Memorias del XVI Congreso Latinoamericano
de Control Automático, CLCA 2014
Octubre 14-17, 2014. Cancún, Quintana Roo, México

Order dependent one-vs-all tree based binary classification scheme for multiclass automatic speech emotion recognition

Rodríguez C. Santiago*, Bastidas O. Manuela*, Quintero M, O. Lucía*

**Departamento de ciencias básicas, Universidad Eafit*
Medellín, Colombia (e-mail:mbastida,srodrig1,oquinte1@eafit.edu.co)

Abstract: Automatic classification of emotional speech is a recent a booming field of research. There are two usual approaches to this problem, the first one being a multi-class classifier and the second one the use of binarization techniques. In this work we propose a binarization technique based on seven one-vs-all classifiers each corresponding to one of the seven basic emotions: anger, anxiety, happiness, sadness, boredom, disgust and neutral. These classifiers are organized in a tree hierarchical structure which is highly order dependent. The Mahalanobis distance and the Chernoff bound are proposed as possible order indicators for the structure. All the orders are tested using the scheme and decision trees as classifiers and compared to the multiclass equivalents and will show the best order for the scheme with a misclassification rate of 34% compared to the 36% of the multi-class equivalent, the distances did not prove to be a good order indicator.

Keywords: Automatic recognition, Classification, Decision trees, Pattern recognition, Speech analysis.

1. INTRODUCTION

Automatic pattern recognition is a scientific discipline that aims to classify certain objects in a preset number of categories or classes. (Theodoridis & Koutroumbas, 2003) There are lots of areas where pattern recognition can be applied, we focus in the automatic speech emotion recognition field.

Understanding the emotional state of a speaker is a great step towards a natural interaction between man and machine. A recent research field concerns automatic speech emotion recognition, which is defined as extracting the emotional state of a speaker from his or her speech (Moataz El Ayadi, 2010).In this work we will present a cascade (tree) based one-vs-all multi-class classifier for automatic speech emtion recognition and the order in which the classification is made. This is an order dependent scheme and so several orders are tested, including the ones given by distance measures between classes.

We will first introduce the field of automatic speech emotion recognition and the general background for it. Then in section 3 we will introduce the data used which includes its generation and the features that were extracted. Sections 4 and 5 presents the methodolody employed in the current work including the classifiers, the distances between classes as possible order indicators and the classification scheme. Section 6 presents the results for the scheme for different sequencial orders for the binary classifiers and the discussion of these results. Finally section 7 will present the conclussions that were made.

2. BACKGROUND

Generally automatic speech emotion recognition systems have three important aspects: The data base which is used for the system's training, the features extracted from the audio signal and the automatic classification system.

The design of automatic classification systems are articulated through the steps mentioned before, these three steps in more detail are:

1. The data bases: in these how natural the emotions are and the quality of the data is really important (Moataz El Ayadi, 2010). There are three broad kinds of data bases; the first consisting of acted emotions, where it is easier to differentiate one emotion from another, but in this case the data is less realistic due to its acted nature. The second kind where emotions are inducted via an experiment; these ones are natural and close to reality but not all emotions can be inducted and there isn't complete certainty regarding which emotion was present in the speech signal. Finally the ones where emotional speech is recorded from a natural setting (e.g. a call center), in this case the emotion is completely natural and closer to reality. The latter are the ones that present a greater challenge due to nature of the audio signal which doesn't come from a controlled environment and due to the fact that the emotion present in the speech signal is not always clear. The most widely used data bases are acted. (Björn Schuller, 2011).

