INTERSPEECH 2009 BRIGHTON

Emotion Recognition Using a Hierarchical Binary Decision Tree Approach

Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory (SAIL)
Electrical Engineering Department
University of Southern California, Los Angeles, CA 90089, USA
[chiclee | mower | busso | sungbokl]@usc.edu, shri@sipi.usc.edu

Abstract

Emotion state tracking is an important aspect of human-computer and human-robot interaction. It is important to design task specific emotion recognition systems for real-world applications. In this work, we propose a hierarchical structure loosely motivated by Appraisal Theory for emotion recognition. The levels in the hierarchical structure are carefully designed to place the *easier* classification task at the top level and delay the decision between highly ambiguous classes to the end. The proposed structure maps an input utterance into one of the five-emotion classes through subsequent layers of binary classifications. We obtain a balanced recall on each of the individual emotion classes using this hierarchical structure. The performance measure of the average unweighted recall percentage on the evaluation data set improves by 3.3% absolute (8.8% relative) over the baseline model.

1. Introduction

User state identification is of the utmost importance in both human-computer and human-robot interaction. A synthetic agent that is unable to recognize, track, and affect the emotional state of its human user may be unable to foster long-term interactions [1]. Emotion expressions do not exist in isolation. They are a function of mood, personality, the environment, and the task at hand [2]. Therefore, task specific emotion models should be created to reduce the inherent affective variability. For example, there are increasing numbers of interactive educational systems commercially available and under development [3, 4]. These systems must be able to accurately identify a child's emotional state to foster long-term interactions and positive evaluations [5, 6, 7]. If a machine cannot detect anger or frustration in a child, such an affective state may continue or increase, resulting in a premature end to an interaction. Likewise, gaging how certain a child is while engaged in problem solving and learning can help scaffold the interaction in a context appropriate way [8]. Furthermore, since the emotion labels of interest differ depending on task, it is important to construct emotion models that take task-specific information into account.

In this paper we present an emotion recognition study using affective speech collected from fifty-one children interacting with an AIBO dog [9]. The five emotion classes of interest are: *angry*, *emphateic*, *neutral*, *positive*, *rest*. This problem is challenging for three main reasons. Firstly, the interactions are natural; the children are not prompted with specific emotion targets. As a result, the emotion expressions are more subtle when compared with those of acted speech. Secondly, there are a large number of children in the database. This results in a high variability in the speech and emotional expressiveness patterns. It should be noted that the greater inherent variability in the speech and spoken interaction patterns of children, compared to those of adults, pose challenges to the automated processing of those data [10]. Finally, the database is heavily biased by one emotion class, *neutral*. When one emotional class is overrepresented in the training data, it may be difficult to form multi-class emotion recognition models based on conventional techniques.

An interactive machine must be able to identify a user's state from a set of emotions. If the machine is optimized on the measure of conventional *accuracy* (number of accurately classified samples by total number of tested samples), it will likely recognize only a few of the dominant states accurately. Therefore, the resulting interaction paradigm will be skewed towards that set of emotions, leading to repeated misclassification. Unweighted recall provides a method for assessing the performance of a classifier in emotionally biased datasets. The performance of the recognition system should be measured and optimized using average unweighted recall over five emotion classes since the emotion class distribution of the AIBO database is highly skewed towards the *neutral* class.

Our framework is motivated by the Appraisal Theory [11] of emotions. Appraisal theory states that emotion perception is a multi-stage conscious and unconscious process. At each stage an individual appraises the situation, reacts, and reappraises. Our presented framework, a hierarchical binary decision tree, shares the same notion of this appraisal and reappraisal process.

We present a hierarchical structure that splits the five-emotion class problem in a series of binary decision classifications. In our framework we first classify between emotional groupings that are easily distinguished, instead of using conventional *emotional* classes vs. *non-emotional* classes as the first step processing. We leave the more ambiguous emotion classes to latter steps. This approach is empirically beneficial as it helps us to alleviate error accumulation by splitting a multi-class problem into a series of binary decisions. We achieve an average unweighted recall of 48.37% using leave-one speaker out (26-fold) cross validation on the training dataset. We have also obtained 41.57% of unweighted recall on the evaluation dataset, which is 3.3% absolute and 8.8% relative over the best baseline results presented in [9].

