Feature selection in multimodal continuous emotion prediction

Shahin Amiriparian, Michael Freitag, Nicholas Cummins, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Amiriparian, Shahin, Michael Freitag, Nicholas Cummins, and Björn Schuller. 2017. "Feature selection in multimodal continuous emotion prediction." In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 23-26 Oct. 2017, San Antonio, TX, USA*, 30–37. Los Alamitos, CA: IEEE. https://doi.org/10.1109/aciiw.2017.8272619.

Feature Selection in Multimodal Continuous Emotion Prediction

Shahin Amiriparian[*†‡], Michael Freitag[†], Nicholas Cummins[*†] and Björn Schuller[*‡]
[*]*Chair of Embedded Intelligence for Health Care & Wellbeing, Augsburg University, Augsburg, Germany*
[†]*Chair of Complex & Intelligent Systems, Universität Passau, Germany*
[‡]*Machine Intelligence & Signal Processing Group, Technische Universität München, Germany*
[§]*GLAM – Group on Language, Audio & Music, Imperial College London, London, UK*

Email: shahin.amiriparian@tum.de

Abstract—**Advances in affective computing have been made by combining information from different modalities, such as audio, video, and physiological signals. However, increasing the number of modalities also grows the dimensionality of the associated feature vectors, leading to higher computational cost and possibly lower prediction performance. In this regard, we present an comparative study of feature reduction methodologies for continuous emotion recognition. We compare dimensionality reduction by principal component analysis, filter-based feature selection using canonical correlation analysis, and correlation-based feature selection, as well as wrapper-based feature selection with sequential forward selection, and competitive swarm optimisation. These approaches are evaluated on the AV+EC-2015 database using support vector regression. Our results demonstrate that the wrapper-based approaches typically outperform the other methodologies, while pruning a large number of irrelevant features.**

1. Introduction

In view of the fundamental importance of showing and recognising emotions in human-human interaction, automatic emotion recognition opens up a multitude of intriguing opportunities, especially for human-computer interaction. Due to its vast range of possible real-life applications, including, for example, in education or gaming, [1], automatic emotion recognition is attracting a growing number of researchers from differing fields [2].

Since emotion is a continuous phenomenon, a natural task in this domain is continuous emotion prediction based on streams of information. Previous research has shown that multiple modalities can significantly improve prediction accuracy [3], [4], [5]. However, with each additional modality, more features need to be considered when training a model, which can quickly lead to issues with training time and the 'curse of dimensionality'. Indeed, it has been shown that the naive fusion approach of concatenating the features from all modalities for training is a suboptimal approach [6].

Most current approaches reduce the effective dimensionality of the feature set, e. g. by fusing predictions on the individual modalities [4], [6], quantising features [7], [8], or dimensionality reduction techniques such as PCA [9]. Filter-based feature selection approaches are also commonly used in the speech processing domain [2], [10]. Such approaches typically employ statistical tests or measures, such as information gain [11], the Kolmogorov-Smirnov test [12], canonical correlation analysis [13], or the chi-squared test [14].

However, to date, few emotion prediction approaches employ wrapper-based feature selection. Heuristic wrapper algorithms, such as stepwise forward selection [15], tend to struggle with the exponential size of the feature set search space for large real-world problems [16, Ch. 17]. Nevertheless, such algorithms have been employed successfully in affective computing [17].

In this paper, we perform a comparative study of several archetypal feature selection approaches for multimodal continuous emotion prediction. In addition to these, we also explore a recent evolutionary wrapper-based feature selection approach, which is based on competitive swarm optimisation (CSO) [18], [19]. This algorithm has previously been shown to perform well in computational paralinguistics [20]. First, a strong baseline is established on the 2015 Audio-Visual Emotion recognition Challenge (AVEC 2015) corpus [3]. Then, we compare dimensionality reduction by principal component analysis (PCA) [21], filter-based feature selection using canonical correlation analysis (CCA) [13], and correlation-based feature subset selection (CFS) [22], as well as wrapper-based feature selection with sequential forward selection (SFS) [15], and competitive swarm optimisation.

The remainder of this paper is organised as follows. In Section 2, we provide an overview of the selected feature selection approaches. Section 3 outlines our experiments, and discusses the results thereof. Finally, we draw conclusions, and provide an outlook on future work in Section 4.

2. Feature Selection Approaches

The approaches investigated are arranged into three categories: 1) principal component analysis for dimensionality reduction, 2) canonical correlation analysis and correlation-based feature selection for filter-based feature selection, and 3) sequential forward selection and competitive swarm optimisation for wrapper-based feature selection.

