MIRUtrecht participation in MediaEval 2013: Emotion in Music task

Anna Aljanaki, Frans Wiering, Remco C. Veltkamp
Utrecht University, Princetonplein 5, Utrecht 3584CC
{ A.Aljanaki@uu.nl, F.Wiering@uu.nl, R.C.Veltkamp@uu.nl}

ABSTRACT

This working notes paper describes the system proposed by the MIRUtrecht team[1] for static emotion recognition from audio (task *Emotion in Music*) in the MediaEval evaluation contest 2013. We approach the problem by proposing a scheme comprising data filtering, feature extraction, attribute selection and multivariate regression. The system is based on state-of-the art research in the field and achieved performance of (in terms of R^2, i.e. proportion of variance explained by the model) 0.64 for arousal and 0.36 for valence.

1. INTRODUCTION

The objective of the static task *Emotion in music* in the MediaEval 2013 evaluation contest is to predict emotion from musical audio. The training dataset consists of 700 music audio files of 45 seconds, belonging to eight different genres, which were annotated using the valence and arousal emotional model by Mechanical Turk workers. In this paper we describe the computational model, built on a training set and evaluated on a test set, which consisted of 300 audio files, annotated in the same way. More details concerning the dataset collection can be found in [4].

The valence-arousal model allows to avoid verbalization problems during data collection and is easily amenable to computational modeling. Two possibilities exist for modeling data using this model. The first possibility is to classify music into one of four quadrants, which correspond to emotions of (from the upper right clockwise) happiness, relaxation, depression and anger. The second possibility is to build a regression model separately for valence and arousal. The latter approach is employed in this paper.

1.1 Related Work

A regressive approach to modeling valence and arousal has already been undertaken by many researchers (see review by Yang [7]), with notable attempts by MacDorman et al. [3] (using kernel ISOMAP or PCA for dimensionality reduction and multiple linear regression for predictions) and Yang et al. [8]. (using PCA for correlation reduction, RReliefF for feature selection and Support Vector Regression for predictions). In [8], the prediction accuracy in terms of R^2 reaches 58.3 for arousal and 28.1 for valence.

2. SYSTEM DESCRIPTION

2.1 Data Filtering

In the dataset provided by MediaEval, it appears that valence and arousal dimensions are highly correlated (Pearson's r = 0.56, see also Figure 1). This is not an unusual situation (in [3], these dimensions correlate with Pearson's r = 0.33, in [8], r = 0.34). The upper left (angry) quadrant contains more data points than the opposite lower right (calm) quadrant. When looking at separate data points in the angry quadrant, we discovered some audio files containing speech or noise. We decided to filter them out. This was done after extracting features (as described in section 2.2). An InterquartileRange filter in Weka [3] was used to detect those outliers using both extracted features and valence-arousal annotations. For each feature, the audio file x is considered to be an outlier, if it satisfies the following criteria:

$$Q3 + 6*IQR < x < Q1 - 6*IQR,$$

where Q1 is the first quartile threshold, i.e. the middle number between the smallest and the median of the data set, *Q3* is the third quartile, i.e. the middle number between the largest and the median of the dataset, and *IQR* = Q3 – Q1.

In total, 13 items were deleted from the dataset based on suggestions from the filter, including, in addition to files containing speech, noise and environmental sounds, 4 files containing contemporary classical music. Figure 1 shows a scatterplot of the dataset, with outliers marked as red crosses.

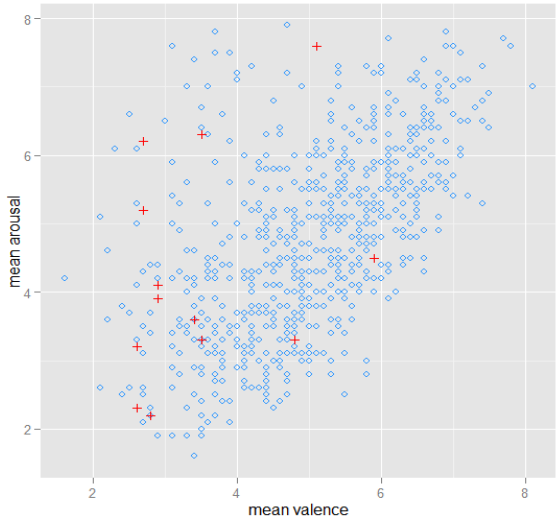

Figure 1. Training dataset plotted on valence-arousal plane. Each point is an audio file, red crosses are outliers.

2.2 Feature Extraction

We used three toolboxes to extract features, namely the MIRToolbox for Matlab [2], the Psysound [1] module for Matlab and the Queen Mary University VAMP plugin for Sonic Annotator [5]. Most of the features were extracted using MIRToolbox (see Table 1).

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

[1] This research is supported by the FES project COMMIT/.

