Bagging Based Feature Selection for Dimensional Affect Recognition in the Continuous Emotion Space

Aihua Chen, Shuai Yuan and Dongmei Jiang[1]

Abstract This paper exploits the Bagging based feature selection method on the baseline audio features provided by AVEC2012 challenge competition. The selected features are input to SVR and RVM regression models, respectively, to estimate the affect dimensions arousal, valence, expectation, and power embedded in the audio speech. Experiments have been carried out on the word based and frame based baseline features, respectively, and the Pearson correlations between the estimated affect dimensions and their ground-truth labels are compared to those from the traditional correlation based feature selection (CFS) method with BestFirst or sequential floating forward selection (SFFS) algorithm. Experimental results show that both on word based and frame based baseline feature selection obtains the best accuracy in estimating the affect dimensions, while keeping the lowest number of features.

Keywords Affect dimensions · Feature selection · Bagging · Correlation based feature selection

1 Introduction

Human-computer interactions will be more natural if computers are able to perceive and respond to the non-verbal actions of human beings, such as emotion. Emotion recognition has attracted increasing attention of researchers from various fields including psychology, cognition and computer science. Most of the proposed systems have focused on the recognition of acted emotions with the six basic emotions introduced in the early 70s by Ekman [1]. However, because the limited categorical emotion description methods are not able to capture the complex relationship and subtle differences between the spontaneous emotions, and neither can they explicitly model the changes of affect over time, analysis of

A.Chen (✉)，S. Yuan (✉)， D. Jiang (✉)
Northwestern Polytechnic University, Xi'an, China
Shaanxi Provincial Key Laboratory on Speech, Image and Information Processing
e-mail: veket201791@126.com, 237079166@qq.com, jiangdm@nwpu.edu.cn

spontaneous affect behavior still remains a challenge which can't be solved with the traditional pattern classification models such as support vector machine (SVM) or hidden Markov model (HMM).

In recent years, some dimensional affect recognition methods have been proposed, where the affective state is characterized in terms of a number of latent dimensions in the psychological emotion space, aiming to improving the understanding of human affect by modeling affect as a small number of continuously valued, continuous time signals [2]. During the competition of the Audio-Visual Emotion Challenge and Workshops AVEC2011 and AVEC2012 [2,3], various audio visual feature extraction methods, regression models and later fusion methods have been proposed to estimate the dimensions arousal, valence, expectation, and power in the emotion space, and promising results have been obtained.

However, in the current affect dimension estimation methods, high dimension of the audio and visual features still remains a problem. To capture the characteristics of the features over a period, statistical functionals, such as arithmetic mean, standard deviation, et al., are normally made on the low-level descriptors (LLD) such as energy, pitch and other spectral features. For example, 1841 features are extracted as the baseline audio feature set in AVEC2012 [2]. In [4], 64988 dimensional visual feature vectors are extracted from the temporal information with the mean and standard deviation over fixed length temporal windows on the face image sequence.

Feature selection is prerequisite before applying the high dimension features in the regression models, besides reducing the computational costs, feature selection could remove the noisy information and therefore lead to a better regression of the affective state. In recent publications, correlation based feature selection (CFS) [5] has been widely adopted on audio or visual features for emotion recognition in discrete or continuous emotion space, with sequential floating forward selection (SFFS) algorithm [6, 7], or BestFirst algorithm [8] as the searching strategy. However, in the above CFS methods, the components of the features are regarded as independent of each other. Moreover, in the regression process of feature selection, only the correlations between the feature-target as well as between the features are considered, while the residuals, which have been proved important in [9] for feature selection, are ignored.

Ensemble learning improves generalization performance of individual learners by combining the outputs of a set of diverse base classifiers. Various works have demonstrated its significant improvements in the accuracy of feature selection [10]. Bagging is a successful ensemble method based on bootstrapping and aggregating concepts, i.e., the training set is randomly sampled many times with replacement to construct several base classifiers which are then aggregated.

