Combining Ranking and Classification to Improve Emotion Recognition in Spontaneous Speech

Houwei Cao[1]*, Ragini Verma*[1]*, Ani Nenkova*[2]

[1]Department of Radiology, Section of Biomedical Image Analysis, University of Pennsylvania
[2]Department of Computer and Information Science, University of Pennsylvania
{Houwei.Cao, Ragini.Verma}@uphs.upenn.edu, nenkova@seas.upenn.edu

Abstract

We introduce a novel emotion recognition approach which integrates ranking models. The approach is speaker independent, yet it is designed to exploit information from utterances from the same speaker in the test set before making predictions. It achieves much higher precision in identifying emotional utterances than a conventional SVM classifier. Furthermore we test several possibilities for combining conventional classification and predictions based on ranking. All combinations improve overall prediction accuracy. All experiments are performed on the FAU AIBO database which contains realistic spontaneous emotional speech. Our best combination system achieves 6.6% absolute improvement over the Interspeech 2009 emotion challenge baseline system on the 5-class classification tasks.

Index Terms: emotion classification, ranking models, spontaneous speech

1. Introduction

Expression of emotion is important in human communication and engineers of human-computer interaction systems need ways to include emotion processing [1, 2]. Early studies of emotion recognition performed classification of preselected, prototypical prompted emotional speech [3]. Many investigations of acted speech identified representative features and developed sophisticated classification paradigms [4, 5]. Recently, researchers have shifted their focus to emotion recognition from spontaneous speech in realistic scenarios [6].

State-of-the-art emotion classification systems rate single test utterance with one prediction score for each test utterance independently. However, in many real applications, classifiers will need to process a recording of complete conversations. For example, a user may wish to extract emotional utterances from recordings of telephone and broadcast conversations or political debates and speeches. In such cases, multiple utterances from the same speaker exist, and our novel emotion detection system utilizes this information.

By ranking emotion estimates we can sort all utterances from a sample of speech with respect to the degree of a particular emotional expression. Compared with conventional classification methods, our method helps to account for beneficial information from other samples in the test set. Particularly when an utterance is compared with all other utterances by the same speaker, emotional content is likely to be identify better depending on the expressivity of that speaker. Nevertheless, this type approach was previously unexplored for emotion recognition.

In this paper we introduce a ranking approach for the emotion recognition. Our results show a clear benefit of using a ranking-based model for prediction on spontaneous, non-prototypical emotional speech. The emotion classification system with ranking information significantly outperforms the conventional SVM classifier.

We implemented conventional multi-class SVM classifiers with the LIBSVM toolkit [8], using radial basis kernels. All systems that we present use utterance features that were extracted by the openSMILE feature extraction library [13]. The library provides a comprehensive set of 988 standard acoustic features, involving of prosody, spectral, and voice quality information, that have been proven to be useful for emotion recognition in many previous studies [12].

We now turn to describe in Section 2 how ranking SVMs are adapted for use in emotion recognition. In Section 3 we discuss several methods for combination of conventional SVM classifiers and ranking-based models. The experimental setup is described in Session 4, and the results from various emotion classification system are compared in Session 5.

2. Ranking SVM for Emotion Recognition

Ranking support vector machines (SVM) are a typical pairwise method for designing ranking models. The basic idea is to formalize *learning to rank* as a problem of binary classification on pairs that define a partial ordering and then to solve the problem using SVM classification [7]. The method has had success in information retrieval, where the task is to sort webpages returned by a search engine by relevance to the query.

In this study, we first adapt ranking SVMs for the emotion recognition tasks. The ranking problem is to sort the utterances with respect to how much they convey a particular emotion. To train a ranker for a target emotion, we specify a set of pairs of instances for which one instance conveys the target emotion better than the other; the optimization problem of the ranker is to minimize the number of incorrectly predicted partially ordered pairs.

