Feature Selection for Speaker Traits

Jouni Pohjalainen[1], *Serdar Kadioglu*[2], *Okko Räsänen*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
[2]Department of Computer Science, Brown University, Providence, RI 02912, USA
jpohjala@acoustics.hut.fi, serdark@cs.brown.edu, okko.rasanen@aalto.fi

Abstract

This study focuses on handling high-dimensional classification problems by means of feature selection. The data sets used are provided by the organizers of the Interspeech 2012 Speaker Trait Challenge. A combination of two feature selection approaches gives results that approach or exceed the challenge baselines using a k-nearest-neighbor classifier. One of the feature selection methods is based on covering the data set with correct unsupervised or supervised classifications according to individual features. The other selection method applies a measure of statistical dependence between discretized features and class labels.

Index Terms: pattern recognition, feature selection, high-dimensional data, speaker characteristics

1. Introduction

The Interspeech 2012 Speaker Trait Challenge provides 6125 utterance-level features (functionals of low-level acoustic descriptors) which can be used in binary speech classification tasks [1]. In each of the seven subtasks, the goal is to identify the presence or absence of speaker traits [1]: openness, conscientiousness, extraversion, agreeableness, neuroticism, likability and intelligibility.

Pattern classification with such high-dimensional data is complicated by the phenomenon referred to as the curse of dimensionality [2] [3]. Large amounts of training data are required to avoid overlearning in a high-dimensional, sparsely populated feature space.

The diversity of the available utterance-level acoustic features suggests that there may exist low-dimensional subspaces in which the classes are separable using simple pattern recognition methods. In order to test this assumption, the approach chosen in the present study is to treat the problem as that of feature selection. The main practical problem is the number of possible feature subsets which is $2^{6125} - 1$. Because this number is intractable, informed feature selection methods are required.

We present two differently based approaches for feature selection, analyze their performance individually and together and compare the obtained results to the challenge baselines.

2. Methods

2.1. Feature selection based on classification

The first approach to feature selection is based on supervised or unsupervised discrimination between two classes using Gaussian mixture models (GMMs) to model the probability distributions of individual features. A feature set is constructed based on the classifications.

In the task of classifying speaker trait T as present or not present, unidimensional GMMs λ_{T} and λ_{NT} are initially trained in the normal fashion, using class-specific data of a feature variable, with five iterations of expectation-maximization (EM) re-estimation for GMMs [4]. Before initial training, the mixture weights of the GMMs are initialized by uniform distributions. The variance parameters of each component are initialized by 0.1 times the global variance of the feature. The initial mean parameters of each component are given by the heuristic approach described in [5]. In applying the GMMs to unseen data in a conventional supervised manner, the class decision for each observation is based on the logarithmic likelihood ratio $L = L_{\mathrm{T}} - L_{\mathrm{NT}}$, where L_{T} and L_{NT} are the logarithmic likelihoods of the observation having been produced by each GMM. The decision threshold is adjusted on the training data set according to the equal error rate (EER) criterion typically used in detection applications. It corresponds to equal misclassification rate for both classes.

The mixture-based class discrimination for individual features is also evaluated using unsupervised learning after the initial supervised training. In this method, the parameters of two J-component GMMs are joined to form one composite GMM with $2J$ components while multiplying the component weight parameters by 0.5. EM iteration is now applied to training the composite GMM, with one modification: the sums of the J weight parameters belonging to the T and NT classes are both normalized back to 0.5 before each expectation step (E step) of the EM iteration, thus ensuring that the prior probabilities of the two classes stay equal in the E step. Otherwise, the GMM parameters are allowed to freely adapt to the complete data set which includes observations from both classes. After five iterations, the sub-GMMs belonging to the two classes are again separated and used to classify

the data using a decision threshold that has been set according to the EER criterion in the training phase. Given that EM is a hill-climbing method guaranteed to converge on at least a local, if not global, maximum of the likelihood function [4], this method favors features for which the solution that discriminates between the two classes is close in likelihood to a local maximum or saddle point of a natural clustering solution.

Classifications are obtained using both of the above methods, i.e. by supervised and unsupervised learning. For both cases, we construct matrices where rows correspond to audio clips, columns correspond to features, and the value is 1 if the clip in question was correctly classified by the feature in question and 0 otherwise.