In this paper, we try to exploit the Bagging based feature selection on the baseline audio features of the AVEC2012 database, with the regression tree as the predictor, and the sum-of-squares of the residuals, as well as the correlations between the predicted labels and ground truth labels to measure the goodness of the subsets. After the audio features of the AVEC2012 training set are selected, they are input to support vector regression (SVR, [11]) or relevance vector

machine for regression (RVR, [11], [12]) to estimate the affect dimensions arousal, valence, expectation, and power. Results are compared to those from the audio features selected by the CFS related methods.

The remainder of this paper is as follows. In Section II, we introduce the continuous emotion space with the dimensions arousal, valence, expectation, and power. Section III describes the Bagging algorithm and the process for feature selection. Section IV briefly introduces the regression models SVR and RVR. The experimental results are analyzed in Section V, and in Section VI, we draw conclusions.

2 Continuous Emotion Space

Different with the categorical emotion description, the continuous dimensional emotion space thinks affect states are related to one another in a systematic manner. According to [13], the majority of affect variability can be covered in a continuous space defined in terms of two orthogonal dimensions, valence and arousal. Valence refers to how positive or negative the emotion is, ranging from miserable feelings to pleasant feelings of happiness. Arousal refers to how excited or apathetic the emotion is, ranging from sleepiness to frantic excitement. Later, [14] proved that four dimensions activity, expectation, power and valence, are needed to satisfactorily represent the similarities and differences in the meaning of emotion words. Activity has the same meaning as arousal. Expectation subsumes various concepts that can be separated as expecting, anticipating, being taken unaware. The power dimension subsumes two related concepts, power and control. In the AVEC2011 and AVEC2012 challenge competition workshops [2, 3], the four dimensions (arousal, expectation, power, and valence) have been labeled for the videos of the training and development set.

3 Bagging Based Feature Selection

3.1 Original Audio Features

For the original features before feature selection, we adopt the baseline audio features of the SEMAINE corpus provided by AVEC2012 [2], which consists of 1841 features, composed of 42 functionals over 25 energy and spectral related low-level descriptors (LLD), 32 functionals over 6 voicing related LLD, 19 functionals over 25 delta coefficients of the energy/spectral LLD, 19 functionals over 6 delta coefficients of the voicing related LLD, and 10 voiced/unvoiced durational features. The detailed description of the SEMAINE corpus, as well as the baseline audio and video features, can be found in [2].

3.2 Bagging Based Feature Selection

Suppose the training set of $\mathcal{L}$ consists of data $\{(y_n, \boldsymbol{x}_n), n = 1, \dots, \mathcal{N}\}$ where the y's are either class labels or a numerical response, and $\boldsymbol{x} = (x_1, \dots, x_{\mathcal{M}})$ are the input features. Suppose there is a predictor $\varphi(\boldsymbol{x}, \mathcal{L})$ predicting y from the input $\boldsymbol{x}$, given a sequence of training sets $\{\mathcal{L}_k\}$ each consisting of $\mathcal{N}$ independent observations from the same underlying distribution as $\mathcal{L}$, the mission is to use the $\{\mathcal{L}_k\}$ to get a better predictor than the single training set predictor $\varphi(\boldsymbol{x}, \mathcal{L})$.

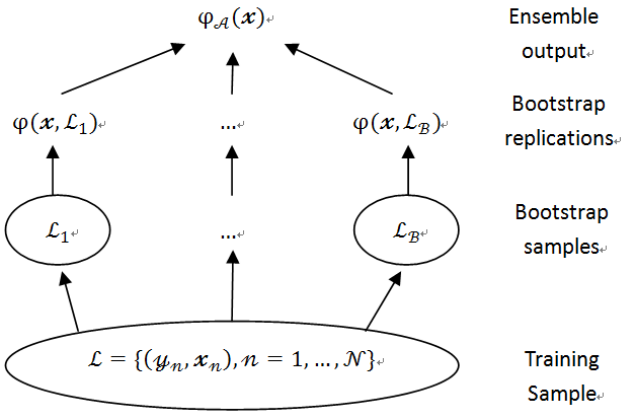

Fig. 1 The bootstrap aggregating (bagging) procedure

Take repeated bootstrap samples $\{\mathcal{L}_k\}$ from $\mathcal{L}$, $k = 1, \dots, \mathcal{B}$, and form a predictor $\{\varphi(\boldsymbol{x}, \mathcal{L}_k)\}$. If y is numerical, take $\varphi_{\mathcal{A}}(\boldsymbol{x})$ as

