UNIZA System for the "Emotion in Music" task at MediaEval 2015

Michal Chmulik, Igor Guoth,
Miroslav Malik, Roman Jarina
Department of Telecommunications
and Multimedia, Faculty of Electrical
Engineering, University of Zilina
Zilina, Slovakia
michal.chmulik@fel.uniza.sk

ABSTRACT

In this working notes paper, we present the UNIZA system for the recognition of dynamic music emotional dimensions arousal and valence. The developed system is based on Support Vector Regression with Radial Basis kernel function. We selected 2 sets of features using stochastic evolutionary optimization algorithms namely Genetic Algorithm and Particle Swarm Algorithm. The models score the average Root Mean Square Error 0.3605 for the valence dimension and 0.2540 for the arousal dimension.

1. INTRODUCTION

The objective of the Emotion in Music task at MediaEval 2015 is to automatically determine temporal dynamics in emotion as a sequence of numerical values in two dimensions: valence and arousal (AV). The task comprises three scenarios: 1) Given a set of baseline audio features, the participants are to return AV scores obtained by machine learning method of their choice; 2) The participants are required to submit their own set of features, they believe that are most discriminative in term of emotion determination; 3) The participants may return AV scores obtained by using any combination of the features and machine learning method. For more details about the task and data set, see the overview paper [1].

2. APPROACH

UNIZA system for the dynamic emotion recognition is based on the Support Vector Regression (SVR) and utilizes the LIBSVM libraries. We follow the approach that we have already applied for emotion recognition from speech [2]. Development of our system has been carried out in Matlab and C++ environments. We have split the development data into 2 approximately equal non-overlapping parts - the first one for the training of regression models while the other part for models testing.

SVR has employed the Radial Basis (RBF) kernel function. Search for optimal kernel parameters has been performed by the grid search method in cooperation with Bat Algorithm (BA) - metaheuristic optimization technique [3]. The parameters of the kernel were individually optimized for the both dimensions and finally selected the one combination resulting in the best evaluation accuracy. The same kernel parameter values were used in all scenarios of the task.

For the second and third scenarios, we have created 2 sets of features, which are selected from the baseline feature set using stochastic evolutionary optimization algorithms. For this purpose, we have used hybrid combination of Genetic Algorithm (GA) and Particle Swarm Optimization algorithm (PSO) [4]. The GA/PSO hybrid approach works as follows. The both optimization algorithms run in parallel and at the end of each iteration, the best individuals from the both algorithms are selected into the next iteration of the optimization process. The Root Mean Square Error (RMSE) between the predicted and ground truth labels was used as a fitness function for the optimization algorithms. The optimization process was running in 50 iterations and repeated 50 times. Two best combinations of features have been selected for the submission. The first set denoted as "*optimal_1*" consists of 139 features and set denoted as "*optimal_2*" consists of 129 features. Both sets include 72 identical features - detailed description is beyond the limit of paper pages but the sets intersection contains mostly a number of auditory spectra coefficients, MFCC coefficients as well as their spectral skewness, slope, flux and delta regression variations.

In the system development stage, we also tested the features extracted by the MIRToolbox [5] - combination of chromagram, onset detection, log-attack time, roughness, tempo, key and tonality. The feature extraction process has been performed on frames with different duration and overlapping depending on the particular feature. As a result, we have obtained 51 features and this set is denoted as "*MIR*". We have used identical feature format as the baseline (non-overlapping segments of 500 ms) and besides the mean values and standard deviations, we have also used the maximal values.

Table 1 states mean evaluation accuracy that we have obtained for the development data using evaluation metrics RMSE [1] and Pearson's correlation coefficient *r*. The "*default*" run represents the first scenario of the task when our system was fed with the baseline feature set. The other runs corresponds to the third scenario where different feature sets were tested with our regression models. Based on acquired preliminary results, we decided to further process and submit only feature sets with the highest ranking (e.g. "*optimal_1*" and "*optimal_2*").

Table 1. Results of the system on the development data for different feature sets.

	Arousal		*Valence*	
run	**RMSE**	***r***	**RMSE**	***r***
default	0.0815	0.4203	0.0724	0.4238
optimal_1	0.0806	0.4553	0.0669	0.4603
optimal_2	0.0794	0.4681	0.0659	0.4718
MIR	0.0988	0.2058	0.1044	0.2104
def.+MIR	0.0887	0.3543	0.0771	0.3709

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany.

3. RESULTS AND DISCUSSION

In Table 2, there is notified the official classification accuracy of our system according to the evaluation metrics of the task [1] for the first scenario ("*default*" run) and the third scenario of the task ("*optimal_1*", "*optimal_2*").

Table 2. Official results of UNIZA team for different runs.

Valence

run	**RMSE**	***r***
default	0.3662±0.1747	-0.0218±0.4011
optimal_1	**0.3605±0.1727**	**-0.0141±0.4007**
optimal_2	0.3613±0.1737	-0.0161±0.3961

Arousal

run	**RMSE**	***r***
default	0.2554±0.0995	0.5100±0.2248
optimal_1	0.2571±0.0997	0.5097±0.2228
optimal_2	**0.2540±0.1028**	**0.4930±0.2326**

As it can be seen, our feature set did not provide any significant improvement of the system efficiency in comparison with the baseline feature set and the differences in RMSE are barely noticeable. Anyhow, the best results represent the "*optimal_1*" run for the valence dimension and the "*optimal_2*" run for the arousal dimension. The arousal dimension acquires better results than the valence dimension as is usual in the emotion recognition tasks. In comparison with the results from the development data, there can be seen a huge drop of the correlation coefficient *r* for the valence dimension.

Although our feature sets do not achieve significantly better score, feature dimension of the sets are greatly reduced (approximately 50% of the baseline) thus the computational demands of the system is also greatly reduced. Based on the results, it seems that there is a great redundancy of data in the baseline feature set. The system efficiency may be improved by finer tuning of the kernel parameters individually for each dimension. Also, application of other regression model could improve the system accuracy.

In the future, we would like to create a vast set of features (baseline + "*MIR*" and other musical-oriented features) and search for the optimal subset giving the best classification accuracy that would be at least equal baseline/full set accuracy. For the searching process, some of the state-of-the-art nature-inspired optimization technique will be applied.

4. CONCLUSION

We developed the SVR-based system for dynamic music emotion recognition. Regrettably, our feature sets suggested according to the evolutionary optimization methods did not cause significant improvement of classification accuracy of the system. On the other hand, the (almost) equal result were obtained using only approximately 50% of the baseline features.

5. REFERENCES

[1] Aljanaki A., Yang Y.-H., Soleymani M. 2015. Emotion in Music Task at MediaEval 2015. In *MediaEval 2015 Workshop*, 2015, Wurzen, Germany.

[2] Hric, M.; Chmulik, M.; Guoth, I.; Jarina, R. 2015. SVM based speaker emotion recognition in continuous scale. In *Proceedings of 25th International Conference Radioelektronika 2015*, 2015, Pardubice, Czech republic, 339-342.

[3] Yang X.-S. 2014. Nature-Inspired Optimization Algorithms. *Elsevier*, London.

[4] Kennedy J., Eberhart R.C. with Shi Y. 2001. Swarm Intelligence. *Morgan Kaufmann Publisher*, San Francisco.

[5] Lartillot O., Toiviainen P. 2007. A Matlab Toolbox for Musical Feature Extraction From Audio. In *International Conference on Digital Audio Effects*, 2007, Bordeaux, France, 237-244.