We make the following observation. Our goal to select a subset of features can be formulated as an Integer Linear Programming (ILP) problem. In particular, we map our problem to the well-known Set Covering Problem (SCP). In the SCP problem, we are given a finite set $S := \{1, ..., m\}$ of items, and a family $F = \{S_1, ..., S_n \subseteq S\}$ of subsets of S, and a cost function $c : F \rightarrow R^+$. The objective is to find a subset $C \subseteq F$ such that $\sum_{S_i \in C} c(S_i)$ is minimized. In our formulation, the family of sets corresponds to individual features and the set of items are audio clips. For each feature we attribute a cost of one, that is, we are dealing with the so-called unicost SCP. Finally, our objective is to *cover* all clips using minimum number of features.

As explained above, before solving this ILP, each feature has been evaluated on each clip, and it has been noted whether the clip can be classified correctly using that feature. This yields knowledge about which items (clips) are covered in which sets (features). Then, our formulation can be written as:

$$\begin{array}{lll} \text{Minimize} & \sum_{s \in F} x & \\ & \sum_{s \in F : e \in s} x_s \geq 1 & \forall\, e \in S \\ & x_s \in \{0, 1\} & \forall\, s \in F \end{array} \quad (1)$$

The decision variables x_s denote whether the feature s is selected while the first constraint ensures that we consider every clip. Our goal is to select the minimum number of such features. This observation has an immediate bearing on our problem. We can leverage general techniques for solving ILP's; namely using a branch-and-bound algorithm based on the linear relaxation of the original problem where the integer decision variables $x_s \in \{0, 1\}$ are replaced with $x_s \in [0, 1]$. While this algorithm solves the problem to optimality, in general, the SCP is NP-hard [6].

In case optimal solutions cannot be computed efficiently, it is possible to trade optimality with efficiency. One alternative is to use a simple and fast greedy algorithm to obtain an approximate solution. The greedy algorithm selects a set that covers the most items using the least cost at each step until all items are covered. In fact, this algorithm achieves an H_n factor approximation algorithm for the minimum set cover problem, where $H_n = 1 + 1/2 + ... + 1/n$. Another approach would be to solve the linear relaxation to optimality, and then to use the rounding up technique to obtain an integral solution [7]. In this work, we consider the rounding-up technique.

Depending on whether the SCP was based on classifications using GMMs trained in a completely supervised or a partially unsupervised manner, the feature selection method is termed supervised or unsupervised classification set covering problem (SSCP or USCP, respectively).

2.2. Feature selection based on statistical dependence

In this feature selection method, each feature is discretized by quantizing it to one of $N = 65$ levels, where the quantization scale is adjusted such that each bin will contain an equal amount of samples. The statistical dependence between the discretized feature y and the class labeling z is evaluated according to the formula

$$D = \sum_{y \in Y} \sum_{z \in Z} p(y, z) \frac{p(y, z)}{p(y)p(z)}. \quad (2)$$

We have found this formulation preferable to the conventional mutual information (MI) measure in assessing statistical dependence in problems like the present one. A previously specified number of features having the highest D value will be selected. This feature selection method is termed the statistical dependence method (SD).

2.3. Classification

We apply k nearest neighbors (kNN) as the classification method [2]. Prior to kNN classification, each feature is normalized to have zero mean and unit variance. While making a decision on an input vector based on its k nearest neighbors according to the Euclidean distance, the counts of different classes are scaled by dividing them by the frequencies of occurrence of the same classes in the training data.

2.4. Evaluation

Evaluation of classification results is carried out according to the guidelines of the Interspeech 2012 Speaker Trait Challenge [1]. In particular, the evaluation is based on three data sets, namely the training, development and evaluation sets. Moreover, the unweighted average recall (UA) is used as the primary measure of performance. Another measure is the weighted average recall (WA).

3. Results

3.1. Dimensionalities

In order to determine the proper size of the feature set (dimensionality of the feature space), SSCP and USCP feature selection were first performed. The size of these

feature sets ranged from 273 to 423. Each possible size of subset of the initial feature set was analyzed. Seven random subsets of each size were generated and evaluated in classification ($k = 40$) according to the UA measure. The results were averaged across the seven random sets. In addition, the SD feature selection method was used to obtain a subset of each size and classified with optimal $k \in \{5, 10, \ldots, 150\}$. Both sequences were smoothed using a 3-tap moving average filter and then averaged together for each subset size. The number of features giving the highest such score was chosen for each case where SCP- and SD-based feature selection methods were used in succession. In the evaluation on development sets, SD by itself used the same dimensionality as the better of SSCP+SD and USCP+SD configured as above.