2. The features: They are measurements extracted from the speech signal which will somehow contain emotional content; these will be the ones that characterize the emotions and help differentiate them. It is possible to extract these

characteristics both from small window frames or big ones, both are valid approaches and the choice depends mostly in the application. Usual features include pitch, formats, energy, timing, voice quality, spectral, etc. (Moataz El Ayadi, 2010)

3. The classifier: it is highly dependent on a well-defined and representative feature set. Several classifiers have been used for the task of automatic speech emotion recognition, among them there are: support vector machines (SVM), hidden markov models (HMM), Gaussian mixture models (GMM), artificial neural networks (ANN), k-NN, etc. (Angkoon Phinyomark, 2009) (Stavros Ntalampiras, 2012) (Björn Schuller, 2011) (Chul Min Lee, 2005) (Moataz El Ayadi, 2010) (Raul Fernandez, 2005) (Theodoridis & Koutroumbas, 2008).

The accuracy of classifiers in the automatic speech emotion recognition task is in average between 51.19% and 70 for ANN and between 74% and 81.9% for the other ones (HMM, GMM, SVM), (Moataz El Ayadi, 2010), also it is important to note that for speaker-independent speech emotion recognition systems the accuracy is less than 80% in most of the mentioned techniques, but for speaker-dependent classification, the recognition accuracy can exceeded 90% (Moataz El Ayadi, 2010).

There are two approaches regarding what classification scheme to use, one is to use a multiclass classifier where the output is the emotion present in the speech segment given the input features. Another one is to construct several binary classifiers that are simpler and combined will yield the present emotion. Our approach is the latter, we attempt to exploit the relationships between classes as was proposed in (Rifkin & Klautau, 2004) using the scheme presented in Section 5, it is similar to the one proposed by (Platt, et al., 2000).

3. DATA

3.1 Data generation

The data used was composed of segments from the Berlin Database of emotional speech (Technical University, s.f.) which were sticked together so each signal contained all seven basic emotions: anger, boredom, disgust, anxiety, happiness, sadness and neutral. Each signal uses a random recording for each emotional state.

In the Fig. 1 one of the speech signals constructed as described before is showed, each segment represents an emotion, they are ordered this way: Anger *(Ang)*, Anxiety *(Anx)*, Happiness *(Hap)*, Sadness *(Sad)*, Boredom *(Bor),* Disgust *(Dis)*, Neutral *(Neu)*.

Also the basic information of the signal in Fig. 1 is:

Number of samples: 34121

Duration of each emotion: 2.1326 secs

Total duration: 14.9279 secs

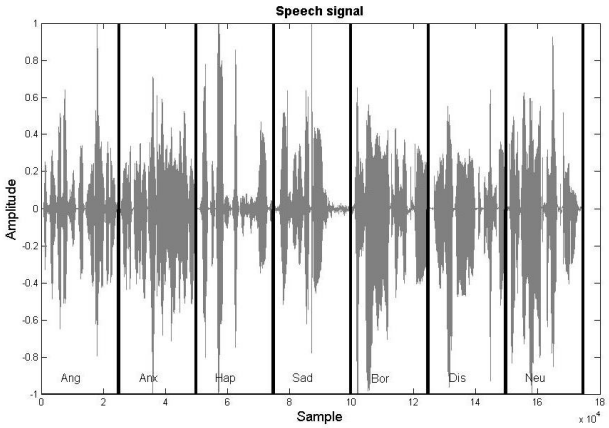

Fig. 1. One of the speech signals with labels corresponding to each emotional speech segment.

3.2 Feature extraction

Table 1. shows the features which were extracted from each signal frame, they broadly include entropy based features (1,2,3,6), frequency based features (4,5,6), energy related features (8,9), dynamic features (7,10,11) and paralinguistic (12). These features are used in emotional speech recognition (Moataz El Ayadi, 2010), in biological signal processing (Angkoon Phinyomark, 2009) and in speech recognition (Rufiner, et al., 2004).

Table 1. Features extracted from each signal window.

1	Shannon entropy
2	q-entropy
3	Log energy entropy
4	Maximum PSD
5	Frequency in kHz with a higher PSD
6	Fourier entropy
7	Order 2 AR parameters
8	Root mean squeare (RMS)
9	Integrated Absolute Value
10	Waveform length
11	Zero crossing rate
12	Pitch

Each feature was extracted from a rectangular window of 256 samples, which corresponds to 16 milliseconds with an overlap between windows of 224 samples or 14 milliseconds.

