Predicting Affect in Music Using Regression Methods on Low Level Features

Rahul Gupta, Shrikanth Narayanan
Signal Analysis and Interpretation Lab (SAIL), University of Southern California, Los Angeles, CA, USA
guptarah@usc.edu, shri@sipi.usc.edu

ABSTRACT

Music has been shown to impact the affective states of the listener. The emotion in music task at the MediaEval challenge 2015 focuses on predicting the affective dimensions of valence and arousal in music using low level features. In particular, this edition of the challenge involves prediction on full length songs given a training set containing smaller 30 second clips. We approach the problem as a regression task and test several regression algorithms. We proposed these regression methods on the dataset from previous edition of the same task (Mediaeval 2014) involving prediction on 30 second clips instead of full length songs. Through evaluation on the 2015 data set, we obtain a point of reference for the model performances on longer song clips. Whereas our models perform relatively well in predicting arousal (root mean square error: .24), we do not obtain good results for valence prediction (root mean square error: .35). We analyze the results and the experimental setup and discuss plausible solutions for a better prediction.

1. INTRODUCTION

Music is an important part of media and considerable research has gone into understanding and indexing the music signal [1, 2]. Music has been shown to impact the affective states of the listeners and in depth analysis of the relation between music and affect can impact both understanding and design of music. Over the past few years, the emotion in music task at various MediaEval challenges [3, 4, 5] has provided a unified platform for understanding the affective characteristics of music signals. The emotion in music task at MediaEval 2015 [5] provides a training set which is a subset of the 2014 challenge, with valence and arousal annotations over 30 second clips. This subset is chosen for better quality annotations as described in the overview paper [5]. However, it is also unique in the sense that the prediction has to be made on a test set containing full length songs. This poses the challenge of generalizing models trained over smaller music segments for prediction on longer segments.

In this work, we present the results on affect prediction in music using our previous models developed on the 2014 challenge data set. We tested multiple regression models followed by a smoothing operation in last year's challenge [6] and more recently developed a Boosted Ensemble of Single feature Filters (BESiF) algorithm [7] for affect prediction in music. In general, the affective signals evolve smoothly over time and do not undergo abrupt changes. Our models take this factor into account by learning the mapping from features to the affective dimensions while also accounting for the smooth temporal evolution of affect. In the 2015 emotion in music task, our best models obtain a root mean square error values of .35 and .24 in valence and arousal prediction, respectively. In the next section we describe our methodology in detail.

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

2. METHODOLOGY

The 2015 challenge task provides a development set consisting of 30 second clips from 431 songs; annotated at a rate of 2 frames per second. The baseline feature set is extracted using OpenSmile [8] and contains 260 features. The test set contains 58 full length songs annotated at the same frame rate as the development set. We use three different regression methods to predict the affective dimensions of valence and arousal from the 260 baseline features. We describe these methods below.

2.1 Linear Regression + Smoothing (LR+S)

In this model, we use the 260 features and learn separate linear regression models to predict arousal and valence. After obtaining the decisions, we perform a smoothing operation by low pass filtering the frame-wise arousal and valence values. We use a moving average filter as the low pass filter with filter length tuned using three fold inner cross validation on the train set (arousal filter length = 13; valence filter length = 38). The smoothing operation not only removes the high frequency noise, but also incorporates the local context into account while making decision for a frame. The decision for a frame is given as an unweighted combination of frame values in a window centered around that frame, thereby incorporating local context.

2.2 Least Squares Boosting + Smoothing (LSB+S)

Least squares boosting [9, 10] is another regression algorithm trained using gradient boosting [9]. We use the "fitensemble" function in Matlab to train a least squares boosting model for predicting valence and arousal. The base learners used for least squares boosting are regression trees [11]. The number of regression trees in the ensemble is tuned using 3 fold cross-validation on the train set. After obtaining the frame-wise decisions from the least squares boosting algorithm, we perform a smoothing operation as explained in the section 2.1.

Method	Valence		Arousal	
	RMSE	ρ	RMSE	ρ
Baseline [5]	0.37	0.01	0.27	0.36
LR+S	0.35	0.01	0.24	0.65
LSB+S	0.35	0.05	0.24	0.59
BESiF	0.37	-0.04	0.28	0.50
Unweighted summation	0.35	0.00	0.24	0.64

Table 1: Results on valence and arousal prediction using the proposed regression systems.

2.3 Boosted Ensemble of Single feature Filters (BESiF)

We proposed another gradient boosting based algorithm on the 2014 emotion in music data set [7]. In this algorithm, we propose the base learners to be filters (analogous to regression trees used in LSB+S algorithm). The motivation behind this algorithm was to perform a joint learning of regression and smoothing unlike previous two methods. The filters not only learn the mapping between low level features and the affective dimensions, but also perform temporal smoothing. A detailed description of the training algorithm can be found in [7].

2.4 Unweighted combination of LS+S, LSB+S and BESiF algorithms

Our final model was an unweighted combination of the previous three models. Unweighted combination of models have been shown to help prediction if and when models capture complementary information from the features [12, 13]. In the next section, we present our results and analysis.

3. RESULTS AND DISCUSSION

We show the results from the four models presented above in Table 1. From the results, we observe that our approach using regression fails for valence prediction with close to no correlation with the ground truth. As this was not the case for at least the LR+S system in the previous edition of the challenge (MediaEval 2014 [4]), we suspect that there are inherent differences in the data sets from MediaEval 2014 and 2015. As previously pointed out, this year's challenge involved prediction over full length song segments with training on 30 second clips. This poses a data mismatch problem, particularly with respect to our BESiF algorithm. The filters in the algorithm are optimized over shorter time series whereas test set prediction is over longer time series.