To train a ranking SVM, we define a partial ordering to form pairs only from utterances from the same speaker and consider all utterances that convey the target emotion to have higher scores than utterances that convey any other emotion. In testing, all utterances from the sample speaker will be given to the ranker for a target emotion. The ranker produces a ranking score for each test utterances and they are sorted by decreasing score. Utterances with higher ranker are considered to express the target emotion more clearly than utterances scored lower. The output from an individual ranker is analogous to a one-vs-all binary classify that attempts to distinguish the target emotion from all others. It should be noted that the original prediction scores given by the rankers can only be used for ordering. They don't have a meaning in an absolute sense and scores predicted from different rankers can not be compared directly in a meaningful way. Normalization should be performed in order to make them comparable. For each test utterance U we define a normalized ranking score for each of the emotions we want to analyze:

$$1 - \frac{rank_m(u)}{N} \quad (1)$$

The score combines information about the number of test utterances in the test set N and the rank of that utterance given by the ranker for emotion m, $rank_m(u)$. These scores fall in the range $[0, 1)$. An utterance will have score 0 if it was the last in the sorted list defined by the SVM ranking (least likely to express the target emotion). The score will be very close to 1 for utterances at the top of the list (which resemble the most the target emotion among all utterances spoken by the speaker).

3. Emotion Classification Systems

3.1. Ranking-based classification system

Given a sample of speech, the emotion rankers indicate how relevant each of the utterances is to a particular emotion. However, the rankers do not directly give a way to decide which particular emotion is expressed by a given utterance. In order to classify the unknown testing utterances as expressing one of the emotion classes, we need to combine ranking scores from different rankers. To do this, we implemented a two-pass system, by training a standard SVM classifier to combine the ranking scores from the individual rankers.

We use a leave-one-subject-out paradigm to train emotion rankers on the training set. The test predictions from each fold are normalized to assign ranking scores for each of the emotional rankers, as described in Section 2. These ranking scores (one for each emotion) are used to represent the utterance and a conventional multi-class SVM classifier is trained, again using leave-one-subject-out paradigm on the training set.

3.2. System Combination

As shown in [9], many combination methods can be effective in emotion recognition tasks. We investigate four approaches to combine the predictions of the ranker-based and conventional SVM classifiers.

We experiment with three rule-based combinations of the classifiers. Each uses a different function to compare or combine the posterior classification probability by each of the two classifiers, for each of the m emotion classes. Assume that p_{ir} is the posterior classification probability for emotion class i by the ranker-derived classifier and p_{ic} is that of the conventional multi-class classifier. Emotion class i is assigned to the utterance where i is defined as

$$\arg\max_{i \in M} (F(p_{1r}, p_{1c}), F(p_{2r}, p_{2c}), ..., F(p_{mr}, p_{mc})) \quad (2)$$

for different choices of the combination function F.

For the *max-combination* rule, the emotion class with highest posterior classification probability, from either classifier, is assigned to an utterance; $F = max$. If two emotions had the same prediction probability, a class is chosen at random to resolve the tie. Here for each emotion we take the higher posterior probability from the conventional and ranking-derived classifiers, and assigned the utterance to the emotion class with highest such probability.

For the *mean-combination* rule, $F = average$ and we combine predictions from different classifiers by averaging their output posterior probabilities. The emotion of an utterance is chosen by comparing the averaging probabilities assigned by the classifies for each emotion classes, and the utterance is classified as conveying the emotion for which it achieved the highest average probability.

For the *multiplication-combination* rule, $F = product$. Instead of averaging posterior probabilities for each class, we multiply them.

Finally, we consider a combination though supervised learning. For each utterance, we generate a new feature representation that consists of the output probabilities for each emotion class, from the conventional and the ranking-derived multi-class SVMs. A new multi-class classifier is trained with these representations.

4. Experimental Setup

We used the FAU AIBO corpus of spontaneous elicited emotional speech. It consists of recordings of 51 Ger-

man children (30 female/21 male), interacting with the AIBO robot controlled by a human invisible to them. The speech data is annotated for the following 11 emotional states: *anger, bored, emphatic, helpless, joyful, motherese, neutral, reprimanding, rest, surprised, touchy*. The whole corpus contains 8.9 hours of speech recording in total and was automatically segmented into turns using a long pause (>1s). The corpus was originally labelled at word-level. Each word was annotated as one of the above eleven states by 5 listeners via majority voting. After that the turn-level labels were derived from word-level labels with confidence scores [11].

We focused on the five-class (**A***ngry,* **E***mphatic,* **N***eutral,* **P***ositive,* **R***est*) classification problem as described in Interspeech 2009 Challenge [10]. The class **R**est contains emotional utterances that do not belong to any of the other classes. There are 18,216 emotional turns corresponding to these five big classes onto which the original 11 emotional states were mapped. They were divided into training and testing sets as shown in Table 1.