2.1. Principal Component Analysis

Principal component analysis (PCA) is a widely used statistical procedure for dimensionality reduction. PCA transforms a set of intercorrelated features into a possibly smaller set of linearly uncorrelated features, called principal components. This is achieved by applying an orthogonal transformation to the original feature space, i. e. the principal components are linear combinations of the original features. Feature selection using PCA is commonly performed by using only the first k principal components, which account for a predefined fraction, e. g. 95 %, of variability in the data [21]. As no label information is used, PCA is an entirely unsupervised approach.

2.2. Canonical Correlation Analysis

Canonical correlation analysis is a statistical method related to PCA, which can also be used for feature selection [13]. Whereas PCA transforms the feature space so as to optimally model variance in a single data set, CCA defines linear transformations which maximise the mutual correlation between two vector spaces representing the same underlying semantic phenomenon. CCA can be formulated as a generalised eigenproblem, in which larger eigenvalues correspond to projection vectors which capture increasingly more correlation between the vector spaces [13]. Therefore, feature selection can be performed by using only the first k projection vectors associated with the largest eigenvalues.

Following [13], we use a CCA based regression scheme, where one vector space represents the training data, and the other vector space represents the target values. Therefore, CCA finds a linear projection of the training data which maximises the correlation with the target values, thus facilitating prediction. This approach has previously been shown to be suitable for emotion and depression recognition systems [13], [23].

2.3. Correlation-based Selection

Correlation-based feature selection is based on the central hypothesis that a good feature subset contains features which are highly correlated with the target value, but uncorrelated with each other [22]. CFS defines an evaluation function for feature subsets, which prefers feature subsets that simultaneously have high predictive ability, i. e. correlation with the target value, and low redundancy, i. e. inter-correlation between features.

Since evaluating all possible feature subsets is infeasible due to the exponential size of the feature set search space, CFS needs to be coupled with a suitable heuristic search algorithm. In this paper, we explore sequential forward selection and sequential backward elimination, both of which have been shown to perform well with CFS [22].

2.4. Sequential Forward Selection

Sequential forward selection is one of the first heuristic search algorithms that has been used for wrapper-based feature selection [15]. Due to its simplicity, and often high efficiency, it remains a highly popular approach in a multitude of fields, including emotion prediction [24]. Feature selection with SFS begins with the empty feature set, and iteratively adds features which yield the highest increase in prediction performance for the feature set. When no further improvement can be achieved by a single feature addition, the algorithm terminates. Even though numerous improvements of the basic SFS wrapper algorithm have been devised [24], we found that even in its most basic form, SFS can produce highly competitive results on the AVEC 2015 database.

Sequential backward elimination (SBE) is closely related to SFS, but starts with the full feature set and iteratively removes single features. SBE commonly produces much larger final feature sets, and requires substantially more computation time than SFS and is not considered for investigation in this paper.

2.5. Competitive Swarm Optimisation

Competitive swarm optimisation [18], [19] is an adaptation of the canonical particle swarm optimisation algorithm [25] for large-scale optimisation. Similar to other evolutionary feature selection algorithms, a set of candidate feature vectors, called particles, is maintained. Each particle has a velocity, and is allowed to move through the feature set search space over several time steps, or generations. Particles are assigned a fitness value, which describes the prediction performance of the feature set they represent. At each time step, particles move a small distance according to their velocity. Furthermore, they are accelerated towards solutions with the currently highest fitness values. Over time, the particle swarm converges toward an optimal solution.

CSO has been developed specifically for large-scale optimisation problems such as high-dimensional feature selection [18], [19], which is a common problem in speech processing [10].

2.6. Early Stopping

Wrapper algorithms, as their optimisation procedure is guided by the prediction performance of the feature sets, as prone to overfitting during optimisation [15]. We therefore test an early stopping mechanism based on a separate validation set, which has previously been shown to be a reasonable technique to combat this effect [26].

In each SFS iteration and, respectively, CSO generation, the best feature set found is evaluated on a separate validation set, which consists of entirely unseen data. If no performance improvement has been observed on the validation set for a certain number of iterations, feature selection is terminated and the feature set with the best performance on the validation set is returned. In particular, given performance scores on the validation set v_i for iterations $i \in \mathbb{N}$, feature selection is stopped in generation

$$i_{stop} := \min\{i \in \mathbb{N} \mid \forall j \leq \max\{20, 2i\} : v_j \leq v_i\}. \quad (1)$$

That is, optimisation stops if the best performance on the validation set so far has been observed in iteration i, and no improvement is achieved until iteration $2i$.A minimum of 20 iterations is evaluated to allow the feature selection algorithms to 'warm up'.

3. Experiments and Results

We first establish a baseline on the AVEC 2015 database with a machine learning system based on support vector regression (SVR). Feature selection and dimensionality reduction is then performed using the approaches presented above, and evaluated using the same machine learning setup.