3.2. Values of k

The value of k, i.e., the number of nearest neighbors based on which the class decisions are made, was chosen in training of each task so as to maximize development set classification performance such that $k \in \{5, 10, \ldots, 150\}$. For the joint training by training and development set data in order to classify the test data, these values were scaled in proportion of the increased training data size.

3.3. Likability sub-challenge

The results are shown in Table 1. Combined with the SD method, both the SSCP and the USCP methods outperform the development set baseline, and the SSCP method combined with the SD method outperforms the test set baseline. In order to better respond to classification of the test sets of each subtask, the SSCP and USCP have been cross-optimized between the training and development sets (classification of the development set is trained on the training set and vice versa). Whether this has a noticeable effect on the development set scores is unknown. The SD and MI methods only analyze the training data set.

Table 1: *Results obtained for the likability sub-challenge. The WA scores are shown for the test set only.*

Data	Method of feature selection	Number of features	k	UA	WA
Devel	Baseline (SVM)	6125		58.5	
Devel	SSCP	406	60	57.4	
Devel	USCP	423	35	59.8	
Devel	SD	349	65	55.3	
Devel	SSCP+SD	349	90	62.0	
Devel	USCP+SD	411	40	60.0	
Devel	SSCP+MI	349	90	58.6	
Test	Baseline (RF)	6125		59.0	59.2
Test	USCP+SD	411	58	53.3	53.9
Test	SSCP+SD	349	131	**61.3**	**61.3**

3.4. Pathology sub-challenge

As shown in Table 2, several methods exceed the development set baseline. In terms of the UA measure, the test set baseline is not exceeded, but in terms of WA, the proposed methods exceed the test set baseline score.

Table 2: *Results obtained for the pathology sub-challenge. WA scores are shown for the test set only.*

Data	Method of feature selection	Number of features	k	UA	WA
Devel	Baseline (RF)	6125		65.1	
Devel	SSCP	318	115	67.2	
Devel	USCP	347	85	61.8	
Devel	SD	300	30	66.2	
Devel	SSCP+SD	299	75	68.3	
Devel	USCP+SD	172	125	63.8	
Test	Baseline (RF)	6125		68.9	67.5
Test	USCP	347	155	65.5	**72.4**
Test	USCP+SD	300	137	65.6	**72.4**
Test	SSCP+SD	299	137	66.3	69.8

3.5. Personality sub-challenge

The results are shown in Table 3. Each development set baseline is outperformed by at least one of the proposed feature selection methods using dimensionality selection (Section 3.1). The best of these methods were used for the test set. In addition, for the test set, the USCP was used followed by the SD method using a threshold of 1.15 on the statistic of Eq. 2, leading to differently sized feature sets. Comparison of, e.g., the two versions of USCP+SD in the "C" task shows potential effect of misinformative features. The test set baseline of the openness task is exceeded by the proposed methods. USCP+SD comes close to the test set baseline of the conscientiousness task. The test set baseline of the neuroticism task is approximately reached by the SSCP+SD method and the WA score of the baseline method is exceeded.

4. Conclusions

This study focused on problems of binary classification of speaker characteristics. Given a large number of initial features but a limited amount of labeled data, the problems were approached from the perspective of feature subset selection. The underlying assumption was that among the vast amount of feature spaces obtainable by selecting a subset of the available 6125 features, there would exist feature spaces in which the class separation is both easy (relative to the general difficulty of the problem at hand) and generalizable. The challenge was to find some of such feature spaces using feature selection approaches that would not overemphasize the performance of particular algorithms on particular datasets, i.e. overlearn the training data.

Two different approaches to feature selection, working on different assumptions, were combined: one based on covering all of the data in terms of correct classifications using Gaussian mixtures, and another based on selecting features whose quantized versions displayed the highest statistical dependence with the labeling. By combining the two different approaches, we were able to discover feature spaces in which high classification accuracy, comparable to state-of-the-art classification methods [1], was obtained using a conceptually simple nearest-neighbor classifier. Moreover, the promising classification performance using the discovered feature sets carried over from the development sets to the test sets, meaning that the feature selection results using the proposed approach are generalizable. The results of the feature selection can conceivably be used in other, similar paralinguistic analysis tasks. The features selected by the methods are available online [8].

