MIREX 2014: MOOD CLASSIFICATION TASKS SUBMISSION

Renato Panda
CISUC
University of Coimbra
panda@dei.uc.pt

Bruno Rocha
CISUC
University of Coimbra
bmrocha@dei.uc.pt

Rui Pedro Paiva
CISUC
University of Coimbra
ruipedro@dei.uc.pt

ABSTRACT

NOTE: this is a draft, will be updated once data is out.

Our participation in the music emotion recognition task in MIREX 2014 consists of two strategies: our best solution from 2012 and a newer one with the addition of new melodic audio features, subject of study during the current year. Three audio frameworks – Marsyas, MIR Toolbox and PsySound3, are used to extract the commonly used audio features from the samples. In addition, MELODIA vamp plugin extracts the pitch contours (melody), which are then used to calculate several melodic features directly from audio. These features are then used with different classification strategies, manly using support vector machines, resulting in our various submissions.

1. INTRODUCTION

The classification system is built on research results obtained during the last [1-2] and current [cmmr-mml] years. First, several audio features are extracted from the existing dataset. To this end, Marsyas [3] framework as well as MIR Toolbox [4] and PsySound3 [5] are employed. Additionally, we calculate melodic audio features [mml] from the pitch contours outputted by MELODIA [juslin] vamp plugin. The remaining process of training a classifier and predicting labels for the test sets is done in MATLAB, using Support Vector Machines with the libSVM library [6].

The main differences between our submissions are the combination of features employed as well as the classification strategy. Namely, our best performing submission from 2012 is submitted unchanged (hierarchical approach [mirex 2012]), as well as a new version using only classification (SVC) instead of the regression (SVR) strategy.

The remaining versions employ the melodic audio features alone or in combination to the standard features used last year, which seem to improve the results in our recent studies [cmmr].

2. FEATURE EXTRACTION

Over the decades, several authors in other fields have studied the relations between music and emotion. Still, many of these relations are still unclear and further research is needed to implement computation models able to capture such characteristics.

Here, three standard audio frameworks were used to extract features from the existent audio files: Marsyas, a fast framework coded in C++; MIR Toolbox, an integrated set of functions written in MATLAB, that are specific to the extraction of musical features and provide a high number of both low and high-level audio features; PsySound3, a MATLAB toolbox for the analysis of sound recordings using physical and psychoacoustical algorithms. Furthermore, a new set of melodic features, extracted directly from audio files is used. These audio features are calculated using MATLAB, resorting to a previous melody transcription step by the MELODIA vamp plugin.

The process results in a total of 410 features: 124 obtained with Marsyas; 177 with MIR Toolbox (using statistics such as mean, standard deviation, kurtosis and skewness); 11 with PsySound3, based on the 15 best features identified in [7]; 98 melodic audio features as described in [cmmr]. A brief description of these features is presented in Table 1.

Framework	Features
Marsyas (124)	Centroid, rolloff, flux, Mel frequency cepstral coefficients (MFCCs), Peak Ratio – Chroma.
MIR Toolbox (177)	Among others: Root mean square (RMS) energy, rhythmic fluctuation, tempo, attack time and slope, zero crossing rate, rolloff, flux, high frequency energy, Mel frequency cepstral coefficients (MFCCs), roughness, spectral peaks variability (irregularity), inharmonicity, pitch, mode, harmonic change and key.
PsySound3 (11)	Loudness, sharpness, timbral width, spectral and tonal dissonances, pure tonalness, multiplicity.
Melodic features (no framework)	Divided in three categories: Pitch and duration, Vibrato, and Contour typology. From each feature (over all contours) mean, standard deviation, skewness and kurtosis are calculated. In addition to these features, we also compute: the melody's highest and lowest pitches; the range between them; the ratio of contours with vibrato to all contours in the melody.

Table 1. Used audio features and respective frameworks.

3. EMOTION CLASSIFICATION

Based on our previous studies, support vector machines are generally the best performing supervised learning techniques. Thus, our approaches are based on it, using the libSVM implementation. The main differences between submissions are all related with the number of features used and the classification approach: a single classifier and a hierarchical strategy using two levels to help discriminate between ambiguous clusters.

3.1 Single SVM Classifier

A single classifier is trained using the radial basis function kernel (RBF) to predict one of the five existent clusters, as studied in [2].

3.2 Hierarchical Classification Model

The goal behind the hierarchical approach is to reduce the semantic and acoustic ambiguity problems between clusters 1-5 and 2-4 that were previously identified [9]. Although our best result used regression at the first level, we submit an approach with a single SVM classifier to discriminate between groups of clusters 1-5, 2-4 and cluster 3. In cases where groups 1-5 and 2-4 are selected, a second level of classification is used to distinguish between the two using a second SVM classifier.

3.3 Feature Selection and SVM parameters

Three types of feature combinations are employed in different submissions. First, we use the entire feature set to train each model. Additionally, the subsets of features considered relevant for emotion are used [cmmr] with only 11 features. These were achieved with a similarly organized "MIREX-like" dataset [1-2]. Finally, a more dynamic submission selects a subset of the features available from the training set using ReliefF [10] selection algorithm. Furthermore, better parameters for each SVM model (cost, and γ, as well as ε for regression) are obtained using grid search.

4. ACKNOWLEDGMENTS

This work was supported by the MOODetector project (PTDC/EIA-EIA/102185/2008), financed by the Fundação para Ciência e a Tecnologia (FCT) and Programa Operacional Temático Factores de Competitividade (COMPETE) – Portugal.

5. REFERENCES

[1] R. Panda and R. P. Paiva, "Music Emotion Classification: Analysis of a Classifier Ensemble Approach," in *5th International Workshop on Machine Learning and Music (in ICML)*, 2012, pp. 1–2.

[2] R. Panda and R. P. Paiva, "Music Emotion Classification: Dataset Acquisition and Comparative Analysis," in *15th International Conference on Digital Audio Effects (DAFx-12)*, 2012.

[3] G. Tzanetakis, "Manipulation, Analysis and Retrieval Systems for Audio Signals," Science and Technology. Princeton University, 2002.

[4] O. Lartillot and P. Toiviainen, "A Matlab Toolbox for Musical Feature Extraction from Audio," in *Proc. 10th Int. Conf. on Digital Audio Effects*, 2007, pp. 237–244.

[5] D. Cabrera, S. Ferguson, and E. Schubert, "'Psysound3': Software for Acoustical and Psychoacoustical Analysis of Sound Recordings," in *Proceedings of the 13th International Conference on Auditory Display (ICAD2007)*, 2007, pp. 356–363.

[6] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *Computer*. pp. 1–30, 2001.

[7] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A Regression Approach to Music Emotion Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, Feb. 2008.

[8] J. Wang, H. Lo, and S. Jeng, "Mirex 2010: Audio Classification Using Semantic Transformation And Classifier Ensemble," in *Proc. of The 6th International WOCMAT & New Media Conference (WOCMAT 2010)*, 2010, pp. 2–5.

[9] C. Laurier and P. Herrera, "Audio music mood classification using support vector machine," in *MIREX task on Audio Mood Classification*, 2007, pp. 2–4.

[10] M. Robnik-Šikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1–2, pp. 23–69, 2003.

NOTE: this is a draft, will be updated once data is out.