3.1. Database

For the purposes of this paper, we choose the 2015 Audio-Visual Emotion recognition Challenge (AVEC 2015) corpus to evaluate the different approaches [3], which is based on the RECOLA database [27]. The challenge corpus contains data from 27 subjects in the RECOLA database, which has been evenly distributed over training, development, and test partitions (9 subjects each). The corpus contains a variety of features representing acoustic, visual, and physiological modalities. It provides the 102-dimensional *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [28], a 84-dimensional set of facial based appearance visual features, and a 316-dimensional set of facial based geometric visual features. Furthermore, it contains two sets of physiological features, namely a 54-dimensional set of *electrocardiogram* (ECG) features and a 62-dimensional set of *electro-dermal activity* (EDA) features.

Note that, similar to Huang et al. [6], we observed the noticeable errors when using the EDA features from one test set participant, we therefore decided to exclude EDA features entirely from our investigation, i. e. the full feature set has a total of 556 features.

3.2. Machine Learning System

We investigate both feature-level (early) fusion, and decision-level (late) fusion in our machine learning setup. In both cases, the selected features are extracted from the training data and standardised as a start.

For feature-level fusion, a single support vector regression machine (SVR) with a linear kernel is then trained on this data. For decision-level fusion, a separate linear SVR is trained on the selected features from each modality. The individual predictions are then fused using a linear regression model. We use one frame out of every twenty for training, in order to reduce computation time.

As in [6], we perform annotation delay compensation, in order to account for the delay between a human annotator's observations and their decisions. We realign the features with the ground truth by dropping the first D frames from the ground truth, and replicating the last frame D times. The optimal delay value D is determined separately per affective dimension, and fusion method. Initially, this results in predictions that are shifted backwards in time by D frames relative to the raw ground truth. Therefore, we replicate the first frame of the predictions D times, and drop the last D frames.

We also apply a post-processing chain similar to that presented in [29]. First, to eliminate high-frequency noise, a median filter with width W is applied to the predictions. Then, the predictions on the training data are centred and scaled to have the same mean and standard deviation as the training gold standard. This transformation is then applied to the validation data.

Finally, the concordance correlation coefficient (CCC)

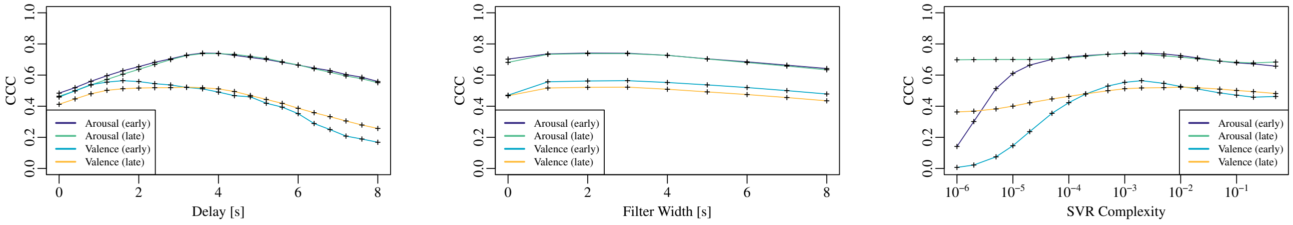
$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (2)$$

between the processed SVR output and the validation data ground truth labels is computed. In the case of wrapper-based feature selection, this value is returned to the feature selection component as the fitness of the supplied feature vector. Note that training and validation data referred to in this section do not necessarily correspond to training and test partitions of the database. For example, feature selection is performed using the training partition as training data, and the development partition as validation data.

3.3. Experimental Settings

The feature selection approaches and the machine learning system have been implemented in Java 1.8, using the WEKA machine learning library (version 3.8.1) [30]. The SVR classifiers are trained using the LibLINEAR library using the L2-regularised L2-loss dual solver with unit bias [31]. While implementing our experiments, we identified serious performance leaks in the LibLINEAR wrapper contained in WEKA (due to instance conversion), and the Java port of LibLINEAR (due to sparse data representation). Thus, we implemented our own LibLINEAR wrapper, and adapted the Java port of LibLINEAR to use a dense data representation. The actual LibLINEAR algorithms have remained entirely unchanged.

3.3.1. Baseline. First, we establish a baseline for further reference, by optimising the temporal shift D, the SVR complexity C, and the post-processing filter width W on the development partition. The temporal shift is evaluated for $0 \leq D \leq 200$ in increments of 10 (0 s to 8 s in increments of 0.4 s), the cost parameter for $C \in [1 \cdot 10^{-6}; 5 \cdot 10^{-1}]$, and the post-processing filter width for $0 \leq W \leq 200$ in increments of 25 (0 s to 8 s in increments of 1 s). For late fusion, the same parameter values are used for each modality, in order to reduce the complexity of our model. We exhaustively evaluated all combinations of these parameters for both affective dimensions and fusion models, the results of which



(a) delay optimisation

(b) filter width optimisation

(c) SVR complexity optimisation

Figure 1: Prediction CCC for different delay values (a), postprocessing filter widths (b), and SVR complexity values (c). For each optimisation parameter, the respective graph shows the maximum prediction CCC on the development partition among all combinations of the remaining two parameters.

