On Instance Selection in Audio Based Emotion Recognition

Sascha Meudt and Friedhelm Schwenker

University of Ulm
Institute of Neural Information Processing
89069 Ulm, Germany
{Sascha.Meudt,Friedhelm.Schwenker}@uni-ulm.de
http://www.uni-ulm.de/in/neuroinformatik.html

Abstract. Affective computing aim to provide simpler and more natural interfaces for human-computer interaction applications, e.g. recognizing automatically the emotional status of the user based on facial expressions or speech is important to model user as complete as possible in order to develop human-computer interfaces that are able to respond to the user's action or behavior in an appropriate manner. In this paper we focus on audio-based emotion recognition. Data sets employed for the statistical evaluation have been collected through Wizard-of-Oz experiments. The emotional labels have been are defined through the experimental set up therefore given on a relatively coarse temporal scale (a few minutes) which This global labeling concept might lead to miss-labeled data at smaller time scales, for instance for window sizes uses in audio analysis (less than a second). Manual labeling at these time scales is very difficult not to say impossible, and therefore our approach is to use the globally defined labels in combination with instance/sample selection methods. In such an instance selection approach the task is to select the most relevant and discriminative data of the training set by using a pre-trained classifier. Mel-Frequency Cepstral Coefficients (MFCC) features are used to extract relevant features, and probabilistic support vector machines (SVM) are applied as base classifiers in our numerical evaluation. Confidence values to the samples of the training set are assigned through the outputs of the probabilistic SVM.

Keywords: Emotion Recognition, Human Computer Interaction, Instance Selection, Active Learning.

1 Introduction and Motivation

In supervised learning a large amount of labeled training data has to be collected in order to construct models of acceptable prediction accuracy, and so in pattern recognition or data mining application the training set design is one of the most important parts of the overall process. Designing a training set means pre-processing the raw data, selecting the relevant features, selecting the representative instances (samples), and labeling the samples for application at hand.

N. Mana, F. Schwenker, and E. Trentin (Eds.): ANNPR 2012, LNAI 7477, pp. 186–192, 2012.

Labeling data is time consuming, expensive (e.g. at least in cases where more than one expert must be asked), and of course error-prone. On the other hand, in many pattern recognition applications, such as classification of text documents, remote sensing, or image/video classification, big pools of unlabeled data are available [5,17].

Emotion classification from audio data is a challenging pattern recognition task. Experiments on acted emotional data sets show the human perception capability is similar to the performance of automatic classifiers, particularly humans produce high error rates on emotional data sets, higher than in many other recognition tasks [4]. Labeling emotional data sets is extremely difficult and time consuming, and therefore specific annotation tools must be applied in this task [12], in particular when real world emotional data has to be analyzed, i.e. the emotional utterances are naturalistic emotions in real human-computer-interaction (HCI) scenarios, for instance in Wizard-of-Oz settings [6,19]. In naturalistic HCI scenarios emotional utterances are mainly *neutral*, typically only a few low intensity emotional patterns can be observed in such data streams. The annotation of such WoZ data is usually driven through the experimental design.

Instance selection deals with searching for a small subset S of the original training set T, such that a classifier trained on S shows similar, or even better classification performance than a classifier trained on the full data set T [9,2,11,7]. We will present confidence-based instance selection criteria for probabilistic support vector machines based on cross-validation. The statistical evaluation of the proposed selection method has been performed on task of affect recognition from speech. Classes are not defined very well in this type of application and therefore lead to data sets with high label noise. Numerical evaluations on these data sets show that classifiers can benefit from instance selection not only in terms of computational costs but even in terms of classification accuracy [3,15,10,13].

The paper is organized as follows: In section 2 the data set and feature extraction procedure are briefly described, then in section 3 an overview on the base classifiers is given. Results are discussed in section 4 and a preliminary conclusion with future work is given in section 5.