$$\varphi_{\mathcal{A}}(\boldsymbol{x}) = av_{\mathcal{B}}\varphi(\boldsymbol{x}, \mathcal{L}_k) \quad (k = 1, \dots, \mathcal{B}) \tag{1}$$

Where the subscript $\mathcal{A}$ in $\varphi_{\mathcal{A}}$ denotes aggregation, and $av_{\mathcal{B}}$ denotes the comprehensive, normally the average, over the $\mathcal{B}$ predictors. If y is a class label, let the $\{\varphi(\boldsymbol{x}, \mathcal{L}_k)\}$ vote to form $\varphi_{\mathcal{A}}(\boldsymbol{x})$. This procedure, as illustrated in Fig.1, is called "bootstrap aggregating", or with the acronym "Bagging" [15].

Referring to [15] and [16], in this paper, we choose the classification and regression trees (CART) as the predictor, and the following procedure has been adopted for feature selection:

1. The baseline audio features of the AVEC2012 training set is taken as the learning set $\mathcal{L}$ for feature selection.
2. A bootstrap sample $\mathcal{L}_k$ is randomly selected from $\mathcal{L}$, and the out-of-bag observations are used as the set $\mathcal{T}_k$ to evaluate the importance of the features. This is repeated $\mathcal{B}$ times ($\mathcal{B} = 100$ in our experiments) giving the regression tree predictors $\{\varphi(\boldsymbol{x}, \mathcal{L}_k)\}$ $(k = 1, \dots, \mathcal{B})$.
3. For $(y_n, \boldsymbol{x}_n) \in \mathcal{T}_k$, where $\boldsymbol{x}_n$ is the dataset containing n samples, and y_n is the set of n labels, the bagged predictor is $\hat{y}_n = av_{\mathcal{B}}\varphi(\boldsymbol{x}, \mathcal{L}_k)$, and the mean squared error (MSE) of the bagged prediction, $e(\mathcal{T}_k)$, is $av_n(y_n - \hat{y}_n)^2$, meaning the average over n targets. For the $\mathcal{B}$ out-of-bag observation

sets $\{\mathcal{T}_k\}$ $(k = 1, \ldots, \mathcal{B})$, the overall out-of-bag MSE $e_{\mathcal{B}}$ is calculated as the average of $\{e(\mathcal{T}_k)$ $(k = 1, \ldots, \mathcal{B})\}$.

4. By randomly permuting the out-of-bag data across one variable x_m at a time, and calculating the overall out-of-bag MSE $e_{\mathcal{B}}$ averaged over all trees in the ensemble, we can obtain the increase $\Delta e_{\mathcal{B}}(m)$ in the out-of-bag MSE due to this permutation, divided by the standard deviation taken over the regression trees for this feature.
5. Finally we rank from large to small the $\{\Delta e_{\mathcal{B}}(m)\}$ obtained by permuting different features. The larger the value, the more important the feature is.

In our experiments, in drawing the bootstrap samples $\{\mathcal{L}_k\}$ from $\mathcal{L}$, we draw $\mathcal{N}$ out of $\mathcal{N}$ observations with replacement. This will omit on average 37% of observations, i.e. "out-of-bag" observations, for each decision tree. The minimal leaf sizes for the bagged trees are set to 5, which are close to optimal for the predictive power of an ensemble, and the number of features selected at random for every decision split is set as one third of the features.