TABLE 1: Optimal values for delay D, post-processing filter width W, and SVR complexity C, determined through evaluation on the development partition. Values for delay and filter width are reported in seconds. We use the prediction CCC without any feature selection on the development and test partitions as a baseline for the feature selection algorithms.

					CCC	
dim.	fusion	D	W	C	devel	test
arousal	early	3.6	2.0	$2 \cdot 10^{-3}$	0.742	0.688
	late	3.6	2.0	$1 \cdot 10^{-3}$	0.738	0.675
valence	early	1.6	3.0	$2 \cdot 10^{-3}$	0.564	0.539
	late	3.2	3.0	$1 \cdot 10^{-2}$	0.522	0.537

are shown in Table 1. Furthermore, Figure 1 displays the prediction CCC on the development partition in relation to the parameter values.

3.3.2. Wrapper-based Feature Selection. Subsequently, we perform wrapper-based feature selection with sequential forward selection, and competitive swarm optimisation. During feature selection, models are trained on the training partition, and evaluated on the development partition. For both approaches, we then select the configurations that perform best on the development partition for evaluation on the test partition. Since wrapper algorithms are inherently susceptible to overfitting [15], we also investigate early stopping for SFS and CSO as outlined above (cf. Section 2.6). Three out of the nine development subjects are used as a validation set, and feature selection is terminated if criterion (1) is met.

Although Cheng and Jin report that the performance of CSO does not depend on a large swarm size [18]. We therefore, evaluate CSO on the development partition with different numbers of generations, and particle swarm sizes (cf. Table 2). In order to eliminate one optimisation dimension, we have limited the total number of particle evaluations, i. e. the product of generations and swarm size, to 40 000. illustrates that this choice is sufficiently large, since no large improvements in prediction performance on the development partition are observed as the number of particle evaluations approaches 40 000.

TABLE 2: Prediction performance on the development partition after feature selection with CSO, for different swarm sizes n_P, number of generations n_G, and with or without early stopping. In all cases, the number of total particle evaluations is limited to at most $n_G \cdot n_P = 40\,000$. When using early stopping, we report results at the generation in which CSO was terminated. Without early stopping, we report results at the end of CSO, i. e. when reaching the maximum number of particle evaluations.

			arousal		valence	
early stopping	n_P	n_G	early	late	early	late
yes	100	400	0.812	0.769	0.591	0.554
	200	200	0.783	0.774	**0.601**	**0.576**
	300	133	0.775	**0.785**	0.589	0.556
	400	100	**0.833**	0.784	0.597	0.563
no	100	400	0.859	0.843	**0.713**	**0.675**
	200	200	**0.862**	**0.844**	0.713	0.666
	300	133	0.855	0.833	0.711	0.645
	400	100	0.853	0.823	0.689	0.627

3.3.3. PCA and Filter-based Feature Selection. Furthermore, we evaluate dimensionality reduction using PCA, and filter-based feature selection using canonical correlation analysis. We use WEKA to compute the PCA transformation separately for each modality on the training partition, and apply it to the development and test partitions. The transformed features are then ranked by their variance, i. e. by the amount of variability in the original data they represent. Since PCA transforms instances to a different feature space, delay, post-processing filter width, and SVR complexity have been re-optimised on the development partition (cf. Table 4).

For CCA, we use a Matlab implementation of the CCA feature selection algorithm presented in [13] to rank the features of each modality on the training partition. We then evaluate system performance on the development partition when using the best k features according to PCA, respectively CCA. The optimal k on the development partition is then used to evaluate PCA, respectively CCA, on the test partition (cf. Table 4).

Finally, we evaluate correlation-based feature selection, using WEKA to perform both sequential forward selection and backward elimination with the CFS attribute evaluator

TABLE 3: Development and test partition scores of the configurations that perform best on the development partition for each feature selection approach. Performance is measured in terms of concordance correlation coefficient between the post-processed predictions and the ground truth labels. Additionally, the overall percentage of selected features s_F is reported. We do not show s_F for PCA, since the PCA features are different from the original features, and they can not be compared easily.

