Signal Processing 88 (2008) 2956–2970

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition

Dimitrios Ververidis, Constantine Kotropoulos *

Artificial Intelligence and Information Analysis Laboratory, Department of Informatics, Aristotle University of Thessaloniki, Box 451, Thessaloniki 541 24, Greece

ARTICLE INFO

Article history:
Received 11 May 2007
Received in revised form
19 June 2008
Accepted 3 July 2008
Available online 5 July 2008

Keywords:
Bayes classifier
Cross-validation
Variance of the correct classification rate of the Bayes classifier
Feature selection
Wrappers

ABSTRACT

This paper addresses subset feature selection performed by the sequential floating forward selection (SFFS). The criterion employed in SFFS is the correct classification rate of the Bayes classifier assuming that the features obey the multivariate Gaussian distribution. A theoretical analysis that models the number of correctly classified utterances as a hypergeometric random variable enables the derivation of an accurate estimate of the variance of the correct classification rate during cross-validation. By employing such variance estimate, we propose a fast SFFS variant. Experimental findings on Danish emotional speech (DES) and speech under simulated and actual stress (SUSAS) databases demonstrate that SFFS computational time is reduced by 50% and the correct classification rate for classifying speech into emotional states for the selected subset of features varies less than the correct classification rate found by the standard SFFS. Although the proposed SFFS variant is tested in the framework of speech emotion recognition, the theoretical results are valid for any classifier in the context of any wrapper algorithm.

1. Introduction

Vocal emotions constitute an important constituent of multimodal human computer interaction [1,2]. Several recent surveys are devoted to the analysis and synthesis of speech emotions from the point of view of pattern recognition and machine learning as well as psychology [3–6]. The main problem in speech emotion recognition is how reliable is the correct classification rate (CCR) achieved by a classifier. This paper derives a number of propositions that govern the estimation of accurate CCRs, a topic that has not been addressed adequately, yet.

The classification of utterances into emotional states is usually accomplished by a classifier that exploits the acoustic features that are extracted from the utterances. Such a scenario is depicted in Fig. 1. Feature extraction consists of two steps, namely the extraction of acoustic feature contours and the estimation of global statistics of feature contours. The global statistics are useful in speech emotion recognition, because they are less sensitive to linguistic information. These global statistics will be called simply as features throughout the paper. One might extract tens to thousands of such features from an utterance. However, the performance of any classifier is not optimized, when all features are used.

Indeed, in such a case, the CCR, usually deteriorates. This problem is often known as ‘curse of dimensionality’, which is due to the fact that a limited set of utterances does not offer sufficient information to train a classifier with many parameters weighing the features. Therefore, the use of an algorithm that selects a subset of features is necessary. An algorithm that selects a subset of features, which optimizes the CCR, is called a *wrapper* [7].

Different feature selection strategies for wrappers have been proposed, namely *exhaustive*, *sequential*, and *random search* [8,9]. In exhaustive search, all possible combinations of features are evaluated. However, this method is

* Corresponding author. Tel./fax: +30 2310 998225.
E-mail addresses: jimver@aiia.csd.auth.gr (D. Ververidis), costas@aiia.csd.auth.gr (C. Kotropoulos).

0165-1684/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2008.07.001

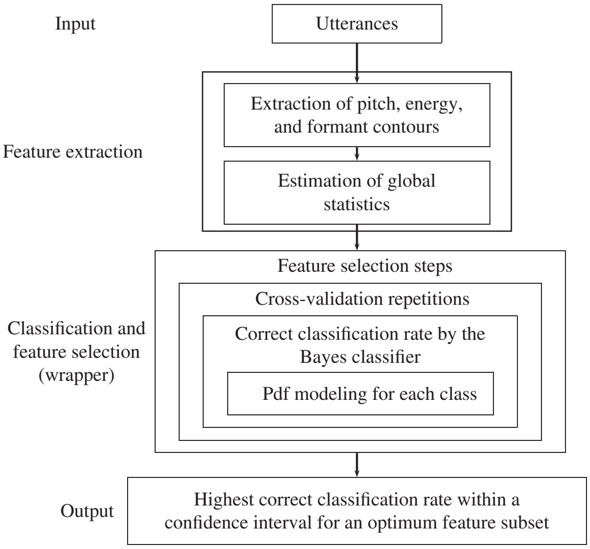
Input
Utterances
Feature extraction
Extraction of pitch, energy, and formant contours
Estimation of global statistics
Classification and feature selection (wrapper)
Feature selection steps
Cross-validation repetitions
Correct classification rate by the Bayes classifier
Pdf modeling for each class
Output
Highest correct classification rate within a confidence interval for an optimum feature subset

