MediaEval 2014: THU-HCSIL Approach to Emotion in Music Task using Multi-level Regression [*]

Yuchao Fan, Mingxing Xu
Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
fyc12@mails.tsinghua.edu.cn, xumx@tsinghua.edu.cn

ABSTRACT

This working notes paper describes the system proposed by THU-HCSIL team for dynamic music emotion recognition. The procedure is divided into two module - feature extraction and regression. Both feature selection and feature combination are used to form the final THU feature set. In regression module, a Booster-based Multi-level Regression method is presented, which outperforms the baseline significantly on test data in RMSE metric for dynamic task.

1. INTRODUCTION

The objective of *Emotion in Music* task in MediaEval 2014 Workshop is to predict 2-dimensional emotion of music for both the whole clip (static) and the sequenced 0.5-second segments (dynamic). For more details about the task and data set, see the task overview paper [1].

2. FEATURE SELECTION AND COMBINATION

This module is based on the 6552-dimensional EMO-LARGE features provided by organizers. These features are extracted by OpenSmile Toolbox [2] and are calculated by low-level features (LLDs) such as MFCC, Spectral, ZCR, loudness. Details of LLDs and statistic functions can be found in [2].

Since the redundancy and insufficiency of the base features, we use both feature selection and feature combination to improve our feature set.

2.1 Feature Selection

Firstly, two algorithms were used for feature importance ranking. The first is Random Forest [3], which constructs a multitude of decision trees at training time and the importance of feature can be computed through permutation over all trees. The second is Extremely Random Trees (Extra-trees) [4], another type of random trees with different node-splitting strategy. Let F_n^{RF} denote the first n important features in Random Forest and F_n^{ET} in Extra-trees.

Then, we define the selected features as $F_n^{SLT} = \{F_n^{RF} \cap F_n^{ET}\}$. Using XGBoost lib [5], the optimized n^* is determined according to the RMSE performance of F_n^{SLT}. In our work, 661-dim F_n^{SLT} is obtained for arousal when $n^* = 1280$ and 522-dim for valence when $n^* = 1280$.

Table 1: Composition of the supplement features.

Tool	Feature	Dim
MIRToolbox [7]	mode, inharmonicity, flatness, centroid, brightness, entropy, kurtosis, rolloff, roughness, spread, skewness, regularity, zerocross, key1, key2	15
Bregman [8]	chroma	12
	log-frequency spectrum	95
	low-frequency spectrum	95

2.2 Feature Combination

Some music-related features, such as mode, timbre, are not included in the EMO-LARGE features, while they are proved to be important in music emotion recognition [6]. Thus, we use two toolboxes - MIRToolbox [7] and Bregman music toolbox [8] - with default settings to extract such additional features for our task.

For each 0.5 second segment, 217 features are extracted in total (see Table 1). Then we combine these supplement features with the selected features to form our final THU feature set, 878 features (661+217) for arousal and 739 features (522+217) for valence.

3. MULTI-LEVEL REGRESSION

With the assumption that the regression accuracy can be improved by reducing the scope of model, we propose a 2-level regression method in this section. Figure 1 shows the framework of this method.

The first-level regression is just a naive procedure commonly used in regression module. For x in test set $\mathbf{X}$, we predict the test result $r_0 = f(x, R^G)$ by model R^G, a global model trained with entire training dataset $\mathbf{X}^{Tr}$.

For each emotion dimension, we divide $[-0.75, 0.75]$ (based on the ground truth distribution of $\mathbf{X}^{Tr}$) into 6 bins with equal length (0.25). For x in $\mathbf{X}$, its bin index $i = Ind(x) = \lceil (f(x, R^G) + 0.75)/0.25 \rceil$. The 2nd-level regression result of x

[*]This work was partially supported by the National Natural Science Foundation of China (No.61171116) and National Basic Research Program of China (973 Program, No.2012CB316401).

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

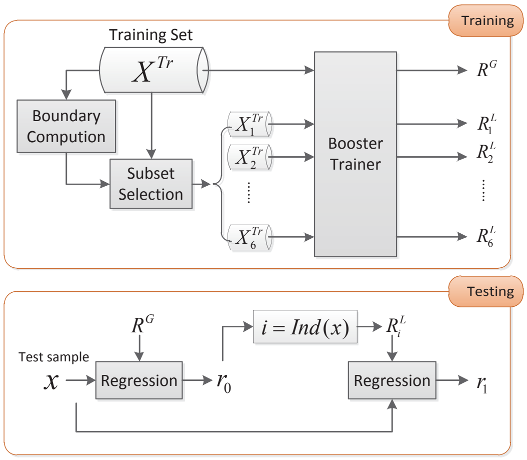

Figure 1: Process framework for multi-level regression system.