	Arousal						Valence					
	early fusion			late fusion			early fusion			late fusion		
model	s_F	devel	test	s_F	devel	test	s_F	devel	test	s_F	devel	test
baseline	–	0.742	0.688	–	0.738	0.675	–	0.564	**0.539**	–	0.522	0.537
PCA	–	0.737	0.674	–	0.742	0.674	–	0.531	0.427	–	0.486	0.471
CCA	80.0	0.751	0.689	95.5	0.748	0.670	89.9	0.570	0.536	94.2	0.538	0.534
CFS (forward)	6.8	0.607	0.643	6.8	0.576	0.597	10.8	0.382	0.395	10.8	0.384	0.406
CFS (backward)	33.1	0.694	0.651	33.1	0.621	0.583	17.4	0.410	0.474	17.4	0.415	0.384
CSO	53.6	0.862	0.692	53.2	0.844	0.676	66.0	0.713	0.512	67.6	0.675	0.503
CSO (early stopping)	49.8	0.833	0.690	51.8	0.785	0.682	79.3	0.601	0.510	88.1	0.576	0.537
SFS	16.0	0.874	**0.714**	22.5	0.870	0.710	23.6	0.756	0.483	17.1	0.745	0.496
SFS (early stopping)	2.5	0.834	0.709	3.6	0.840	0.705	18.2	0.681	0.467	5.6	0.679	0.489

TABLE 4: Results of optimising the number of selected features k on the development partition for PCA and CCA. Since PCA transforms instances to a different feature space, delay D, post-processing filter width W, and SVR complexity C have also been optimised on the development partition prior to optimising k.

		arousal		valence	
model	parameter	early	late	early	late
PCA	D	3.6	4.0	2.0	2.0
	W	2.0	2.0	3.0	1.0
	C	$5 \cdot 10^{-3}$	$1 \cdot 10^{-1}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-6}$
	k	170	169	171	172
CCA	k	445	531	500	524

on the training partition. The resulting feature subset is then used for evaluation on the development, and test partitions. We have also investigated CFS in conjunction with best-first search, but observed no substantial improvements in performance at the cost of higher computation time.

3.4. Results

The results of parameter optimisation on the development partition for both affective dimensions and fusion models are given in Table 1. For arousal, early fusion results achieved a prediction CCC of 0.742 on the development partition, and 0.688 on the test partition. Likewise, early fusion outperforms late fusion for valence, resulting in a prediction CCC of 0.564 on the development partition, and 0.539 on the test partition. The largest difference between early and late fusion can be observed for valence on the development partition, where the early fusion prediction CCC exceeds late fusion by 0.042. Apart from that, early and late fusion actually perform rather similar, with the difference in CCC ranging from 0.002 for valence on the test partition to 0.013 for arousal on the test partition.

The strongest development partition scores, and the corresponding test set scores for each feature selection approach are displayed in Table 3. Wrapper-based feature selection yields consistently good results for arousal prediction, where all investigated wrapper approaches are able to improve the prediction CCC over the baseline. Scores on the development partition range from 0.785 for CSO with early stopping and late fusion to 0.874 with SFS and early fusion, equivalent to an improvement between 0.047 and 0.132 over the baseline. On the test partition, we observe scores between 0.676 for CSO with late fusion, and 0.714 for SFS with early fusion, constituting improvements between 0.001 and 0.026 over the baseline.

PCA, and the filter-based approaches attain lower performance on arousal prediction. On the development partition, performance sometimes even decreases in comparison to the baseline, with scores between 0.576 for forward CFS with late fusion, and 0.751 for CCA with early fusion. Thus, the prediction CCC decreases by as much as 0.162. Only CCA with both late and early fusion, and PCA with late fusion are able to outperform the baseline on the development partition, by up to 0.009. On the test partition, scores range from 0.583 for backward CFS with late fusion to 0.689 for CCA with early fusion. In fact, the latter configuration is the only one able to outperform the baseline, with an improvement of 0.001. The remaining approaches all fall short of the baseline, decreasing prediction CCC from 0.001 for PCA with late fusion to 0.092 for backward CFS with late fusion.

On the development partition, the wrapper-based approaches also perform well for valence prediction. The prediction CCC ranges from 0.576 for CSO with early stopping and late fusion to 0.756 for SFS with early fusion. That is, we achieve improvements between 0.054 and 0.192 over the baseline. However, no approach is able to beat the baseline on the test partition, with performance ranging from 0.467 for SFS with early stopping and early fusion to 0.537 for CSO with early stopping and late fusion. The former score equals a performance decrease of 0.070, whereas the latter matches the baseline for late fusion (but falls short of the early fusion baseline by 0.002).

Finally, PCA and the filter-based approaches once again perform worse than the wrapper-based approaches. The prediction CCC on the development partition ranges between 0.382 for forward CFS with early fusion, and 0.570 for CCA with early fusion. Only CCA is able to outperform the baseline on the development partition, by up to 0.016. The remaining approaches decrease performance by up to 0.182. No approach was able to improve on the test partition baseline, where scores range between 0.384 for backward CFS with late fusion and 0.536 for CCA with early fusion, i. e. performance decreases between 0.003 and 0.153 in comparison to the baseline.