The paper is organized as follows. Our research methodology is described in Section 2. The experimental results and discussion are presented in Section 3. Conclusion and future work are given in Section 4.

2. Methodology and Approach

2.1. Classifier Framework

We formulate the design of our classification framework based on the following two main ideas. Our goal is to optimize the

Copyright © 2009 ISCA

performance metric of unweighted recall percentage.

- We use a combination of binary classifiers instead of a single multi-class classifier
- We propose a classification framework composed of a hierarchical tree, where the top level classification is performed on the *easiest* emotion recognition task

The framework is shown in the Figure 1. The main idea behind the proposed classification scheme is to split the five-class problem into a set of two-class problems. We start with the relatively easy classification task at the top level and leave the harder tasks for the end. The proposed approach can propagate fewer classification errors down the tree when compared to the conventional intuitive approach of classifying *non-emotional* classes vs. *emotional* classes as the first step then splitting the broad *emotional* class further to identify specific emotion classes of interest. Also by splitting the five-class problem into a set of two-class problems, we can obtain a more balanced recall percentage for each emotion classes.

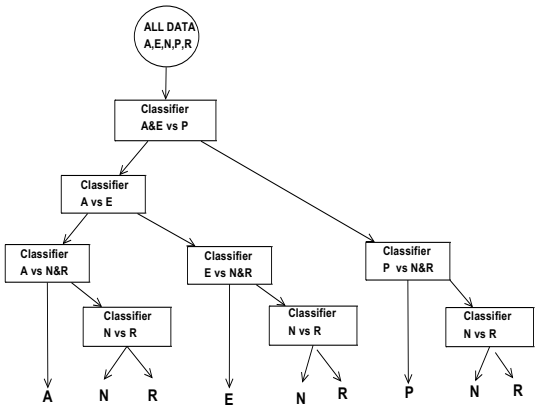

Figure 1: Proposed Classification Framework: A Hierarchical Binary Decision Tree. The emotion classes are A: Angry, E: Emphatic, P: Positive, N: Neutral, R: Rest.

Each classifier box shown in Figure 1 is a binary classifier. At each level, the hard output label of the test sample is fed into the next level of classifier to perform another set of binary classifications. This sequence of binary classifications allows us to take advantage of the variability inherent in the data by creating initial classifications with high recall and identifying classification tasks with a high levels of discriminability. The order of the classification is important in this framework. We want to ensure a maximum separation between any two chosen classes at each level. We propose the sequence of classification as shown in Figure 1 based on a combination of empirical results, which will be discussed in the later section, and prior knowledge regarding the nature of the emotion classes.

The classes considered in this task are: **A**ngry, **E**mphatic, **P**ositive, **N**eutral, and **R**est. We can consider the classes A, E, and P as the *emotional* classes and the classes N and R as the *non-emotional* classes. We placed A/E vs. P at the first classification stage because empirical testing indicated that audio features allowed us to accurately discriminate between these two groups of classes. We delay the decision between N and R until the end based on the empirical observation regarding the high level of similarity and ambiguity between N and R. We trained a total of six classifiers listed as follows (the classifiers were trained using all the data from the training set with class labels relevant to the task):

- Angry/Emphatic vs. Positive (A&E vs. P)
- Angry vs. Emphatic (A vs. E)
- Angry vs. Neutral/Rest (A vs. N&R)
- Emphatic vs. Neutral/Rest (E vs. N&R)
- Positive vs. Neutral/Rest (P vs. N&R)
- Neutral vs. Rest (N vs. R)

2.2. Classifier Type

We propose two classification schemes. The first uses Bayesian Logistic Regression, the second, Support Vector Machine (SVM) classification. We compare the performance of our system using these two classification schemes in each of the classification boxes shown in Figure 1 described in the previous section. The Bayesian Logistic Regression [12] and the Support Vector Machine [13] have both shown to be effective in classification tasks.

Single class bias is a problem in this emotion recognition task as it may bias the results towards the over-represented class, in this case - *neutral*. Prior work has shown the effectiveness of using Synthetic Minority Oversampling Technique (SMOTE) [14] in dealing with over-representation of a single class. However, in this paper, instead of generating artificial data samples to balance classes, we exploit our known knowledge about the class distribution of the AIBO database to adjust the decision threshold on both Bayesian Logistic Regression and Support Vector Machine to obtain a balanced recall accuracy across five emotion classes.