Table 2: Regression results with different models. ET = Extra-Trees, RF = Random Forest.

Model	Arousal		Valence		Tool
	MAE	RMSE	MAE	RMSE	
Linear	0.159	0.203	0.158	0.199	WEKA
Pace	0.153	0.195	0.154	0.197	
M5Rule	0.154	0.195	0.159	0.205	
ET	0.150	0.188	0.149	0.191	scikit-learn
RF	0.142	0.177	0.143	0.183	
Booster	**0.141**	**0.176**	**0.138**	**0.178**	XGBoost

is $r_1 = f(x, R_i^L)$, where R_i^L is the local model for the i^{th} bin trained by the subset of training dataset $\mathbf{X}_i^{Tr}$, which is determined by $\mathbf{X}_i^{Tr} = \{x|x \in \mathbf{X}^{Tr}\ and\ g(x) \in [a_i, b_i]\}$, where $g(x)$ is the ground truth of x and the boundaries $\{a_i, b_i\}$ are computed on dev set $\mathbf{X}^D$ as follows:

We first define $G_i = \{g(x)|Ind(x) = i, x \in \mathbf{X}^D\}$. After investigating the distribution of G_i, we then select $a_i = min\{g|CDF_i(g) \geq 1.5\%\}$ and $b_i = max\{g|CDF_i(g) \leq 98.5\%\}$, where CDF_i is the cumulative distribution function for G_i, to eliminate outliers and fix the confidence level of dev set $\mathbf{X}^D$ at 97%.

4. RESULTS AND CONCLUSIONS

Results in this section are all modeled on THU feature set and experiments are all for dynamic task.

4.1 Experiment Results

5-fold cross validation is used and is song independent (separated by song IDs randomly). 1/5 of training set is used as dev set (as mentioned in section 3).

Multiple regression algorithms were compared (Table 2). We choose tree booster in XGBoost for the following experiments, in which max tree-depth $\tau = 4$, step size shrinkage $\eta = 0.1$ and minimum loss reduction $\gamma = 1.0$.

Table 3 illustrates that the performance of multi-level regression system is improved slightly in both arousal and valence case.

4.2 Results on Required Test Data

Table 3: Comparison of RMSE result between naive regression and multi-level regression.

Regression Strategy	RMSE	
	Arousal	Valence
One-level	0.176	0.178
Multi-level	**0.175**	**0.177**

Table 4: Results on required test data.

Task	Run	Arousal		Valence	
		APC	RMSE	APC	RMSE
Dynamic	baseline	**0.180**	0.270	**0.110**	0.190
	1	0.128	0.124	0.064	0.100
	2	0.127	0.125	0.074	0.099
	3	0.170	**0.119**	0.091	0.094
	4	0.167	0.120	0.098	**0.094**
Static	5	0.770	0.107	0.459	0.097

Each clip from test data is split into 90 0.5-second segments and predicted in segment level. Table 4 shows the results of 5 runs submitted, run 1-4 are for dynamic task and run 5 is for static task. XGBoost lib is employed for regression in all runs.

- Dynamic task (subtask 2):

Run 1. One-level regression.

Run 2. Two-level regression.

Run 3. Run 1 + smooth. For one clip, we replace segment result in Run 1 with the average of 4 adjacent segment results and itself as following:

$$r^{run3}(i) = mean(r^{run1}_{i-2}, ..., r^{run1}_{i+2}), i = 3, 4, ..., 88.$$

Run 4. Run 2 + smooth.

- Static task (subtask 1):

Run 5. For each clip, average over 90 segment results from Run 2.

Result shows that our result outperforms baseline on RMSE metric. The lower APC results might have been as a result of not considering the correlation between adjacent segments. Performance of Run 3 and 4 on APC shows that the smooth module can increase the inter-clip correlation effectively.

5. REFERENCES

[1] A. Aljanaki, Y-H. Yang, M. Soleymani. Emotion in Music Task at MediaEval 2014. In *MediaEval 2014 Workshop*, Barcelona, Spain, October 16-17 2014.

[2] Eyben, Florian, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia.* ACM, 2010.

[3] Breiman, Leo. Random forests. In *Machine learning* 45.1 (2001): 5-32.

[4] Geurts, Pierre, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. In *Machine learning* 63.1 (2006): 3-42.

[5] https://github.com/tqchen/xgboost/

[6] Yang, Yi-Hsuan, and Homer H. Chen. Machine recognition of music emotion: A review. In *ACM Transactions on Intelligent Systems and Technology (TIST)* 3.3 (2012): 40.

[7] Lartillot, Olivier, and Petri Toiviainen. A Matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects.* 2007.

[8] http://bregman.dartmouth.edu/bregman/