4. CLASSIFIERS AND DISTANCES BETWEEN CLASSES

In this section the Linear Discriminant Analysis (LDA) and the classification tree will be introduced, then the concepts of Chernoff bound and Mahalanobis distance which were taken as possible good order for the automatic classification scheme presented in the last part of this section, these are good candidates for the classification scheme's order since they measure distance between classes given their features.

Using the data from the features and knowing to which of the seven basic emotions corresponds each set of 12 features corresponds; the scheme will be evaluated against two multiclass classifiers, which are part of the Matlab® environment. The idea is to compare our scheme of binary classifiers to the multiclass equivalents.

4.1 Classifiers

4.1.1 Linear discriminant analysis

The LDA is a statistical technique which allows us to see the difference between certain groups with respect to their features. This technique was born with Fisher (1936) and in general consists of reducing the number of features to new variables which are a combination of the original ones; they are expressed as a discriminant function which is capable of dividing all the data in the desired number of classes (Seco, 1992), in our case this corresponds to the seven basic emotions. It is worth noting that an LDA is a linear classifiear and as so it's expected no to have the best results, this is fine since our aim is not to have a great performance of a classifier but to stablish and test the proposed scheme.

4.1.2 Classification tree

Classification trees and regression trees predict an output given a certain feature set; in particular, decision trees are systems where all the possible classifications are evaluated and one is accepted (Theodoridis & Koutroumbas, 2003).

Trees work in a sequential manner and are used to give an output given some inputs that pass through sequence of decisions or nodes, going from the root (beginning) to the leafs (the last levels). Decision trees have an advantage when dealing with multi-class problems. Each step consists of deciding in which direction to keep going given the value of a feature, this step is vital for the classifier as it depends on how a given feature can discern between classes. When a tree is trained it is this decision in each node which is decided upon. It's important to note that at the end of a branch a final rule will determine which class was decided by that given path. (Theodoridis & Koutroumbas, 2003)

4.2 Distances

Since this work will evaluate an order dependent scheme, it is important to introduce two distances which might give insight into the appropriate order. These two measures are the Mahalanobis distance and the Chernoff bound; they measure the differentiation between classes given a set of features. The hypothesis is that given the emotion most different from the other one, feature wise, then this should either be the first one or the last one to be classified and so on.

4.2.1 Chernoff bound (Bhattacharya distance)

The Chernoff bound is defined from probability theory as the upper limit of the classification error of a Bayesian classifier for two classes and with assumptions that are not in the scope of this work (see (Theodoridis & Koutroumbas, 2003)). This bound generates what is called the Bhattacharya distance which is defined as:

Let p, q be two classes with means μ_p μ_q and standard deviation Σ_p and Σ_q then:

$$d_{p,q} = \frac{1}{8}(\mu_p - \mu_q)'\left(\frac{\Sigma_p + \Sigma_q}{2}\right)^{-1}(\mu_p - \mu_q) + \frac{1}{2}\ln\frac{\frac{\Sigma_p + \Sigma_q}{2}}{\sqrt{|\Sigma_p \| \Sigma_q|}} \quad (1)$$

This gives us an approximation to the distance between two classes with different means and standard deviations; we will use it as means to define the classes which will be more easily classified as their distances are greater.

4.2.2 Mahalanobis distance

As a direct consequence of the Bhattacharya distance when the classes have the same standard deviation we have the Mahalanobis distance, it is also used as a measure of separation between classes and is defined as (Theodoridis & Koutroumbas, 2003):

Let p, q be two classes with means μ_p, μ_q and standard deviation $\Sigma_p = \Sigma_q$ then the Mahalanobis distance is:

$$d_{p,q} = \frac{1}{8}(\mu_p - \mu_q)'(\Sigma_p)^{-1}(\mu_p - \mu_q) \quad (2)$$

This also gives us an approximation to the distance between two classes and it will be used as means to define the classes which will be more easily.