2.2.1. Bayesian Logistic Regression

A general binary logistic regression model is a discriminative model of the form shown in Equation 1.

$$p(y = 1|\beta, x) = \psi(B^T x) \tag{1}$$

where y is the class label (+1,−1), x is the input feature vector, β's are the model parameters, and ψ is the logistic function defined in Equation 2

$$\psi(z) = \frac{\exp(z)}{1 + \exp(z)} \tag{2}$$

In the Bayesian Logistic Regression, we place a Gaussian prior with $\mu = 0$ and covariance $\sigma^2 I$ on the model parameters β's shown in Equation 3 and perform a *maximum a posteriori estimation* of the model parameters to prevent overfitting of the parameters on the training data. This has the same effect as the ridge logistic regression where the model parameters's $||L_2||$ is constrained.

$$p(\beta_j|\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{\beta_j^2}{2\sigma^2}) \tag{3}$$

The baseline decision rule for binary classification is that whenever $p(y = 1|\beta, x) \geq 0.5$, we assign the class label as 1. However, since we aim to obtain a more balanced recall percentage across all five emotion classes, the threshold on the probability can be tuned to ensure a balanced error rate can be achieved between any two group of emotion classes in a binary classification. The BBR software [12] was used for Bayesian Logistic Regression model training and threshold tuning.

2.2.2. Support Vector Machine

Support Vector Machine (SVM) seeks a classification rule for maximum-margin separation between two classes (+1,−1). Because of the imbalance of data, the cost of error C, which is defined as the cost of misclassifying positive samples can be incorporated and specified roughly corresponding to the priors of the class samples. By including this bias turn, we can obtain a more balanced recall percentage of the five emotion classes. The optimization problem including the cost of error C is set up as shown in Equation 4 below.

$$\begin{aligned} &\text{minimize} : ||\vec{w}||^2 + C_+ \sum_{i:y_i=1} \xi_i + C_- \sum_{j:y_j=-1} \xi_j \\ &\text{subject to} : \forall k : y_k[\vec{w}^T \vec{x} + b] \geq 1 - \xi_k \end{aligned} \tag{4}$$

The training algorithm is implemented using *SVM-Light* [15] with linear kernel, and C was specified according to the class prior on the whole training dataset.

2.3. Feature Normalization & Selection

Features are *z-normalized* with respect to the neutral utterances in the training dataset. We assume that the average characteristics of neutral utterances across the 51 children do not vary extensively. The testing examples are *z-normalized* with μ and σ^2 of neutral utterances from the training data. The normalization allows us to mitigate the problem of speaker-specific emotional patterns. We perform feature selection on the 384 features provided with the AIBO database described in [9] using the statistics software SPSS. We perform feature selections for each of the six classifiers listed in the previous section. A total of six sets of features were selected to use with each of the specified classifiers. We used binary logistic regression with forward selection as feature selection. Our stopping criterion was based on the conditional likelihood of the model.

3. Experimental Results and Discussion

Two different datasets were used in this work, a training dataset and an unlabeled evaluation dataset. We developed and tested our algorithm using Experiment I with training dataset described below. During the development phase, the leave one speaker out (26-fold) cross-validation was used to estimate the classification performance measured by the average unweighted recall. The method was used to simulate the scenario in which the unlabeled evaluation dataset consists of a disjoints speaker sets.

- **Experiment I** : Leave one speaker out (26-fold) cross-validation on the training dataset
- **Experiment II** : Evaluate performance on the unlabeled evaluation dataset

3.1. Result of Experiment I

The unweighted recall for Bayesian Logistic Regression was 48.27%. The unweighted recall for Support Vector Machines was 47.44%. Please see Table 1 for a summary of the results. The columns of the confusion matrix represent our hypothesized class labels and the rows of the matrix are the ground truth class labels. Several observations can be made from examining the result. First, if we look at the recall for A/E vs. P at the first step, they are at 94.82% and 92.28% (Bayesian Logistic Regression) respectively. Both of the classifiers are able to identify these two emotion groups with high accuracy. Therefore, we are able to retain the majority of the members of the two groups of emotion classes by placing this classification task as the first step in the proposed structure. Since our structure outputs hard labels at every stage, the error is likely to accumulate and propagate down the tree. Therefore, it is crucial to maintain a high level of recall starting from the beginning of the structure to prevent further degradation on the performance measure - unweighted recall.