Table 1: *Number of instances for the 5-class problem*

	Speaker	A	E	N	P	R
Train	26	881	2093	5590	674	721
Test	25	611	1508	5377	215	546

5. Experimental Results

We evaluate the performance of the proposed emotion recognition systems on the 5-class classification task using the testing set of FAU AIBO corpus. Since the number of instances per emotion class varied widely in the FAU AIBO database as shown in Table 1, we used the unweighted average (UA) recall, also known as balanced accuracy, as performance metric for the emotion recognition experiments presented below. We investigated performance of each of the classifiers, as well as various fusion systems introduced in Section 3. The results were summarized in Table 2.

Two classifiers were trained with the training dataset at the first stage. They are (1) the conventional SVM multi-class classifier and (2) the ranking-SVM based classifiers. From these, we can obtain fusion systems (3), (4), and (5) by model-level combination of the classifiers in terms of the *mean-combination*, *max-combination*, and *multiplication-combination* rule respectively. In addition, a new classifier (6) was trained based on the posterior class probabilities from the two classifiers we wish to combine. Finally, system (7) further combined classifiers (1), (2), and (6) in a *multiplication-combination* way.

The UA recall for the conventional and ranking-derived classifiers were 41.5% and 39.4% respectively. The ranking-derived classifier has 10% better accuracy in identifying *Anger* and 4% better for *Emphatic* utterances. It does however make more mistakes on the *Neutral* class compared to the conventional classifier. Given that the two classifiers perform well on different class, we expect that the combination of the two approaches will further improve performance. Both classifiers have very low accuracy for *Rest*. This is not surprising because *Rest* is a catch-all class combining different infrequent emotions, unlike the other four classes which contained utterances expressing the same emotion.

As expected, the performance clearly improves when the two reference classifiers (conventional and ranking-derived) are combined. The performance of different combination approaches varies by as much as 2%. Combination system (7) achieves the best UA recall of 44.8%, by taking advantage of both model-level and feature-level information. Compared with the baseline SVM classifiers, it exhibits much higher recall rate on all clean emotion classes and maintained a reasonable recall rate on *Neutral*. On the other hand, if we compare with the benchmark result of 38.2% reported in Interspeech 2009 emotion challenge [10], remarkable improvement was achieved [1].

6. Discussion

We carried out further analysis to better understand how the conventional and ranking-derived classifiers complemented each other. We divided the test instances into two groups according to their classification results by the two classifiers:

Group I contains instances classified as conveying the same emotion by the ranking SVM and the conventional SVM classifiers.

Group II contains instance classified as conveying different emotions by the two classifiers.

Table 3 summarizes the results of these two groups with different classification systems. We first observed that there exists significant difference between Group I and Group II in terms of classification performance for the conventional SVM classifier. The UA recall of the conventional SVM classifier for Group I is 50.1%, while the result for Group II for the same system is only 32.2%. This suggests that ranking-based classifiers may help us to retrieve more reliable emotional instances from a large sample of data.

In addition, for Group II, we can get significant improvement from 32.2% to 36.8% by integrating the ranking SVM classifier. Particularly, the UA recall for our best combination system (7) on the three clean emotional classes (*anger, emphatic, positive*) was 46.0%. Compared with the 28.6% obtained by baseline SVM classifiers, UA recall improved by 17.4% absolute (60.8% relative). This promising improvement suggests that the advantage of our proposed approach will be more significant in realistic situations, where the most typical tasks are retriev-

[1]The winner of the challenge achieved 41.65% UA recall [14], and by the democratic fusion of all 17 participants' results, the organizer obtained 44.01% UA recall.

Table 2: *Multi-class classification rate (%) on the FAU AIBO datasets for the 5-class task.*

Individual Systems							
		Anger	Emphatic	Neutral	Positive	Rest	UA
(1)	SVM	50.1	42.6	40.9	53.5	20.5	41.5
(2)	Ranking SVM	60.1	46.5	27.0	52.6	11.0	39.4
Combination Systems							
		Anger	Emphatic	Neutral	Positive	Rest	UA
(3)	F_max{(1), (2)}	58.8	48.4	35.0	58.1	14.8	43.0
(4)	F_average{(1), (2)}	58.1	48.5	37.2	56.3	17.4	43.5
(5)	F_product{(1), (2)}	58.1	48.5	37.6	57.2	19.4	44.2
(6)	F_feature{svm_probability, ranking_score}	58.1	46.7	35.6	69.8	11.9	44.4
(7)	F_product{(1), (2),(6)}	58.1	48.4	37.0	65.1	15.2	**44.8**