4 Regression Models for Dimensional Affect Recognition

Sparse kernel machines SVR and RVR are two state-of-the-art machine learning techniques used in the target problem. In our work, the epsilon-SVR with a radial basis function kernel, implemented by LibSVM for matlab [17], is adopted for regression. As an alternative sparse kernel technique, RVR, which is based on a Bayesian formulation and provides posterior probabilistic outputs, has much sparser solutions than the SVR. In our experiments, we adopt the SparseBayes package for matlab [18] for the implementation of RVM.

5 Experiments and Analysis

5.1 Experimental Setup

In our experiments, we used the pre-processed version of AVEC2012 data set as described in [2]. The data set contains two types of baseline audio features according to the length of the episodes: 1) word based with episodes of whole words (WL), in which one audio feature per word is extracted; 2) frame based with episodes of 2 second sliding windows, where the audio features are extracted at 0.5 second intervals, but only during speech.

In our experiments, affect dimension estimation experiments are performed on the above two levels, respectively. For each of the affect dimensions, we firstly perform the CFS with BestFirst or SFFS algorithm, and then use the selected

features, together with the labels of the corresponding affect dimension, to train the SVR or RVM models. Selected features of the testing set are input into the trained SVR or RVM models to estimate the continuous values of the affect dimensions embedded in the audio speech. For the experiments on the Bagging based feature selection, we choose the subsets whose numbers of features do not exceed or are very close to those from the CFS methods.

For each affect dimension, the performance of the feature selection method with the regression models is measured via the Pearson correlation between the estimated values and the ground truth labels in the testing set. After the Pearson correlations of all sessions in the testing set are calculated, we take the average over the testing sessions as the final result to evaluate the estimation of the affect dimension.

5.2 Results and Analysis

Table 1 Pearson's correlations with the word based baseline features

SVR	Arousal	Expectation	Power	Valence	Mean
AVEC(baseline)	0.054	0.020	0.019	0.062	0.039 (1841)
Original	0.052	0.016	0.002	0.069	0.035 (1841)
CFS+BestFirst	0.073 (52)	0.039 (65)	0.001 (62)	0.055 (47)	0.042 (57)
CFS+SFFS	0.070 (44)	0.059 (53)	0.003 (64)	0.064 (56)	0.049 (55)
Bagging	*0.078 (32)*	*0.066 (7)*	*0.011 (41)*	*0.070 (28)*	*0.057 (27)*
RVR	Arousal	Expectation	Power	Valence	Mean
Original	0.073	0.063	0.009	0.078	0.056(1841)
CFS+BestFirst	0.070 (52)	0.036 (65)	0.022 (62)	0.053 (47)	0.045 (57)
CFS+SFFS	0.068 (44)	0.019 (53)	0.034 (64)	0.067 (56)	0.047 (55)
Bagging	*0.081 (32)*	*0.071 (7)*	*0.010 (41)*	*0.082 (28)*	*0.061 (27)*

Table 2 Pearson's correlations with the frame based baseline features

SVR	Arousal	Expectation	Power	Valence	Mean
Original	0.164	0.199	0.099	0.091	0.138 (1841)
CFS+BestFirst	0.173 (59)	0.206 (64)	0.092 (67)	0.071 (43)	0.136 (59)
CFS+SFFS	0.185 (28)	0.209 (44)	0.088 (48)	0.074 (38)	0.139 (40)
Bagging	*0.209 (27)*	*0.220 (40)*	*0.097 (32)*	*0.077 (46)*	*0.150 (37)*
RVR	Arousal	Expectation	Power	Valence	Mean
Original	0.169	0.205	0.095	0.091	0.140 (1841)
CFS+BestFirst	0.171 (59)	0.210 (64)	0.090 (67)	0.082 (43)	0.138 (59)
CFS+SFFS	0.184 (28)	0.211 (44)	0.085 (48)	0.090 (38)	0.143 (40)
Bagging	*0.201 (27)*	*0.234 (40)*	*0.115 (32)*	*0.108 (46)*	*0.165 (37)*

The results on the word based baseline audio features are shown in Table 1, where the Pearson correlations of the four affect dimensions arousal, expectation,

power, and valence, obtained from SVR and RVM, are listed respectively. The number between brackets is the number of features being selected. The results with the frame based baseline features are listed in Table 2.

