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, jouni.pohjalainen@aalto.fi, okko.rasanen@aalto.fi

[2] Department of Computer Science, Brown University, Providence, RI 02912, USA, serdark@cs.brown.edu

September, 10

InterSpeech 2012

Aalto University
School of Science
and Technology

Overview

- Introduction

- Feature selection methods
- Classification method
- Evaluation procedure

- Likability sub-challenge
- Pathology sub-challenge
- Personality sub-challenge

- Conclusions

Introduction

- 6125 utterance-level features were provided or each sub-challenge
- Instead of focusing on classification methods for high-dimensional data, can we find feature spaces where classification is "easy"?
- There are are $2^{6125} - 1$ possible feature spaces

Introduction

- Compared to state-of-the-art classification methods for high-dimensional data, can we reach their performance by means of just feature selection and a basic classification method?

- The results should be generalizable
  - avoid overlearning any single feature selection objective

Approach I - Classification using individual features

	Feature 1	Feature 2		…		Feature m
Audio Clip 1						
Audio Clip 2						
…						
Audio Clip n						

- We start with a classification phase

Approach I - Classification using individual features

	Feature 1	Feature 2		…		Feature m
Audio Clip 1						
Audio Clip 2						
…						
Audio Clip n						

- We start with a classification phase
- Goal: classify the speech audio clips using each individual feature

Approach I - Classification using individual features

	Feature 1	Feature 2		…		Feature m
Audio Clip 1						
Audio Clip 2						
…						
Audio Clip n						

- Supervised training
  - Train 8-component GMMs for both classes
  - Trait is present or not

Approach I - Classification using individual features

	Feature 1	Feature 2		…		Feature m
Audio Clip 1						
Audio Clip 2						
…						
Audio Clip n						

- Unsupervised training
  - After initial supervised training, combine the two GMMs into one and let it freely adapt to the training data during additional EM iterations, then again separate the two GMMs

Approach I - Classification using individual features

	Feature 1	Feature 2		…		Feature m
Audio Clip 1						
Audio Clip 2						
…						
Audio Clip n						

- Use the GMMs to classify the clips according to the Bayes rule
- Classify the development set using GMMs trained on the training set and vice versa

Approach I - Classification using individual features

	Feature 1	Feature 2		…		Feature m
Audio Clip 1	0 / 1					
Audio Clip 2						
				…		
…						
Audio Clip n						0 / 1

- for each audio clip
- for each feature
  - note whether the classification was correct or not

Feature Selection

	Feature 1	Feature 2		…		Feature m
Audio Clip 1	0 / 1					
Audio Clip 2						
				…		
…						
Audio Clip n						0 / 1

- Properties we want
  - For each clip there must be at least one correct classification
  - Number of selected features minimized

Observation

	Feature 1	Feature 2		…		Feature m
Audio Clip 1	0 / 1					
Audio Clip 2						
				…		
…						
Audio Clip n						0 / 1

- We can formulate this as the Set-Covering Problem
  - Features as sets
  - Audio clips as items

Observation

	Feature 1	Feature 2		…		Feature m
Audio Clip 1	0 / 1					
Audio Clip 2						
				…		
…						
Audio Clip n						0 / 1

- We can formulate this as the Set-Covering Problem
  - Features as sets
  - Audio clips as items
- Each set has a cost of one: unicost-SCP

Set-Covering Problem (SCP)

	Feature 1	Feature 2		…		Feature m
Audio Clip 1	0 / 1					
Audio Clip 2						
				…		
…						
Audio Clip n						0 / 1

- Goal:
  - Select features/sets/columns such that each clip/item/row is covered at least once
  - Total cost (number of selected features) is minimized

Solving the SCP

- This observation has an immediate bearing on the feature selection problem
- We can leverage general techniques for solving the SCP

Solving the SCP

- Exact solution method
  - Branch-and-bound algorithm using the linear relaxation

Solving the SCP

- Exact solution method
  - Branch-and-bound algorithm using the linear relaxation
- In general, SCP is NP-hard