Table 3: *Multi-class classification rate (%) on FAU AIBO datasets for 5-class task of two different groups.*

Group I						
	A	E	N	P	R	UA
	68.7	58.3	38.2	73.1	12.4	**50.1**
Group II						
	A	E	N	P	R	UA
System (1)	22.7	31.8	41.6	31.3	33.8	32.2
System (5)	43.4	40.4	37.1	37.5	25.8	36.8
System (7)	43.4	40.5	36.6	54.2	18.1	38.6

ing emotional instances from large sample sets where the majority of expression is likely to be *Neutral*.

7. Conclusions

In this paper, we introduced a novel ranking model for emotion recognition. Emotional rankers are trained to give soft indications of the relevance of an utterance to a particular emotion and a final multi-class classifier predicts emotion based on the output predictions generated by these emotion rankers. Compared with conventional SVM classifiers, ranking-based classifiers recognize many emotional instances better but have lower accuracy for neutral utterances. By combining the two approaches, we achieve performance better than that of either individual classifier. The results indicate that our proposed combination system significantly outperforms the conventional SVM classifier in terms of UA recall. In addition, we observed that performing ranking-SVM based classification in parallel with conventional SVM classification and examining their predictions may help us to differentiate highly reliable predictions from relatively poor ones.

The FAU AIBO corpus contains spontaneous speech and the majority of utterances in it are neutral. Only a relatively small portion of the utterances in the corpus are expressions of spontaneous emotional speech. Our promising results on the FAU AIBO dataset suggest that the proposed emotion recognition system should be able to retrieve emotional instances from large samples of data in many realistic situations. In future work, we will further explore the ranking SVM for acquisition of emotion speech from existing pre-recorded corpora.

8. References

[1] Petrushin, V., 1999. Emotion in speech: recognition and application to call centers. In: Proceedings of Artificial Neural Networks in Engineering. pp. 7–10.

[2] Lee, C. M., Narayanan, S. S., Pieraccini, R., 2002. Classifying emotions in human-machine spoken dialogs. In: Proceedings of the IEEE International Conference on Multimedia & Expo (ICME). pp. 737–740.

[3] Cowie, R., 2000. Describing the emotional states expressed in speech. In: Proceedings of the ISCA Workshop on Speech and Emotion.

[4] McGilloway, S., Cowie, S., Douglas-Cowie, E., Gielen, S., Westerdijk, M., Stroeve, S., 2000. Aproaching automatic recognition of emotion from voice: a rough benchmark. In: ISCA Workshop on Speech and Emotion. pp. 200–205.

[5] Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: Resources, features, and methods. In: Speech Communication. pp. 1162–1181.

[6] Truong, K.P., van Leeuwen, D.A., Neerincx, M.A., de Jong, F.M.G., 2009. Arousal and Valence Prediction in Spontaneous Emotional Speech: Felt versus Perceived Emotion. In: Proceeding of Interspeech. pp. 2027–2030.

[7] Joachims T., 2002. Optimizing Search Engines Using Clickthrough Data. In: Proceedings of ACM SIGKDD Conference on KDD.

[8] Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a library for support vector machines.

[9] López-Cózar, R., Silovsky, J., Griol, D., 2010. F^2 – New Technique for Recognition of User Emotional States in Spoken Dialogue Systems. In: Proceedings of SIGDIAL 2010.

[10] Planet, S. et.al, 2009. GTM-URL Contribution to the INTERSPEECH 2009 Emotion Challenge. In: Proceedings of Interspeech.

[11] Steidl, S., 2009. Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech. Logos Verlag.

[12] Dellaert, F., Polzin, T., Waibel, A., 1996. Recognizing emotion in speech. In: Proceedings of the CMC. pp. 1970–1973.

[13] Eyben, F., Wöllmer, M., Schuller, B., 2010. openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In: Proceedings of the ACM Multimedia (MM). pp. 1459–1462.

[14] Marcel Kockmann, Lukás Burget, Jan Cernocký, 2009. Brno University of Technology system for Interspeech 2009 emotion challenge. In: Proceedings of Interspeech.