2 Data Collection and Feature Extraction

The data used to validate the architecture was collected in a Wizard-of-Oz experiment where human-computer interaction (HCI) is simulated [8]. Within the study, the computer interacts as a mental trainer of the popular game "Concentration" while the subjects are able to control the system using short speech commands. The setup induces emotions according to the Valence-Arousal-Dominance (VAD) model [16] using the following affective factors:

- Delaying the response of a command
- Non-execution of the command
- Simulating incorrect speech recognition

- Offering technical assistance
- Lack of technical assistance
- Propose to quit the game ahead of time
- Positive feedback

The procedure of emotion induction is structured in different experimental sequences (ES-4 and ES-6) in which the user is passed through VAD octants by the investigator. Within this study we focus on the recognition of the emotional octants in ES-4 and ES-6 (positive valence, low arousal, high dominance versus negative valence, high arousal, low dominance). The database used consists of 6 subjects with an average age of 63.5 years. Audio, video and physiological data was recorded. In this approach only the audio part was used for classification.

In order to extract the speech non-speech segments from the recorded audio, an energy-based threshold was defined. The energy was determined using a window of 40ms. From this signal, Mel-Frequency Cepstral Coefficients (MFCC) features with a 20 dimensional filter bank were calculated [18]. The windows are shifted with an offset of half the respective window size. This results in about 3000 feature vectors per individual, which in average are equally balanced on the both experimental parts.

3 Methodology

In our architecture we propose the utilization of a Support Vector Machine (SVM). The SVM is a supervised learning method following the maximum margin paradigm. The kernel trick increases the dimensionality of the feature space and therefore allows non-linear non-linear separation surfaces. Within our study we used the Gaussian Radial Basis Function (RBF) kernel, which transforms the input data into the Hilbert space of infinite dimensions and is calibrated by the parameter γ. Due to noise or wrong annotations it is convenient to have a non-rigid hyper-plane, being less sensitive to outliers in training. Therefore, an extension to the SVM introduces a so-called slack term that tolerates the amount of misclassified data using the control parameter C. A probabilistic classification output can be obtained using the method proposed by Platt et al. [14]. Detailed information of the algorithm can be found for instance in [1].

Instance selection Algorithm

```
Input: Dataset T, Reduction amount r, Number of classes l

split Dataset T into N bags
FOR EACH t_i in T (i=1..N)
   train SVM on T without t_i
   classify each x in t_i
END FOR

remove all misclassified examples x from T
```

```
build S by taking most confident r/l examples x of each class

train SVM on S

Output: Reduced Dataset S, SVM
```

First we use a cross-validation based approach to reclassify the complete dataset. We split the original feature dataset T into N bags, using $N-1$ bags for training a probabilistic SVM. This SVM then is used to reclassify the left out bag and mark each feature vector in it with its reclassification decision and confidence. These procedure is repeated for each of the N bags. After marking all feature vectors of the original dataset, the dataset is reduced by keeping only the highest confident and correct reclassified instances. In addition a balancing constraint of the two classes is achieve by taking the same amount of features of each class, even if this implies that the chosen confidence thresholds for the two classes differ. The reduced dataset S then is used to train a second stage classifier where we again chose a probabilistic SVM.

Due to the fact that emotion recognition is a highly individual task, the instance selection and SVM training task is done separately for each individual.

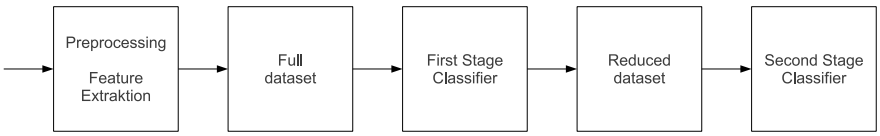

Fig. 1. Architecture of the instance selection based architecture

4 Statistical Evaluation

In this section, the statistical evaluation of the instance selection architecture is described for the above mentioned data. First the first stage classifier is evaluated which is used for the selection. In a second part the second stage classifiers that where trained with the reduced dataset are evaluated. We used three different reduction intensities of 98%, 90% and 80% reduction from the original dataset. The reduced dataset contains an equal balanced amount of feature vectors from both classes. The SVM parameters where set to $C = 2$ and $\gamma = 4 \cdot 10^{-3}$. Each individual result was evaluated by a 5 fold cross-validation procedure. The individual results where combined by averaging the individual results.

