Restricted domain speech classification using automatic transcription and SVMs

Arthur Kantor
Mark Hasegawa-Johnson

Introduction

This paper considers utterance classification with a very large number of classes (138 classes) in a limited domain (physical activity diary). Methods common in call routing applications (e.g., [3,4,5,6]) are applied to this large-search-space system, with good results: our system first transcribes speech with a commercial speech recognition package, and then uses an SVM classifier to classify the transcriptions. With sufficient training examples, the errors in the transcription do not affect the classification accuracy. High accuracy of the classifier may be related to the relatively small vocabulary used by subjects: although subjects were not instructed to limit their vocabulary in any way, in practice, a vocabulary of only 500 keywords was sufficient to accurately distinguish among the 138 physical activity classes.

While general automatic speech transcription and natural language understanding are difficult tasks, we demonstrate that natural speech restricted to a particular topic can be classified into a large number of classes with high accuracy.

Motivation and problem description

This system was built to classify and analyze physical activity as part of a kinesiology experiment. In this experiment, each subject recorded the type of physical activity he or she has been performing for the past 30 minutes, at 30 minute intervals. The record was a short sentence spoken in unconstrained English, and recorded on digital media. The physical activity was then classified into one of 612 possible activity classes by human experts. Each activity class has associated with it an energy expenditure rate, called a Metabolic Equivalent Task-hour, or a MET score. The goal for our system was to automatically classify the utterances so as to minimize (1) the classification error, and (2) the mean-squared MET score error.

Even though 612 classes were possible, all labels actually assigned to any training token by any labeler came from a 138-class subset of the possible labels. Those 138 activity classes have a lot of semantic overlap, so frequently two different human experts would assign the same utterance to two different classes, but the MET score for these classes would be similar.

We can treat the MET score as weighted error, which can be used to denote the similarity between misclassified examples.

Methods

Each sentence was treated as an independent example, and was transcribed manually and automatically, using the commercial speech recognition software Dragon Naturally Speaking (DNS) v7. The DNS was first adapted to the voice of each talker. The word accuracy rate was around 90%. In 68% of the utterances, manual transcriptions matched the automatic transcriptions (sentence error rate = 32%).

The transcriptions were then used as input for the classification stage. The features for each example were obtained by treating each transcription as a bag of words giving us a binary vector which specified a subset of the English dictionary. On our feature vectors we have trained 138 binary one-vs.-all SVM [2] classifiers, one for each class. We used the standard linear kernel (RBF kernels were found not to give any accuracy improvement). We have also experimented with the error correcting output codes [1], and even though we could achieve the same accuracy with a smaller number of binary classifiers, each classifier took much longer to train, so that it was more computationally efficient to simply use the one-vs.-all multi-class scheme.

Results:

We have tested the classification accuracy with a 0-1 loss function and a mean squared MET score error.

Error rate on small dataset:

		Test (1001 examples)	
		Manual Transcription	**Automatic Transcription**
Train (4628 examples)	**Manual Transcription**	29% error rate (.67 METS error)	38% error rate (.60 METS error)
	Automatic Transcription	29% error rate (.86 METS error)	36% error rate (.80 METS error)

Error rate on large dataset:

		Test (1102 examples)	
		Manual Transcription	**Automatic Transcription**
Train (13746 examples)	**Manual Transcription**	20% error rate (.40 METS error)	24% error rate (.37 METS error)
	Automatic Transcription	19% error rate (.31 METS error)	21% error rate (.22 METS error)

With one-vs.-all classification it is possible for an example to be classified into more than one class, or into none at all, and such cases did happen with our data. If an example was claimed by multiple classes, we chose the class that claimed it most strongly, i.e. its classification margin was most positive.

About half of the classification errors are due to the system refusing to classify the example into any class (reject utterances). Reject utterances were not used in METS error calculation. In our application, reject utterances are preferable to misclassification errors, since these ‘difficult’ examples can be passed on to a human for manual classification, and can then be used to improve the accuracy of the automatic classifier.

Using the small dataset, manually transcribed test data is always classified better than automatically transcribed data, regardless of the type of data used for training. As the number of training examples increases, not only does the error rate drop, but the error rate for the automatically transcribed speech approaches the error rate of the manually transcribed speech. Using the large dataset, a classifier trained on manual transcriptions is still unable to classify automatically transcribed test data, but a classifier trained on automatic transcriptions produces statistically identical classification error rates, regardless of whether the test transcriptions are manually or automatically generated. The METS error rate of the best classifier is slightly better using automatically transcribed test data, but the difference is not statistically significant ($p>0.05$).

One of the surprising outcomes of this experiment was that even though an open vocabulary was allowed, only 1640 unique words were used in the entire dataset. When we trimmed this small number of keywords to 500 keywords by removing the 1140 keywords having least mutual information with the classes, there was no noticeable drop in classification accuracy.

Future work:

To determine the possible upper limit of recognition accuracy, it would be interesting to test the agreement between humans performing this classification task. Also, it may be possible to build a high accuracy keyword-spotting speech recognizer for 500 keywords, and to tune the keyword-spotter for use in a portable or telephone-based electronic physical activity diary.

References:

[1] Rennie and Rifkin. "Improving Multiclass Text Classification with the Support Vector Machine." A.I. Memo #2001-026, C.B.C.L. Memo #210, October 2001.
[2] Christopher J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition (1998)
[3] Min Tang, Bryan Pellom, and Kadri Hacioglu, “Call-Type Classification and Unsupervised Training for the Call Center Domain,” in IEEE Workshop on ASRU, 2003.
[4] Stephen Cox, “Discriminative Techniques in Call Routing,” in Proc. ICASSP 2003.
[5] Lee Begeja, Bernard Renger, David Gibbon, Zhu Liu, and Behzad Shahraray, “Interactive Machine Learning Techniques for Improving SLU Models,” in HLT/NAACL 2004.
[6] Sheila Garfield and Stefan Wermter, “Comparing Support Vector Machines, Recurrent Networks and Finite State Transducers for Classifying Spoken Utterances,” in Proc. ICANN 2003.