Solving the SCP

- Approximate solution methods
  - Greedy algorithm: select iteratively the features that cover the most remaining clips
  - Rounding-up technique: solve the linear relaxation. Then, to get an integer solution, round-up the values of every fractional value to 1 (i.e. select these sets)

Solve the SCP

	Feature 1	Feature 2		…		Feature m
Audio Clip 1	0 / 1					
Audio Clip 2						
				…		
…						
Audio Clip n						0 / 1

- Based on supervised classification: SSCP

Solve the SCP

	Feature 1	Feature 2		…		Feature m
Audio Clip 1	0 / 1					
Audio Clip 2						
				…		
…						
Audio Clip n						0 / 1

- Based on supervised classification: SSCP
- Based on unsupervised classification: USCP

Approach II – Statistical dependence between features and labels

- Discretize each feature by quantizing it to one of N = 65 levels, where the quantization scale is adjusted s.t. each bin will contain an equal amount of samples

- Measure the statistical dependence (SD) using the formula

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

where y is the discretized feature and z is the class labeling

Approach II – Statistical dependence between features and labels

- For comparison, the conventional mutual information (MI) measure is given by

$$D = \sum_{y \in Y} \sum_{z \in Z} p(y,z) \log\left(\frac{p(y,z)}{p(y)p(z)} \right)$$

- In order to determine the number of features to be chosen according to the SD measure in the combined selection methods, subsets of each size of the SCP-based feature sets were obtained by random selection and by SD (using *training* data)

Combined feature selection methods

- First select a few hundred features using SSCP or USCP
- Use the SD or MI measure to select features among this subset

Classification method: *k* nearest neighbors (*k*NN)

- Each feature is normalized to have zero mean and unit variance
- Find *k* nearest neighbors according to the Euclidean distance measure
- Among the *k* nearest neighbors, the counts of different classes are scaled by dividing them by the frequencies of occurrence of those classes in the training data
- Value of *k* determined experimentally

Likability sub-challenge

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**

Pathology sub-challenge

Data	Method of feature selection	Number of features	k	UA	WA
Devel	Baseline (RF)	6125		*64.8*	
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**

Personality sub-challenge

	Data	Method of feature selection	Number of features	k	UA	WA
O	Devel	Baseline (SVM)	6125		*60.4*	
	Devel	SSCP+SD	277	50	**72.3**	
	Test	Baseline (RF)	6125		*59.0*	*63.7*
	Test	SSCP+SD	277	86	**62.3**	**58.2**
C	Devel	Baseline (RF)	6125		*74.9*	
	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

Personality sub-challenge

	Data	Method of feature selection	Number of features	k	UA	WA
E	Devel	Baseline (RF)	6125		*82.8*	
	Devel	USCP+SD	291	70	**85.8**	
	Test	Baseline (SVM)	6125		*76.2*	*76.6*
	Test	USCP+SD	291	120	**71.6**	**72.6**
A	Devel	Baseline (SVM)	6125		*67.6*	
	Devel	SSCP+SD	217	15	**72.1**	
	Test	Baseline (RF)	6125		*64.2*	*64.2*
	Test	SSCP+SD	217	26	**59.7**	**60.2**
N	Devel	Baseline (RF)	6125		*68.9*	
	Devel	SSCP+SD	124	40	**74.8**	
	Test	Baseline (SVM)	6125		*65.9*	*65.7*
	Test	SSCP+SD	124	69	**65.3**	**67.6**

Conclusions

- Using feature selection and simple *k*NN classification, it was possible to achieve classification accuracy comparable to state-of-the-art classification methods using the full high-dimensional feature space

- Good results were obtained by two approaches:
  - Set covering problem (SCP) based on classification using single features (different classification method)
  - Statistical dependence (SD) between discretized features and class labeling

Conclusions

- The results generalized from one dataset to another
- By combining differently based feature selection criteria (SCP and SD), better results were obtained than by using a single criterion