Ensemble Methods for Affective Recognition

Rafooneh Jafarian Bahri

Montreal, Québec, Canada

SYS843
Neural Network and
Fuzzy Systems

1-Introduction

- Emotion recognition is an actively emerging area of research due to its vast application in many areas such as in business, education, detecting deception, brain-based diseases and etc..
- Basic research on human emotions assist people in many professions requiring face-to-face interactions improving their skills in reading the emotions of others.

- The overarching goal is the prediction of emotion in multidimensional emotion space i.e. Arousal and Valence .
- We aim to compare the state of the art technique for recognition of expression based on RECOLA audio.
- We are interested in assessment of the impact of dynamic selection and LSTM.

2-Technique

- In this work we applied two methods in ensemble : Multiple homogenous and heterogenous regressors without selection (case 2), and dynamic selection using Euclidean and cosine distance with pool of homogenous regressor. (case 1)

Case 1:

- Dynamic Selection: a pool of regressors is generated using each annotation label. Using NN-rule in testing step with cosine distance as similarity measure, we selected the best regressor for each query in test set.

Case 2:

- Approach I: For each query, we made predictions with all regressors in the pool, then we trained a SVR as fusion decision to learn from prediction from training set and the true labels, and predict the predictions results from test set
- Approach II: Similar to approach I, yet instead of using predictions from validation set, we use 9-fold to create a block of prediction using the predictions from test set
- Approach III: In this approach we take an average between the predictions made from LSTM and SVR.

3-Methodology

- **Preprocessing:** we rescale the range of features to scale the range in [0, 1], (Zero One Normalization):

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- temporal shift is applied for each file in the dataset due to annotation delay (Delay compensation).

- **Learning algorithms** used are SVR ε-insensitive loss function with linear kernel and Long short-term Memory architecture in recurrent neural network.
- **Postprocessing:** prediction are normalized using standard deviation ratio between the prediction and labels.

- For each query sample x_i in prediction vector, we multiply it by the ratio below:

$$ratio = \frac{\sigma(Prediction)}{\sigma(Label)}$$

- Median filtering is optimized for the dataset and applied to smooth the prediction:
- A median filter of size n on a sequence $\{x_i,\ i \in \mathbb{Z}\}$:

$$y_i = Median x_i \triangleq Median(x_{i-v}, \dots, x_i, \dots, x_{i+v})$$

$$v = \frac{n-1}{2}$$

4-Experimental set up

- **Database**: The dataset RECOLA consist of five signals i.e. audio, EDA, ECG, video geometric and video appearance from 27 subjects.
- We used audio features in this work and the labels for arousal.
- The annotation of dataset is performed by six gender balanced French-speaking assistants.
- The baseline uses Gold Standard which is average of all ratings for each subject. We use all 6 rating to generate separate models for ensembles.

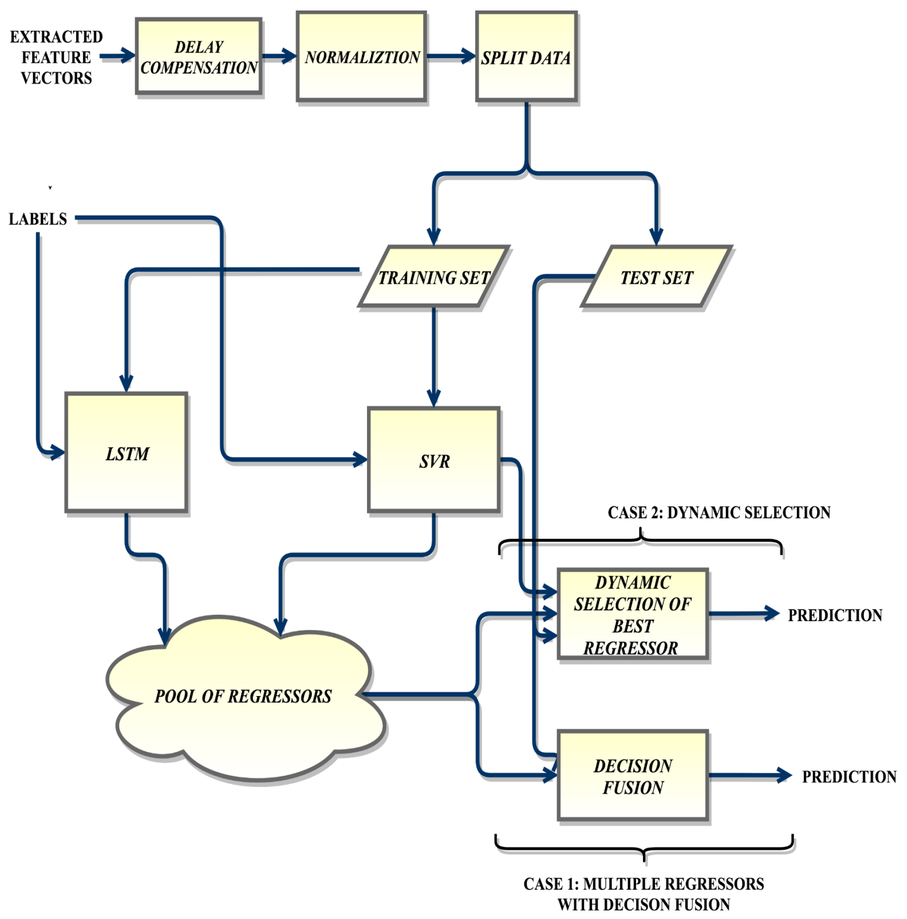
- **Performance Measure:** The performance measure adopted is the concordance correlation coefficient (CCC), which combines the Pearson correlation coefficient (ρ) and the mean square error (MSE).