3.5. Discussion

First of all, we observe that we have established strong baselines for both arousal and valence. On arousal, we achieve test set scores that are comparable to the winner of the AVEC 2015 challenge [32], despite using a less complex SVR approach instead of a deep recurrent neural network. On valence, on the other hand, our test set scores are substantially lower than those reported in [32], initially suggesting weaker performance. However, this is due to the fact that we do not use features representing electro-dermal activity in our investigation, which have been shown to be important for valence prediction [4].

3.5.1. Arousal. Sequential forward selection is clearly superior to the remaining feature selection approaches for arousal prediction, achieving the best scores on both the development and the test partition, with just 16 % of the original feature set (cf. Table 3). Furthermore, when combined with early stopping, comparably high performance is attained with only 2.5 % of the original features. This result impressively demonstrates the number of features in the database which are redundant or irrelevant for predicting arousal. In particular, as shown in Figure 2, about half of the features selected by SFS with early stopping are audio features. This constitutes a very high emphasis on acoustic information, since audio features comprise only 18 % of the original feature set. A similarly high focus on acoustic information for arousal prediction has been observed in related literature [3], [4]

Competitive swarm optimisation achieves minor improvements on the test partition while selecting roughly half of the features from each modality. Initially, these results are surprising insofar as CSO carries out a much more intensive search than SFS, yet selects larger feature subsets with worse performance on the test partition. Still, the performance on the development partition is rather similar for CSO and SFS, indicating that they have just selected equally feasible optima. As CSO is guided purely by performance on the development partition, it is not biased toward selecting small feature sets like SFS. Therefore, we suspect that local optima with fewer features have smaller basins of attraction than those with more features, for which reason we can speculate that CSO eventually converges at a local optimum representing a larger feature set.

Correlation-based feature selection, notwithstanding its consistent selection of small feature sets, achieved inferior performance compared to the other approaches. This shows that the score which CFS assigns to feature sets does not accurately reflect their quality for arousal prediction. For instance, we found that the 14 features selected by SFS with early fusion and early stopping are both not maximally correlated with the ground truth labels, and not minimally correlated with each other. Therefore, this feature subset would be suboptimal according to CFS.

Although PCA is unable to outperform the baseline, it still achieves comparable performance on both the development and test partitions with substantially fewer features, since the dimensionality of the feature space is reduced from 556 to 173. The best scores on the development partition are achieved when using almost all PCA features, but fewer than 100 features could be used without notable impacts on performance. Therefore, we still consider PCA a viable option for dimensionality reduction in this setting.

Finally, CCA results in very large feature sets compared to the remaining approaches, without substantial improvements over the baseline. The prediction performance on the development partition rises steadily as more features are included, until around 350 features are selected. Even this smaller feature set would still contain more features than any feature set selected by the other approaches. Furthermore, CCA selects fewer audio features than features from other modalities (cf. Figure 2), which is inconsistent with previous findings on the importance of audio features for arousal prediction [3], [4].

In summary, there is no clear relation between the number of selected features and the development partition score (Pearson, $\rho = 0.13$), or the test partition score (Pearson, $\rho = 0.07$). We do not include PCA features in these calculations, since they reside in a different feature space which can not be compared to the original feature space easily.

3.5.2. Valence. No feature selection approach is able to outperform the test partition baseline on valence prediction, and only few approaches achieve at least comparable performance. In particular, only CCA with early and late fusion, as well as CSO with early stopping and late fusion perform within 0.01 of the baseline, in terms of prediction CCC. All of these approaches select large feature sets containing more than 85 % of the original features. In fact, the number of selected features is mostly uncorrelated with the development partition score (Pearson, $\rho = 0.17$), but it is strongly related to the test partition score (Pearson, $\rho = 0.76$). The latter result indicates that small feature sets generally perform worse than larger ones on the test partition.

Furthermore, while we observe a high correlation between the development and test partition scores for arousal prediction (Pearson, $\rho = 0.89$), this relation is much weaker for valence prediction (Pearson, $\rho = 0.52$). For example, SFS results in substantially improved performance on the development partition, but the test partition score actually decreases. Similar outcomes can be observed for all wrapper-based approaches. Since early stopping does not resolve this problem, and SFS in particular selects very small feature subsets, we are confident that this behaviour is not caused