5. CLASSIFICATION SCHEME

It is easier to distinguish between two classes than it is to do between more. Thus when presented with a multiclass problem, like automatic speech emotion recognition, we opted to deal with the problem into several binary classification tasks.

Usual approaches to multiclass binary classification include (Galar, et al., 2011):

- One-vs.-one (OVO): A binary classifier for each of the possible combination of pairs of classes, then the outputs of all the classifiers is aggregated for the final decision.

- One-vs.-all (OVA): A binary classifier for each class, where the output is a soft classification between the class and all the other possible ones.

Usually the decision is done through some kind of aggregation on all the outputs of the classifiers (Galar, et al., 2011). A scheme presented in (Platt, et al., 2000) was based on a graph structure and consisted of OVO classifiers. Our Approach is to have a tree-like scheme where seven one-vs-all classifiers determine the emotion of the speech segment; is the structure is presented in Fig 2 We take as input a set of features, in our case the ones described in Section 3.2, these are then classified by the first OVA classifier which decides if the frame corresponds to the emotion of that classifier or

one of the other ones. If it is one of the other ones then the process is repeated by the second classifier and so on.

It is clear from the way the scheme works that the order in which the binary one-vs-all classifiers for each emotion are set has a big impact. This order is one of the main things that will be evaluated.

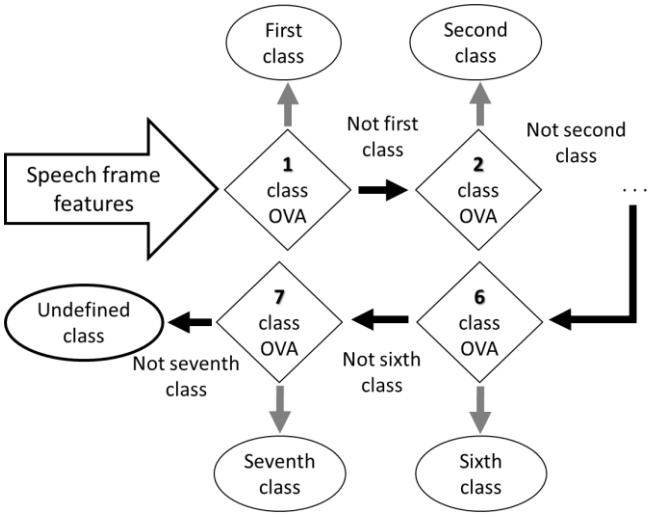

Fig 2. Classification Scheme where seven one-vs-all classifiers determine the emotion of the speech segment.

6. RESULTS AND DISCUSSION

Tests for the scheme were conducted using twenty signals as described in Section 3; this led to a total of 122451 sets of 12 features each. Errors presented in this Section correspond to average misclassification percentage of each of the classes; that is the percentage of sets that were not classified correctly, either globally or per class

The first, two multi-class classifiers were trained using a random sample of 70% of the 122451 sets and validated with the remaining 30%, Table 2. presents the results for this, it will serve as comparison for the proposed scheme.

Table 2. Results for multi-class LDA and tree classifiers.

LDA	
Class	**Error**
Ang	0,950
Anx	0,501
Hap	0,640
Sad	0,909
Bor	0,547
Dis	0,438
Neu	0,722
Mean	
0,672	

Tree	
Class	**Error**
Ang	0,315
Anx	0,484
Hap	0,395
Sad	0,195
Bor	0,433
Dis	0,316
Neu	0,387
Mean	
0,361	

Between the multi-class classifiers the Tree gives a better result as was expected. It is interesting to note that while anxiety and boredom were the ones with a higher error in the Tree, in the LDA they were among the lowest. The opposite is also true for anger and sadness.