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

- **Key experimental setting**: The parameter C and ε for each SVR regressor are optimised using K-fold cross-validation. The SVRs are trained with same training sample but different annotation label providing 6 regressors.

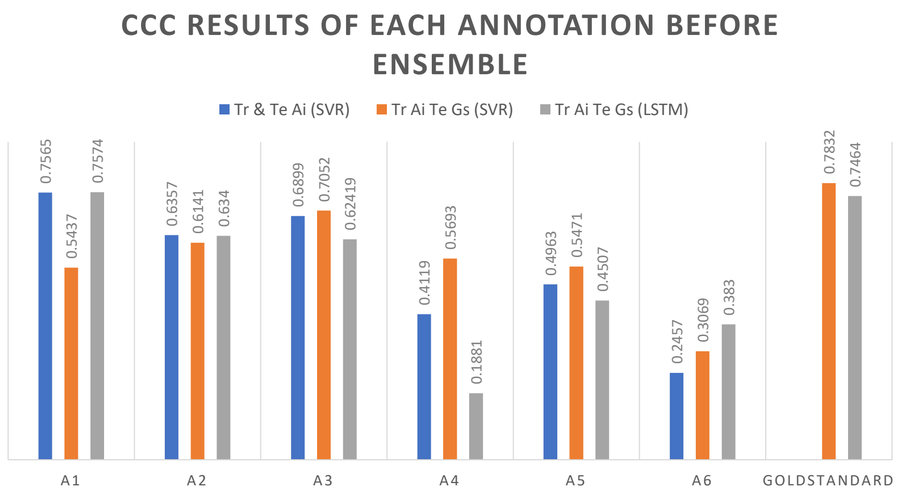
- For LSTM, we optimised the number of neurons to 100 units and transformed the dataset into time series with optimized window size 30. Lastly, we added a 100 unit layer before output layer. The optimized learning rate was 0.01, momentum 0.9. The most effective optimization method used is Dropout. We chose 0.5, giving each neuron 50% chance of drop out.

5-Experimental flow chart



6-Experimental Results:

- The chart below depicts the results before the decision fusion. We can see the difference between the models with annotation labels separately compared with the models with gold standard. We can also notice the different results from two algorithms, SVR and LSTM.



7-Ensemble Results:

- In th following table we decpit the results from case senario I and II.

Algorithm	Method	CCC
Linear SVR	Dynamic Selection (k=2)	0.6712
LSTM+DNN	Train and Test with Gold standard	0.7464
Linear SVR	SVR as decision fusion (train on test prediction)	0.7689
SVR+LSTM+DNN	SVR as decision fusion (train on train prediction)	0.6605
SVR+LSTM+DNN	SVR as decision fusion (trained on test prediction)	0.7784
SVR+LSTM+DNN	Average as decision fusion	**0.8050**
SVR	Dynamic selection Oracle	**0.9450**

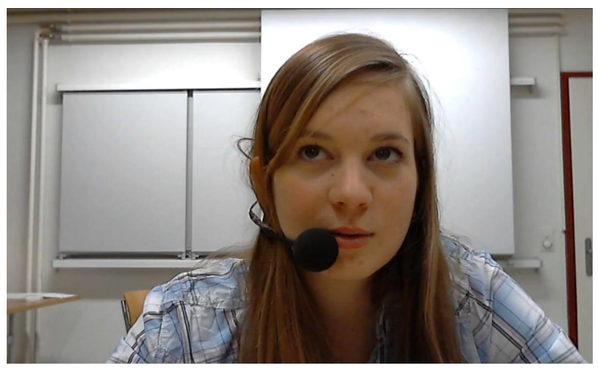
- **Dynamic selection (case 2)**:

- We obtained the first result based on dynamic selection, that use cosine distance as similarity measure between the query and validation set. Using Oracle for dynamic selection, we obtained 0.945 Concordance Correlation Coefficient. This means that for future work we need to adopt a different distance metric measure since we can achieve higher CCC than baseline with oracle.

- **Ensemble, decision fusion (case 1):**

- Approach I (4th row) compared to Approach II achieved lower CCC. This is because in approach II we tried to find the upper bound of decision fusion technique we adopt. (3rd row shows the result using only SVR regressors from the pool)

- In Approach III, the result is promising compared to the rest of approaches.



8-Conclusion:

- In this work our objective is to find a robust system for dynamic emotion prediction as well and ensembles in arousal dimension. We generated several models using annotator's label and different machine learning algorithm, SVR and LSTM.

- The interpretation of the results shows that using different annotator's label can result in a very high accurate prediction as we achieved 0.945 CCC for the oracle. However, using dynamic selection with Cosine distance as similarity measure does not perform well as expected. One can see, that the change in distance metric learning may result in higher accuracy.

- LSTM and SVR are complementary in sense that we achieve the highest score by the average all predictions made by SVR and LSTM regressors. This supports the idea of using different annotation label to diversify generated regressors.