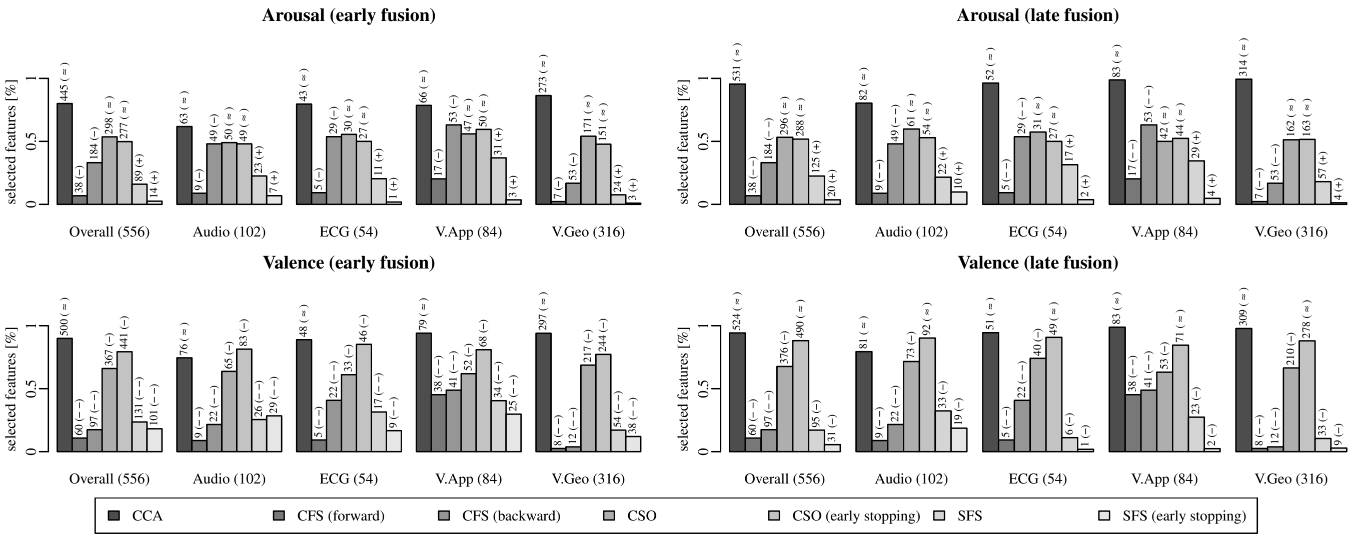


Figure 2: Percentage of selected features (y-axis) from the individual modalities (ECG: electrocardiogram, V.App: video appearance, V.Geo: video geometric) for all feature selection approaches except PCA. The total number of features in each modality is shown in parentheses after the x-axis label, and the number of selected features from the modalities is shown above the individual bars. Furthermore, we indicate the performance of the respective approach relative to the baseline in parentheses above the individual bars. The relative difference is expressed through four categories: approximately equal ($\approx$) if the difference is less than ± 0.01, better ($+$) if it is more than $+0.01$, worse ($-$) if it is less than -0.01, and much worse ($--$) if it is less than -0.05.

by overfitting due to overly intensive search. Nevertheless, we have not been able to identify a compelling reason why we observe such behaviour.

Valence predictions is generally a harder task than arousal prediction, which has been confirmed repeatedly in related literature, where lower prediction accuracy for valence than for arousal has been observed [3], [4], [6]. Our results indicate that in addition to being a harder problem than arousal prediction, valence prediction requires more features as well. For valence, on the other hand, prediction performance increases gradually as more features are included, without any sudden jumps. Moreover, past research has shown that visual information is more important than other modalities for valence prediction [3], [4]. Thus, since the majority of features in the AVEC 2015 database represent visual information, a large fraction of features can be pruned easily for arousal prediction, but not for valence prediction.

4. Conclusions

In this paper, we have performed an in-depth comparative study on dimensionality reduction and feature selection for multimodal continuous emotion prediction. For arousal prediction, our results demonstrate three things: 1) wrapper-based feature selection is highly effective, and clearly outperforms filter-based feature selection and dimensionality reduction; 2) the vast majority of features in the AVEC 2015 database is irrelevant or redundant for arousal prediction; 3) sequential forward selection, in particular, achieves performance comparable to the current state-of-the art while using just 2.5 % of the original feature set. For valence prediction, we have found that feature selection is unable to improve performance over our baseline, partly due to valence prediction being more challenging than arousal prediction, and thus requiring more features. We also plan to utilise dimensionality reduction in our cross-modal representation learning research [33], [34].

Based on the results presented in this paper, we suggest several directions for future research on feature selection in emotion recognition. First, evolutionary algorithms which explicitly optimise the feature set size in addition to prediction performance may improve over the competitive swarm algorithm used in this paper. Second, as there is an abundant supply of unlabelled audiovisual data, unsupervised or semi-supervised feature selection methodologies are worth investigating. Finally, concerning valence prediction, further research is required to clarify our results, and to identify a feature selection approach which yields high performance on unseen data.

5. Acknowledgements

This work was supported by the European Unions's seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu) and No. 688835 (RIA DE-ENIGMA).

References

[1] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wrobel, "Emotion recognition and its applications," in *Human-Computer Systems Interaction: Backgrounds and Applications 3*. Springer, 2014, pp. 51–62.