Table 1. Extracted Features

Source	Features
MIRToolbox	rms, attack time, attack slope, spectral features (centroid, brightness, spread, skewness, kurtosis, flux), tempo, rolloff85, rolloff95, entropy, flatness, roughness, mfcc1-13, zero crossing rate, low energy, key clarity, mode, HCDF, inharmonicity, irregularity
PsySound	loudness
SonicAnnotator	mode

As we were predicting the emotion of the long (45 seconds) audio file, both the average values and the their standard deviations of the features were calculated, where applicable. From Psysound, the dynamic loudness (using the loudness model of Chalupper and Fastl) was employed. Sonic Annotator was used to extract an alternative estimation of mode. In MIRToolbox, the mode of the piece is calculated as a key strength difference between the best major and best minor key. In SonicAnnotator, modulation boundaries are detected, a certain key is predicted for each segment, and mode is estimated according to the amount of time the music is in major or minor mode. In total we extracted 44 features.

2.3 Feature Selection

The features we extracted are not necessarily all of equal importance to our task, and the feature set might contain redundant data. To select important features, we applied the ReliefF feature selection algorithm in WEKA. Table 2 shows the top 10 most important features for valence and for arousal according to ReliefF, where merit is the quality of the attribute, estimated using the probability of the predicted values of two neighbour instances being different.

Table 2. Feature importance

Rank	Arousal		Valence	
	Feature	Merit	Feature	Merit
1	loudness	0.016	roughness	0.011
2	spectral flux	0.013	spectral flux	0.08
3	HCDF	0.09	zero crossing rate	0.07
4	MFCC4	0.07	loudness	0.06
5	attack time	0.06	MFCC8	0.06
6	attack slope	0.06	std roughness	0.06
7	brightness	0.05	MFCC5	0.06
8	MFCC9	0.05	MFCC6	0.05
9	roughness	0.04	HCDF	0.05
10	keyclarity	0.04	brightness	0.05

As we can see, the most important features both for valence and arousal are loudness, spectral flux (as an average distance between each successive frames), roughness (average of all the dissonance between all possible pairs of peaks), and HCDF (harmonic change detection function, which is a flux of a tonal centroid).

Trying to maximize the R^2 value for model predictions, we selected 26 top attributes for arousal and 27 for valence.

2.4 Model fitting

With the selected attributes, we modeled the data using multiple regression, Support Vector Regression, M5Rules, Multilayer Perceptron and other regressive techniques available in WEKA, and evaluated them on the training set with 10-fold cross validation. The two systems that performed best were submitted for evaluation and are described below.

3. Results and Evaluation

The submitted systems were evaluated on 300 test items. Table 3 shows the results of the runs for multiple regression and for M5Rules, which are equal. Three metrics are provided: R^2 is the metric showing the goodness of fit of the model and is often described as the proportion of variance explained by the model, *MAE* is the Mean Average Error and *AE-STD* is its standard deviation.

Table 3. Evaluation

Evaluation metric	M5Rules & Multiple regression	
	arousal	valence
R^2	0.64	0.36
MAE	0.08	0.10
AE-STD	0.06	0.07

From the evaluation results we can conclude that such a simple technique as multiple regression performs as good as more sophisticated models, achieving a sufficiently good performance on a new dataset. The prediction accuracy of valence is, as one would expect from other attempts to model it [3,7,8], lower than that for arousal, though it is higher than in previous research, which might be the outcome of high degree of correlation between valence and arousal in this particular dataset.

4. REFERENCES

[1] Cabrera, D., 1999. PSYSOUND: A computer program for psychoacoustical analysis, in *Proc. Australian Acoust. Soc. Conf*., 1999, pp. 47–54

[2] Lartillot, O., Toiviainen, P., 2007. A Matlab Toolbox for Musical Feature Extraction FromAudio, *International Conference on Digital Audio Effects*, Bordeaux, 2007.

[3] MacDorman, K. F., Ough, S., and Ho, C.-C. 2007. Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. *J. New Music Res. 36*, 4, 281–299.

[4] Soleymani, M., Caro, M., Schmidt, E. M., Sha, C. , and Yang, Y. 2013. 1000 Songs for Emotional Analysis of Music. In *Proceedings of the ACM multimedia 2013 workshop on Crowdsourcing for Multimedia*. ACM, ACM, 2013.

[5] Sonic Annotator. http://www.omras2.org/SonicAnnotator

[6] Weka: Data mining software

http://www.cs.waikato.ac.nz/ml/weka/

[7] Yi-Hsuan Yang and Homer H. Chen. 2012. Machine Recognition of Music Emotion: A Review. *ACM Trans. Intell. Syst. Technol.* 3, 3, Article 40 (May 2012), 30 pages.

[8] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and H. H. Chen. 2008. A Regression Approach to Music Emotion Recognition. *Trans. Audio, Speech and Lang. Proc.* 16, 2 (February 2008), 448-45