In case of the default classification approach without reduction or rejection we reached a classification accuracy of 65.9% and a F1 measure of 0.628. Adding a rejection option of at least 0.95 classification confidence the classification accuracy increases to 81.8% (F1 0.790). The amount of rejection is nearly linear to the confidence threshold. This means that in case of 0.95 minimum confidence about 90% of the decisions had been rejected.

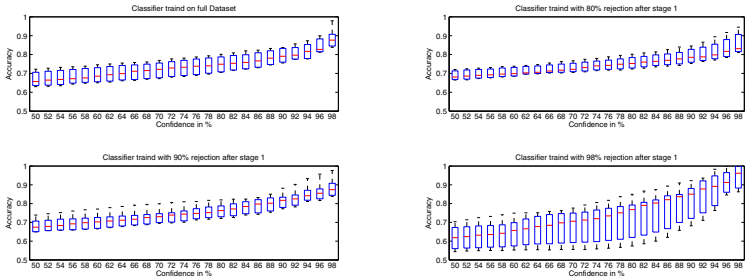

Fig. 2. Accuracy depending on confidence threshold of rejection. Average of all subjects.

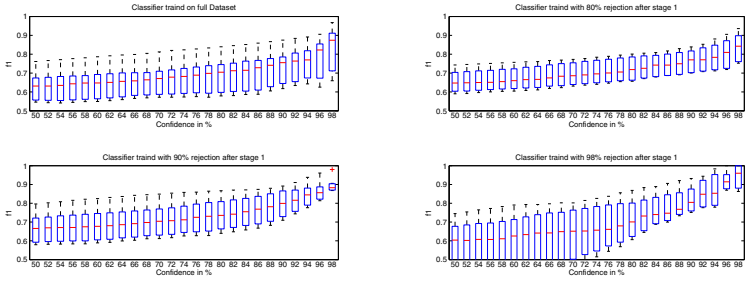

Fig. 3. F1 measure depending on confidence threshold of rejection. Average of all subjects.

Adding the second stage trained with a 80% reduced dataset improved this value. In case of rejectionless classification we reached a classification accuracy of 68.5% with nearly half standard derivation compared to the classifier trained on the whole dataset. Also the F1 measure got up to 0.681 with better standard derivation.

Reducing the dataset to 2% of the whole set and adding a rejection option with minimum confidence of 0.95 we achieved the best result. The classification accuracy got up to 96% also the F1 measure improved to 0.965. In addition the confidence derivation got better, more SVM classification decisions got high confidence values. In case of rejection decision with less than 0.95 confidence we got nearly twice the amount of decisions compared with the not reduced classifier. Keeping in mind that the feature extraction is based on a shifted window with offset of a half window, this implies that around 40% of the time-line is covered with decisions. Or on the other hand the confidence threshold could be set up to nearly 0.9 (Accuracy 87%) by still covering nearly the whole time-line with decisions.

In all cases we could find an instance selection based architecture that outperforms the well known standard SVM training approach in all cases (Accuracy, F1 measure, Standard Derivation and Confidence Derivation). Finally the

computational cost for training and classification on a heavy reduced dataset is highly decreased compared to training on a large dataset with lots of noisy data in it.

5 Conclusion and Future Work

Classifying the emotion is generally a difficult task when leaping from overacted data to realistic human computer interaction. In this study the problem was investigated with respect to the fact that only less parts of the dataset contain intense emotion. The result of the evaluation is that the usage of instance selection can reduce the testing error, or reducing the standard derivation depending on the chosen parameters.

Rejecting samples when classifying such kind of data turns out to be a sound approach. Especially when the distribution of the classes in the data is heavily overlapping. For future work, it could be promising to implement an iterative classifier training procedure, were the training data can be rejected.

Using more than just the audio part of the dataset could also improve the results, by using a co training multi classifier approach.