Table 1: ***Experiment I:*** *Summary of Result*

Bayesian Logistic Regression (BLG)					
Unweighted Recall (UA)			Weighted Recall (WA)		
48.27%			**48.82%**		
	Angry	Emphatic	Neutral	Positive	Rest
Angry	**504**	145	126	53	53
Emphatic	395	**1078**	412	101	107
Neutral	506	1020	**2703**	776	585
Positive	21	31	121	**439**	62
Rest	97	130	185	171	**138**
Support Vector Machine (SVM)					
Unweighted Recall (UA)			Weighted Recall (WA)		
47.44%			**46.84%**		
	Angry	Emphatic	Neutral	Positive	Rest
Angry	**463**	159	123	57	79
Emphatic	322	**1041**	424	156	150
Neutral	386	930	**2548**	958	768
Positive	27	29	103	**446**	69
Rest	80	123	159	192	**167**

We are classifying the emotion class, *rest*, at about the chance level. This is expected because this class is not as strictly defined as the emotions in the other four classes. *Rest* is classified more often as either *neutral* or *positive* compared with *angry* or *emphatic* (Table 1). We hypothesize that this unequal confusion occurs because *positive* and *neutral* are more similar to *rest* because the three emotion categories have similar levels of activation (an attribute measuring the calmness or excitation of an emotion).

We obtain a fairly comparable recall for all four of the emotion classes, except for the class, *rest*. This indicates that the structure of our framework is able to handle the highly skewed database to obtain a more balanced retrieval rate. This is essential in emotion recognition where in natural human interaction, *neutral* is more likely be the majority of expressed emotions. In order to identify several other less frequently expressed but informative emotional classes, the balancing of the recognition accuracy using the proposed structure can be advantageous.

3.2. Result of Experiment II

In Experiment II, we evaluated our classification framework on the unlabeled evaluation dataset. The six classifiers were trained on the whole training dataset. The unweighted recall for Bayesian Logistic Regression was 41.57%. The unweighted recall for Support Vector Machines was 40.84%. The summary of the results is shown in Table 2.

Our proposed framework using Bayesian Logistic Regression achieved the highest average unweighted recall. It improves the accuracy measure of the baseline model presented in [9] by 3.37% absolute (8.82% relative). The average unweighted recall rate on the three *emotional* classes (*angry, emphatic*, and *positive*) is at about 52% where the average unweighted recall rate on *non-emotional* (*neutral* and *rest*) classes is only at about 25%. It shows that our proposed framework is capable of retrieving the emotional utterances even given that

Table 2: ***Experiment II:*** *Summary of Result*

Bayesian Logistic Regression (BLG)					
	Unweighted Recall (UA)		Weighted Recall (WA)		
Baseline	38.2%		39.2%		
BLG	**41.57%**		**39.87%**		
	Angry	Emphatic	Neutral	Positive	Rest
Angry	**290**	171	65	63	22
Emphatic	210	**752**	325	136	85
Neutral	748	1094	**2057**	1109	369
Positive	23	13	39	**131**	9
Rest	95	58	134	197	**62**
Support Vector Machine (SVM)					
	Unweighted Recall (UA)		Weighted Recall (WA)		
Baseline	38.2%		39.2%		
SVM	**40.84%**		**38.05%**		
	Angry	Emphatic	Neutral	Positive	Rest
Angry	**249**	191	71	69	31
Emphatic	153	**753**	322	165	115
Neutral	525	1108	**1925**	1152	667
Positive	15	14	35	**136**	15
Rest	73	60	124	210	**79**

some of these emotion classes are only a small portion of the database. This is arguably more significant and can be more advantageous in real world applications where the majority of expressions is likely to be *neutral*.