Outside the context of the present challenge, we consider high-dimensional pattern recognition problems, similar to the present ones, in which a large and comprehensive set of features is readily available. Instead of focusing on classification methods, the results suggest an alternative approach. Our results demonstrate the potential of tackling these problems as feature selection problems as long as care is taken to avoid overlearning the training data, e.g., by combining different feature selection objectives. Future research directions include further development of the feature selection methodology as well as the application of the proposed feature selection methods to new problems in pattern recognition.

Table 3: *Results obtained for the personality sub-challenge. WA scores are shown for the test set only.*

	Data	Method of feature selection	Number of features	k	UA	WA
O	Devel	Baseline (SVM)	6125		60.4	
	Devel	SSCP	297	50	67.6	
	Devel	USCP	314	125	65.4	
	Devel	SD	277	135	66.8	
	Devel	SSCP+SD	277	50	72.3	
	Devel	USCP+SD	93	105	67.6	
	Test	Baseline (RF)	6125		59.0	63.7
	Test	USCP+SD	149	135	60.4	57.7
	Test	SSCP+SD	277	86	**62.3**	58.2
C	Devel	Baseline (RF)	6125		74.9	
	Devel	SSCP	287	35	71.9	
	Devel	USCP	293	110	74.4	
	Devel	SD	285	10	73.5	
	Devel	SSCP+SD	239	40	74.8	
	Devel	USCP+SD	285	40	75.8	
	Test	Baseline (SVM)	6125		80.1	80.1
	Test	USCP+SD	176	70	79.7	79.6
	Test	USCP+SD	285	69	76.8	76.6
E	Devel	Baseline (RF)	6125		82.8	
	Devel	SSCP	273	30	76.9	
	Devel	USCP	291	70	85.8	
	Devel	SD	291	15	81.4	
	Devel	SSCP+SD	75	130	81.9	
	Devel	USCP+SD	291	70	85.8	
	Test	Baseline (SVM)	6125		76.2	76.6
	Test	USCP+SD	197	15	71.9	72.6
	Test	USCP+SD	291	120	71.6	72.6
A	Devel	Baseline (SVM)	6125		67.6	
	Devel	SSCP	322	50	65.3	
	Devel	USCP	337	50	66.4	
	Devel	SD	217	15	66.1	
	Devel	SSCP+SD	217	15	72.1	
	Devel	USCP+SD	246	15	69.2	
	Test	Baseline (RF)	6125		64.2	64.2
	Test	USCP+SD	199	50	55.9	56.2
	Test	SSCP+SD	217	26	59.7	60.2
N	Devel	Baseline (RF)	6125		68.9	
	Devel	SSCP	312	10	67.9	
	Devel	USCP	320	5	67.9	
	Devel	SD	124	15	69.1	
	Devel	SSCP+SD	124	40	74.8	
	Devel	USCP+SD	149	35	73.0	
	Test	Baseline (SVM)	6125		65.9	65.7
	Test	USCP+SD	200	25	62.4	63.2
	Test	SSCP+SD	124	69	65.3	**67.6**

5. Acknowledgements

This work was supported by Academy of Finland (127345).

6. References

[1] Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G. and Weiss, B., "The Interspeech 2012 Speaker Trait Challenge", Proc. Interspeech 2012, ISCA, Portland, OR, USA, 2012.

[2] Duda, R. O., Hart, P. E. and Stork, D. G., "Pattern Classification", 2nd ed., John Wiley and Sons Inc., 2001.

[3] Theodoridis, S. and Koutroumbas, K., "Pattern Recognition", 2nd ed., Academic Press, 2003.

[4] Xu, L. and Jordan, M. I., "On Convergence Properties of the EM Algorithm for Gaussian Mixtures", Neural Computation, 8(1):129–151, 1996.

[5] Katsavounidis, I., Kuo, C.-C. J. and Zhang, Z., "A New Initialization Technique for Generalized Lloyd Iteration", IEEE Signal Processing Letters, 1(10):144–146, 1994.

[6] Nemhauser, G. L. and Wolsey, L. A., "Integer and Combinatorial Optimization", John Wiley and Sons Inc., 1988.

[7] Vazirani, V. V., "Approximation Algorithms", Springer 2001: I-IXI, pp. 1–378, 2001.

[8] Feature selection results: http://www.acoustics.hut.fi/∼jpohjala/is2012stc/, June 22, 2012.