Fig. 1. Flowchart of the approach used for speech emotion recognition.

practically useless even for small feature sets, as the algorithm complexity is $O(2^D)$, where D is the cardinality of the complete feature set. Sequential search algorithms add or remove features one at a time. For example, either starting from an empty set they add incrementally features (*forward*) or starting from the whole set they delete one feature at a time (*backward*), or starting from a randomly chosen subset they add or delete features one at a time. Sequential algorithms are simple to implement and provide results fast, since their complexity is $O(D + (D-1) + (D-2) + \cdots + (D - D_1 + 1))$, where D_1 is the cardinality of the selected feature set. However, sequential algorithms are frequently trapped at local optima of the criterion function. Random search algorithms start from a randomly selected subset and randomly insert or delete feature sets. The use of randomness helps to escape from local optima. Nevertheless, their performance deteriorates for large feature sets [10]. Since all selection strategies, except the exhaustive search, yield local optima, they are often known as sub-optimum selection algorithms for wrappers. In the following, the term optimum will be used to maintain simplicity. One of the most promising feature selection methods for wrappers is the *sequential floating forward selection algorithm* (SFFS) [11]. The SFFS consists of a forward (insertion) step and a conditional backward (deletion) step that partially avoids the local optima of CCR. In this paper, the execution time will be reduced and the accuracy of SFFS will be improved by theoretically driven modifications of the original algorithm. The execution time is reduced by a preliminary statistical test that helps skipping features, which potentially have no discrimination information. The accuracy is improved by another statistical test, known as tenative test, that selects features that yield a statistically significant improvement of CCR.

A popular method for estimating the CCR of a classifier is the s-fold *cross-validation*. In this method, the available data-set is divided into a set used for classifier design (i.e. the training set) and a set used for testing the classifier (i.e. the test set). To focus the discussion on the application examined in this paper, the emotional states of the utterances that belong to the design set are considered known, whereas we pretend that the emotional states of the utterances of the test set are unknown. The classifier estimates the emotional state of the utterances that belong to the test set. From the comparison of the estimated with the actual (ground truth) emotional state of the test utterances, an estimate of CCR is obtained. By repeating this procedure several times, the mean CCR over repetitions is estimated and returned as the CCR estimate, that is referred to as MCCR. The parameter s in s-fold refers to the division of the available data-set into design and test sets. That is, the available data-set is divided into s roughly equal subsets, the samples of the $s-1$ subsets are used to train the classifier, and the samples of the remaining subset are used to estimate CCR during testing. The procedure is repeated for each one of the s subsets in a cyclic fashion and the average CCR over the s repetitions constitutes the MCCR [12]. Burman proposed the *repeated s-fold cross-validation* for model selection, which is simply the s-fold cross-validation repeated many times [13]. The variance of the MCCR estimated by the repeated s-fold cross-validation varies less than that measured by the s-fold cross-validation. Throughout the paper, the repeated s-fold cross-validation is simply denoted as cross-validation, since it is the only cross-validation variant studied. It will be assumed that the number of correctly classified utterances during cross-validation repetitions is a random variable that follows the hypergeometric distribution. Therefore, according to the central limit theorem (CLT), the more realizations of the random variable are obtained, the less varies the MCCR. The large number of repetitions required to obtain an MCCR with a narrow confidence interval prolongs the execution time of a wrapper. We will prove a lemma that uses the variance of the hypergeometric r.v. to find an accurate estimate of the variance of CCR without many needless cross-validation repetitions. By estimating the variance of CCR, the width of the confidence interval of CCR for a certain number of cross-validation repetitions can be predicted. By reversing the problem, if the user selects a fixed confidence interval, the number of cross-validation repetitions is obtained.

The core of the theoretical analysis is not limited to the Bayes classifier within SFFS, but it can be applied to any classifier used in the context of any wrapper. To validate the theoretical results, experiments were conducted for speech emotion recognition. However, the scope of this paper is not limited to this particular application.

The outline of this paper is as follows. In Section 2, we make a theoretical analysis that concludes with Lemma 2, which estimates the variance of the number of correctly classified utterances. Section 3 describes the Bayes classifier. In Section 4, statistical tests employing Lemma 2 are used to improve the speed and the accuracy of SFFS, when the criterion for feature selection is the CCR of the Bayes classifier. In Section 5.1, experiments are conducted in order to demonstrate the benefits of the proposed