Acknowledgments. This research was supported in part by grants from the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

1. Bishop, C.: Pattern recognition and machine learning, vol. 4. Springer, New York (2006)
2. Brighton, H., Mellish, C.: Advances in instance selection for instance-based learning algorithms. Data Mining and Knowledge Discovery 6(2), 153–172 (2002)
3. Domingo, C., Gavaldà, R., Watanabe, O.: Adaptive sampling methods for scaling up knowledge discovery algorithms. Data Mining and Knowledge Discovery 6(2), 131–152 (2002)
4. Esparza, J., Scherer, S., Brechmann, A., Schwenker, F.: Automatic emotion classification vs. human perception: Comparing machine performance to the human benchmark. In: International Conference on Information Science, Signal Processing and Their Applications (ISSPA 2012), pp. 1286–1291 (2012)
5. Esparza, J., Scherer, S., Schwenker, F.: Studying Self- and Active-Training Methods for Multi-feature Set Emotion Recognition. In: Schwenker, F., Trentin, E. (eds.) PSL 2011. LNCS, vol. 7081, pp. 19–31. Springer, Heidelberg (2012)
6. Glodek, M., Tschechne, S., Layher, G., Schels, M., Brosch, T., Scherer, S., Kchele, M., Schmidt, M., Neumann, H., Palm, G., Schwenker, F.: Multiple classifier systems for the classification of audio-visual emotional states. In: 1st International Audio/Visual Emotion Challenge and Workshop (2011)
7. de Haro-García, A., García-Pedrajas, N., del Castillo, J.A.R.: Large scale instance selection by means of federal instance selection. Data Mining and Knowledge Engineering 75, 58–77 (2012)

8. Kelley, J.: An iterative design methodology for user-friendly natural language office information applications. ACM Transactions on Information Systems (TOIS) 2(1), 26–41 (1984)
9. Liu, H., Motoda, H.: Instance Selection and Construction for Data Mining. Kluwer Academic Publishers, Norwell (2001)
10. Liu, H., Motoda, H.: On issues of instance selection. Data Mining and Knowledge Discovery, 115–130 (2002)
11. Madigan, D., Raghavan, N., DuMouchel, W., Nason, M., Posse, C., Ridgeway, G.: Likelihood-based data squashing: A modeling approach to instance construction. Data Mining and Knowledge Discovery 6(2), 173–190 (2002)
12. Meudt, S., Bigalke, L., Schwenker, F.: ATLAS – an annotation tool for HCI data utilizing machine learning methods. In: Proceedings of the 4th Internantional Conference on Applied Human Factors and Ergonomics, AHFE 2012 (in print, 2012)
13. Olvera-Lpez, J.A., Carrasco-Ochoa, J.A., Trinidad, J.F.M., Kittler, J.: A review of instance selection methods. Artificial Intelligence Reviews, 133–143 (2010)
14. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers 10(3), 61–74 (1999)
15. Reinartz, T.: A unifying view on instance selection. Data Mining and Knowledge Discovery 6(2), 191–210 (2002)
16. Russell, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. Journal of Research in Personality 11(3), 273–294 (1977)
17. Schels, M., Kächele, M., Hrabal, D., Walter, S., Traue, H.C., Schwenker, F.: Classification of Emotional States in a Woz Scenario Exploiting Labeled and Unlabeled Bio-physiological Data. In: Schwenker, F., Trentin, E. (eds.) PSL 2011. LNCS, vol. 7081, pp. 138–147. Springer, Heidelberg (2012)
18. Scherer, S., Glodek, M., Schwenker, F., Campbell, N., Palm, G.: Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data. TiiS 2(1), 4 (2012)
19. Walter, S., Scherer, S., Schels, M., Glodek, M., Hrabal, D., Schmidt, M., Böck, R., Limbrecht, K., Traue, H.C., Schwenker, F.: Multimodal Emotion Classification in Naturalistic User Behavior. In: Jacko, J.A. (ed.) HCII 2011, Part III. LNCS, vol. 6763, pp. 603–611. Springer, Heidelberg (2011), http://www.springerlink.com/content/606237v0u5225w50/