In case of arousal, our systems perform relatively well. The linear regression system performs the best. The BESiF algorithm again fails to perform better than the other algorithms primarily because of the data mismatch problem. The filters in the BESiF algorithm when trained on smaller duration annotation time series may not capture the dynamics that can exist over longer duration annotations. The success of linear regression in arousal prediction offers some promise in case of problems involving such temporal mismatch between train and test set. In the next section, we talk about modifying our current approach to improve the results.

4. FUTURE WORK

Given that our systems do not perform well for valence prediction, we aim to perform a detailed analysis to understand the reasons behind the poor performance. Despite the presence of features correlated with valence in the train set and our success in the last edition of the challenge, a low performance on valence prediction poses a challenge in form of understanding prediction over longer song segments. We suspect that providing annotators with small song segments versus longer segments may have an impact on the annotation itself. Listening to longer clips may alter affective perceptions and introduce other annotator biases. In particular, we aim to investigate the performance of our BESiF algorithm and modify for the given problem setting. This may involve including adaptation schemes [14, 15] to model differences in annotation over the train and the test set and other mismatch that may exist.

Also several previous works have reported differences in performances for arousal and valence prediction using acoustic features similar to the ones used in this work [16, 17, 18]. This is worth investigating into as it may imply that valence prediction may involve other features not considered in the baseline set of features. In the case of continuous emotion tracking involving human interaction, video modality has been shown to add complementarity and even outperform audio signals [18, 19, 20]. This poses a very interesting problem for the valence prediction in music as emotion annotations are made using music audio only. Whereas there can exist videos for certain songs, it has not been investigated if videos can be associated with and even alter the perceived affective evolution of the song. Along similar lines, several works propose the use of song lyrics in predicting affect [21, 22]. Hence textual content of the song can also be incorporated towards the development of an enhanced multi-modal affect prediction system.

5. CONCLUSION

In this work, we use several previously proposed regression methods on the emotion in music task at MediaEval challenge 2015. We note that despite our success in the previous edition of the challenge, our methods fail, particularly for valence prediction. Our methods perform relatively well for arousal prediction, however the trends in performance across models are not as expected. We suspect that there could be several reasons for the unexpected results. Primarily, the differences in lengths of the train and test sets could lead to a mismatched model for test set prediction. We also suspect that it may cause differences in perception of affect in music, leading to differences in affect annotation.

Instead of providing answers to relation between low level features and affective dimensions, our work in this paper opens up more questions regarding the affective evolution of music signal. With regards to the future work, differences in perception of short clips of music signal versus longer clips, differences between affective dimensions of valence and arousal with regards to model development and investigating algorithmic designs will be our initial steps.

6. REFERENCES

[1] Mira Balaban, Kemal Ebcioglu, and Otto E Laske. Understanding music with ai: perspectives on music cognition. 1992.

[2] Michael Tanner and Malcolm Budd. Understanding music. *Proceedings of the Aristotelian Society, Supplementary Volumes*, pages 215–248, 1985.

[3] Mohammad Soleymani, Michael Caro, Erik Schmidt, and Yi-Hsuan Yang. The mediaeval 2013 brave new task: Emotion in music. In *MediaEval 2013 Workshop, Barcelona, Spain*, 2013.

[4] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. Emotion in music task at mediaeval 2014. In *MediaEval 2014 Workshop, Barcelona, Spain*, 2014.

[5] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. Emotion in music task at mediaeval 2015. In *MediaEval 2015 Workshop, Wurzen, Germany*, 2015.

[6] Naveen Kumar, Rahul Gupta, Tanaya Guha, Colin Vaz, Maarten Van Segbroeck, Jangwon Kim, and Shrikanth S Narayanan. Affective feature design and predicting continuous affective dimensions from music. In *MediaEval Workshop, Barcelona, Spain*, 2014.

[7] Rahul Gupta, Naveen Kumar, and Shrikanth Narayanan. Affect prediction in music using boosted ensemble of filters. In *The 2015 European Signal Processing Conference, Nice, France*, 2015.

[8] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM, 2010.

[9] Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[10] Gene H Golub and Charles F Van Loan. An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis*, 17(6):883–893, 1980.

[11] Jane Elith, John R Leathwick, and Trevor Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813, 2008.

[12] Thomas G Dietterich. Ensemble methods in machine learning. In *Multiple classifier systems*, pages 1–15. Springer, 2000.

[13] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[14] George Foster and Roland Kuhn. Mixture-model adaptation for smt. 2007.

[15] WA Ainsworth. Mechanisms of selective feature adaptation. *Perception & Psychophysics*, 21(4):365–370, 1977.

[16] Mihalis Nicolaou, Hatice Gunes, Maja Pantic, et al. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *Affective Computing, IEEE Transactions on*, 2(2):92–105, 2011.

[17] Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth Narayanan. Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing*, 31(2):137–152, 2013.

[18] Rahul Gupta, Nikolaos Malandrakis, Bo Xiao, Tanaya Guha, Maarten Van Segbroeck, Matthew Black, Alexandros Potamianos, and Shrikanth Narayanan. Multimodal prediction of affective dimensions and depression in human-computer interactions. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 33–40. ACM, 2014.

[19] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2014.

[20] Vikramjit Mitra, Elizabeth Shriberg, Mitchell McLaren, Andreas Kathol, Colleen Richey, Dimitra Vergyri, and Martin Graciarena. The sri avec-2014 evaluation system. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 93–101. ACM, 2014.

[21] S Omar Ali and Zehra F Peynircioğlu. Songs and emotions: are lyrics and melodies equal partners? *Psychology of Music*, 34(4):511–534, 2006.

[22] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, pages 255–266. Citeseer, 2010.