From Table 1 and Table 2, one can notice that: 1) overall, the Pearson correlations from the frame based baseline features are higher than those from the word based baseline features, showing that frame based features are more suitable for the estimation of affect dimensions in the continuous emotion space. 2) Either for word based or frame based baseline features, and either on SVR or RVM regression models, for each affect dimension, compared to those from the CFS method with BestFirst or SFFS algorithm, the Bagging based feature selection method obtains the highest Pearson correlations, while keeping the lowest feature numbers (except for that in the estimation of valence on frame based features). This means that the Bagging based feature selection can obtain better performance than the CFS method with less selected features.

6 Discussion

This paper exploits the Bagging based method on the feature selection of the baseline audio features provided by AVEC2012 challenge competition. The selected features are input to SVR and RVM regression models, respectively, to estimate the affect dimensions arousal, valence, expectation, and power embedded in the audio speech. Experiments have been carried out on the word based and frame based baseline features, respectively, and results are compared to those from the traditional correlation based feature selection (CFS) method. Experimental results show that both on word based and frame based baseline features, for each affect dimension, compared to the CFS method with BestFirst or SFFS searching algorithm, the Bagging based feature selection obtains the best accuracy in estimating the affect dimensions, while keeping the lowest feature numbers.

In our future work, we would like to expand the Bagging based feature selection method on the visual features from the face images, and test its performance on more regression models such as long short term memory recurrent neural networks [8].

Acknowledgments This work is supported within the framework of the national natural science foundation project of China (grant 61273265), the Shaanxi provincial key international cooperation project (2011KW-04).

References

1. Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, 301-320.
2. Schuller, B., Valster, M., Eyben, F., Cowie, R., & Pantic, M. (2012, October). AVEC 2012:

the continuous audio/visual emotion challenge. *Proceedings of the 14th ACM international conference on Multimodal interaction*, 449-456.

3. Schuller, B., Valstar, et al. (2011). AVEC 2011–the first international audio/visual emotion challenge. *Affective Computing and Intelligent Interaction*, 415-424.
4. Savran, A., Cao, et al. (2012, October). Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. *Proceedings of the 14th ACM international conference on Multimodal interaction*, 485-492.
5. Hall, M. A. (1999). Feature selection for discrete and numeric class machine learning.
6. Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern recognition letters*, *15*(11), 1119-1125.
7. D. Ververidis & C. Kotropoulos (2006, September). Fast Sequential Floating Forward Selection Applied to Emotional Speech Features Estimated on DES and SUSAS Date Collections. *Proceedings of the 14th European Signal Proceesing Conference*, 4-8. Florence, Italy.
8. Wöllmer, M., Schuller, B., Eyben, F., & Rigoll, G. (2010). Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *Selected Topics in Signal Processing, IEEE Journal of*, *4*(5), 867-881.
9. Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17-21.
10. Wagner, J., Kim, J., & André, E. (2005, July). From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, 940-943.
11. Bishop, C. M. (2006). *Pattern recognition and machine learning*, Vol. 1, 740.
12. Tipping, M. E., & Faul, A. C. (2003, January). Fast marginal likelihood maximisation for sparse Bayesian models. *Proceedings of the ninth international workshop on artificial intelligence and statistics*, Vol. 1, No. 3.
13. Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161-1178.
14. Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological science*, *18*(12), 1050-1057.
15. Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.
16. Jones, B. (1993). *MATLAB: Statistics Toolbox User's Guide*. MathWorks, http://www.mathworks.cn/cn/help/stats/ensemble-methods.html. Accessed on 19 Jan 2013.
17. Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 27.
18. M.E. Tipping (2009). An Efficient Matlab Implementation of the Sparse Bayesian Modelling Algorithm (Version 2.0), *http://www.relevancevector.com*. Accessed 2 Jan 2013.