The two distances that were introduced in Section 4.2 present possible orders for the scheme, this are either from most distant to least or the opposite, they are shown in Table 3 and Table 4. These distances are thought to provide insight regarding the scheme's order because they measure proximity or separation between classes, and this could be a criterion for the order. The distances don't present the same order; they only coincide in the position of sadness and anger, this indicates that the classes don't have the same standard deviation as the Mahalanobis distance supposes.

Table 3. Mahalanobis order of class from most distant class to least.

Order	*Sad*	*Neu*	*Ang*	*Anx*	*Hap*	*Dis*	*Bor*
Inverse	*Bor*	*Dis*	*Hap*	*Anx*	*Ang*	*Neu*	*Sad*

Table 4. Chernoff order of class from most distant class to least.

Order	*Sad*	*Dis*	*Ang*	*Bad*	*Neu*	*Anx*	*Hap*
Inverse	*Hap*	*Anx*	*Neu*	*Bor*	*Ang*	*Dis*	*Sad*

Tests for the proposed scheme using both LDA and Tree based classification were made, only the Tree results will be presented since the LDA yielded unsatisfactory classification rates, not even comparable to the multi-class case. Table 5 shows the best results for the orders given by the distances, in both cases the order given by the distance was better than the inverse one and thus it is the only one shown. It is clear that the error is worse than the one given by the multi-class tree.

Table 5. Best error results with the scheme using a Tree for the Mahalanobis distance and the Chernoff bound.

Order	**Error**
Mahalanobis	0.446
Chernoff	0.445

Table 6. Best order found for the scheme of all the possible orders with the mean error of classifications.

Order	**Error**
1:*Neu*	0,343
2:*Anx*	
3:*Bor*	
4:*Dis*	
5:*Ang*	
6:*Hap*	
7:*Sad*	

Finally all the possible orders, a total of 5040 combinations were tested to find the best one and see if it was the same as one of the distances. Table 6. shows the best order found for the scheme and the error it yielded.

The results of the scheme compared the multi-class equivalent did not yield a significant improvement in classification rate, other advantages might come from using different feature sets for each binary classifier or combining classifiers for each step. Several improvements can be made in hybridization with other methods or classifiers.

From the results of different orders it is clear that the scheme is order dependent, there is a misclassification of around 45% with the proposed distances and of 34% with the best possible order, 11% is a considerable gap.

7. CONCLUSSIONS

The proposed scheme can yield results comparable to the literature in the field of automatic speech emotion recognition and is a new approach with room for improvement. Its flexibility and easy hybridization are an advantage as with other OVA or OVO structures. It would be interesting to use other classifiers such as perceptrons to test the scheme against the multi-class equivalent while preserving the simplicity of the classifier.

The orders given by the Mahalanobis distance and the Chernoff bound were not the best ones and they show that the distance between classes is not the appropriate criterion for the choice of order.

REFERENCES

Angkoon Phinyomark, C. L. P. P., 2009. A Novel Feature Extraction for Robust EMG Pattern Recognition.

Björn Schuller, A. B. S. S. D. S., 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge.

Chul Min Lee, S. S. N., 2005. Toward Detecting Emotions in Spoken Dialogs.

Galar, M, 2011. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition,* p. 1761–1776.

Moataz El Ayadi, M. S. K. F. K., 2010. Survey on speech emotion recognition: Features, classification schemes, and databases.

Raul Fernandez, R. W. P., 2005. Classical and Novel Discriminant Features for Affect Recognition from Speech.

Rifkin, R. & Klautau, A., 2004. In defense of One Vs All clasification. *Jounal of Machine learning research* .

Seco, G. V., 1992. *Técnicas multivariadas aplicadas a las ciencias del comportamiento.* Universidad de Oviedo: s.n.

Stavros Ntalampiras, N. F., 2012. Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition.

Technical University, I. f. S. a. C. D. o. C. S. B., s.f. *Berlin Database of Emotional Speech.* [En línea].

Theodoridis, S. & Koutroumbas, K., 2003. *Pattern recognition.* s.l.:Elsevier.