[2] H. Gunes and B. Schuller, "Categorical and Dimensional Affect Analysis in Continuous Input: Current Trends and Future Directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[3] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AV+EC 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proc. of AVEC'15*. Brisbane, AU: ACM, 2015, pp. 3–8.

[4] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. of AVEC'16*. Amsterdam, NL: ACM, 2016, pp. 3–10.

[5] F. Ringeval, S. Amiriparian, F. Eyben, K. Scherer, and B. Schuller, "Emotion recognition in the wild: Incorporating voice and lip activity in multimodal decision-level fusion," in *Proc. of ICMI*. Istanbul, TR: ACM, 2014, pp. 473–480.

[6] Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu, and J. Epps, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in *Proc. of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC'15, co-located with the 23rd ACM International Conference on Multimedia, MM 2015*. Brisbane, AU: ACM, 2015, pp. 41–48.

[7] M. Schmitt, F. Ringeval, and B. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech," in *Proc. of INTERSPEECH*. San Francisco, CA, US: ISCA, 2016, pp. 495–499.

[8] S. Amiriparian, F. Pohjalainen, E. Marchi, S. Pugachevskiy, and B. Schuller, "Is deception emotional? an emotion-driven predictive approach," in *Proc. of INTERSPEECH*. San Francisco, CA, US: ISCA, 2016, pp. 2011–2015.

[9] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri, "Detecting depression using vocal, facial and semantic communication cues," in *Proc. of AVEC'16*. Amsterdam, NL: ACM, 2016, pp. 11–18.

[10] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, and F. Eyben, "Medium-term speaker states – a review on intoxication, sleepiness and the first challenge," *Computer Speech & Language*, vol. 28, no. 2, pp. 346–374, 2014.

[11] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals," in *Proc. of INTERSPEECH*. Antwerp, BE: ISCA, 2007, pp. 2253–2256.

[12] A. Ivanov and G. Riccardi, "Kolmogorov-smirnov test for feature selection in emotion recognition from speech," in *Proc. of ICASSP*. Kyoto, JP: IEEE, 2012, pp. 5125–5128.

[13] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble cca for continuous emotion prediction," in *Proc. of AVEC'14*. Orlando, FL, US: ACM, 2014, pp. 19–26.

[14] T. Rahman, S. Mariooryad, S. Keshavamurthy, G. Liu, J. H. L. Hansen, and C. Busso, "Detecting sleepiness by fusing classifiers trained with novel acoustic features." in *Proc. of INTERSPEECH*. Florence, IT: ISCA, 2011, pp. 3285–3288.

[15] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[16] R. Bekkerman, M. Bilenko, and J. Langford, *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press, 2011.

[17] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from eeg," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 327–339, 2014.

[18] R. Cheng and Y. Jin, "A competitive swarm optimizer for large scale optimization," *IEEE Transactions on Cybernetics*, vol. 45, no. 2, pp. 191–204, 2015.

[19] S. Gu, R. Cheng, and Y. Jin, "Feature selection for high-dimensional classification using a competitive swarm optimizer," *Soft Computing*, pp. 1–12, 2016.

[20] A. for review, "An 'End-to-Evolution' Hybrid Approach for Snore Sound Classification," 2017, 5 pages, submitted to INTERSPEECH.

[21] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[22] M. A. Hall, "Correlation-based feature selection of discrete and numeric class machine learning," 2000.

[23] J. Deng, N. Cummins, J. Han, X. Xu, Z. Ren, V. Pandit, Z. Zhang, and B. Schuller, "The university of passau open emotion recognition system for the multimodal emotion challenge," in *Proc. of CCPR*. Chengdu, CN: Springer, 2016, pp. 652–666.

[24] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[25] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *IEEE International Conference on Neural Networks*. Perth, AU: IEEE, 1995, pp. 1942–1948.

[26] J. Reunanen, "Overfitting in making comparisons between variable selection methods," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1371–1382, 2003.

[27] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Proc. of Automatic Face and Gesture Recognition*. Shanghai, CN: IEEE, 2013, pp. 1–8.

[28] F. Eyben, K. R. Scherer, B. Schuller, J. Sundberg, E. Andr, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[29] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. of ICASSP*. Shanghai, CN: IEEE, 2016, pp. 5200–5204.

[30] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[31] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[32] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proc. of AVEC'15*. Brisbane, AU: ACM, 2015, pp. 73–80.

[33] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller, "Snore Sound Classification Using Image-based Deep Spectrum Features," in *Procc. of INTERSPEECH*. Stockholm, Sweden: ISCA, August 2017, 5 pages.

[34] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech," in *Proc. of the 25th ACM International Conference on Multimedia, MM 2017*. Mountain View, CA: ACM, October 2017, 7 pages.