N \tag{2}$$

and such that $\mathbf{w}$ and β_i minimize the cost function:

$$\Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\beta_i \tag{3}$$

Here $\mathbf{z}_i$ is the feature vector of ith training example $\mathbf{x}_i$ and d_i is the corresponding desired output. C is a user-specified positive parameter. This constrained optimization problem is solved using the method of Lagrange multipliers. The training examples for which the optimum Lagrange multipliers take nonzero values are called the *support vectors*. For the optimal values of $\mathbf{w}$ and b, the support vectors satisfy the constraint in (2) with the equality sign. For optimum Lagrange multipliers $\alpha_{o,i}$, the optimum weight vector $\mathbf{w}_o$ is given by

$$\mathbf{w}_o = \sum_{i=1}^{N_s}\alpha_{o,i}d_i\mathbf{z}_i \tag{4}$$

where N_s is the number of support vectors. The discriminant function of the optimal decision surface constructed by the SVM is then defined in terms of support vectors as follows:

$$\sum_{i=1}^{N_s}\alpha_{o,i}d_iK(\mathbf{x}, \mathbf{x}_i) + b_o \tag{5}$$

where the inner-product kernel $K(\mathbf{x}, \mathbf{x}_i)$ is defined as:

$$K(\mathbf{x}, \mathbf{x}_i) = \mathbf{z}^T\mathbf{z}_i \tag{6}$$

Therefore the architecture of an SVM consists of an input layer, a hidden layer of N_s nodes and an output node. The pattern vector $\mathbf{x}$ is given as input to each hidden node to compute the corresponding kernel function. The output of the SVM is then computed using (5). For multi-class pattern recognition, an SVM is built for each of the M classes and a test pattern is recognized by assigning the class of the SVM with the maximum output value. The SVM for each class can be built using the negative examples from $(M-1)$ classes or from the classes close to it only. In the next section, we study the performance of SVMs built using these methods for CV recognition.

4. Studies on CV recognition

For each of the 145 CV classes, 12 repetitions of isolated utterances are collected from each of eight male speakers. Out of these, eight for each class and speaker are used for training. The remaining four utterances for each class and speaker are used as test data. Data is collected in a laboratory environment using a sampling frequency of 16 kHz. A fixed size pattern is extracted from the speech data of a CV utterance using the following method: The point at which the consonant ends and the vowel begins in a CV utterance is defined as the Vowel Onset Point (VOP). The VOP for a CV utterance is automatically detected using the method given in [4]. We consider 60 msec of data before and 140 msec of data after the VOP for analysis. The 200 msec segment of speech data around the VOP contains most of the information necessary for CV recognition. This segment is analyzed frame by frame, with a frame duration of 20 msec and a frame shift of 5 msec. Each frame of data is represented using 12 weighted cepstral coefficients derived from eight linear prediction coefficients. The cepstral coefficient vectors of adjacent frames are averaged. Thus a CV utterance is represented by a pattern consisting of 20x12=240 parameters.

A multilayer perceptron model has been trained for CV recognition so that its performance can be compared with that of support vector machines. The model consist of 240 nodes in the input layer, 140 nodes in the first hidden layer, 75 nodes in the second hidden layer and 145 nodes in the output layer. The *SVMTorch* software [5] has been used to build the support vector machines for CV recognition. In the conventional *one-against-the-rest* method to build the SVMs, the examples of 144 classes are used as negative examples for each class. In the *close-class-set* discrimination method, the SVMs are built for each of the 145 classes separately using different *close-class-sets*. Three nonlinear kernel functions considered in our studies are as follows:

$$\begin{array}{ll} \text{Sigmoidal kernel:} & K(\mathbf{x},\mathbf{x}_i) = tanh(0.001\mathbf{x}^T\mathbf{x}_i - 1.0) \\ \text{Polynomial kernel:} & K(\mathbf{x},\mathbf{x}_i) = (\mathbf{x}^T\mathbf{x}_i + 1)^2 \\ \text{Gaussian kernel:} & K(\mathbf{x},\mathbf{x}_i) = exp(\frac{-|\mathbf{x}-\mathbf{x}_i|^2}{0.01}) \end{array}$$

Table 4 gives the performance of CV recognition systems trained using the *one-against-rest-method.* The Gaussian kernel SVM gives an accuracy of about 80% that is significantly better than that of the multilayer perceptron.

Table 4: Classification accuracy of CV recognition systems on test data set of 4640 patterns.

Classification model	Accuracy (in %)
Multilayer perceptron	58.38
Sigmoidal kernel SVM	51.32
Polynomial kernel SVM	60.82
Gaussian kernel SVM	80.46

Table 5 shows the classification accuracy and the average number of support vectors per class for Gaussian kernel SVMs built using different discrimination sets of classes. When the *close-class-set*$_1$ is used to build the SVM for each class, the accuracy decreases by 0.9% only. When the *close-class-set*$_2$ is used, the accuracy of Gaussian kernel SVMs decreases by 4.4% only. These results show that by using only a small number (14 to 18) of classes close to a CV class, it is possible to construct a decision surface for it and achieve a good accuracy. There is about 17% decrease in the average number of support vectors per class when the *close-class-set*$_2$ is used for discrimination. This result demonstrates the effectiveness of the *close-class-set* discrimination method in reducing the complexity of support vector machines for large class set pattern recognition problems.

Table 5: Classification accuracy (in %) and the average number of support vectors (N_s) per class for Gaussian kernel SVMs built using different discrimination sets.

	Discrimination method used to build the SVMs		
	One-against-the-rest	*Close-class-set*$_1$	*Close-class-set*$_2$
Accuracy	80.46	79.54	76.05
Average N_s	329	311	272

5. Summary and Conclusions

In this paper, we proposed a *close-class-set* discrimination method suitable for solving large class set pattern recognition problems using support vector machines. The effectiveness of the method has been demonstrated for recognition of isolated utterances belonging to 145 Consonant-Vowel classes. Approaches for detection of acoustic events [4] and the proposed method for classification of CV units using support vector machines can be combined for recognition of subword units in continuous speech.

References

[1] C.J.C.Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[2] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, 1999.

[3] J.Weston and C.Watkins, “Multi-class support vector machines,” Tech. Rep. CSD-TR-98-04, Royal Holloway, University of London, May 1998.

[4] J.Y.Siva Rama Krishna Rao, C.Chandra Sekhar, and B.Yegnanarayana, “Neural networks based approach for detection of vowel onset points,” in *International Conference on Advances in Pattern Recognition and Digital Techniques*, Calcutta, Dec. 1999, pp. 316–320.

[5] R.Collobert and S.Bengio, “Support vector machines for large-scale regression problems,” Tech. Rep. IDIAP-RR-00-17, IDIAP, Switzerland, 2000.