In summary, our proposed framework for the five-class emotion recognition as a sequence of binary classification tasks is able to improve the unweighted recall by 3.37% absolute (8.82% relative) compared with using Support Vector Machine with SMOTE baseline on the unlabeled evaluation dataset. Since the AIBO database contains very realistic and spontaneous interactions, it is encouraging to see that the framework has some potential to overcome the class imbalance problem in the database and to achieve a comparable recall percentage especially on the *emotional* classes.

4. Conclusion and Future Work

Tracking users' emotion state is essential in promoting an efficient and effective human-robot or human-computer interactions. The AIBO child-machine interaction database consists of five emotion classes and is heavily skewed toward *neutral* speech. A task specific emotion recognition model that targets a balanced emotional retrieval is important in many applications. In this work, we propose a hierarchical binary decision tree that focuses on *easier* subsets of *easier* classification problems at the top level to reduce the accumulation of error. The result indicated in (Section 3) that the average unweighted recall has improved, and the classification framework is capable of recognizing much of the *emotional* classes even though these classes may actually be the less frequently occurred emotions.

Several immediate refinements can be pursued to further enhance the proposed hierarchical framework. Instead of outputting hard labels at every step, a soft label, such as a measure of probability, can provide the framework with more modeling power. Our results could also be improved by performing classifier optimization with feature selection, such as large-margin feature selection for Support Vector Machine. Finally by implementing different ensemble learning techniques with multiple classifiers (combining the Bayesian Logistic Regression and Support Vector Machine), we would expect to see the classification accuracies further improve.

Emotion recognition has become popular in many of the research fields in recent years. It is important to have a well-designed emotion recognition system that can achieve high accuracy even in realistic spontaneous interaction setting for reliable real world applications. Having a reliable automatic emotion recognition system will allow us to progress with many more research hypotheses that can enrich our knowledge about human communication. Further, these improved insights can inform the design of a more robust human-machine spoken interface.

5. Acknowledgements

The paper was supported in part by funds from NSF, Army, USC Annenberg Fellowship, and Intel Foundation PHD Fellowship.

6. References

[1] M. Pantic, N. Sebe, J. Cohn, and T. Huang, "Affective multimodal human-computer interaction," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM New York, NY, USA, 2005, pp. 669–676.

[2] L. Brody and J. Hall, "Gender and emotion in context," *Handbook of Emotions*, pp. 395–408, 2008.

[3] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive robots as social partners and peer tutors for children: A field trial," *Human-Computer Interaction*, vol. 19, no. 1, pp. 61–84, 2004.

[4] A. Kapoor and R. Picard, "Multimodal affect recognition in learning environments," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM New York, NY, USA, 2005, pp. 677–682.

[5] S. Brave, C. Nass, and K. Hutchinson, "Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent," *International journal of human-computer studies*, vol. 62, no. 2, pp. 161–178, 2005.

[6] H. Prendinger, J. Mori, and M. Ishizuka, "Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game," *International journal of human-computer studies*, vol. 62, no. 2, pp. 231–245, 2005.

[7] S. Yildirim, C. M. Lee, S. Lee, A. Potamianos, and S. Narayanan, "Detecting politeness and frustration state of a child in a detecting politeness and frustration state of a child in a conversational computer game," in *In Proc. Eurospeech*, Lisbon, Portugal, October 2005.

[8] M. Black, J. Chang, and S. Narayanan, "An empirical analysis of user uncertainty in problem-solving child-machine interactions: Linguistic analysis of spontaneous children speech," in *Proceedings of the Workshop on Child, Computer and Interaction*, Chnia, Greece, October 2008.

[9] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Interspeech (2009), ISCA*, Brighton, UK, 2009.

[10] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 2, pp. 65–78, 2002.

[11] R. Lazarus, "Relational meaning and discrete emotions," *Appraisal processes in emotion: Theory, methods, research*, pp. 37–67, 2001.

[12] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale bayesian logistic regression for text categorization," *Technometrics*, vol. 49(3), pp. 291–304, 2007.

[13] V. N. Vapnik, *The nature of Statisitcal Learning theory*. New York : Springer, 1995.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[15] T. Joachims, "Making large-scale SVM learning practical," in *Advances in kernel methods - Support Vector Learning, MIT PRESS*, 1998.