Music Emotion Classification: Analysis of a Classifier Ensemble Approach

Renato Panda PANDA@DEI.UC.PT

Rui Pedro Paiva RUIPEDRO@DEI.UC.PT

CISUC, Department of Informatics Engineering, University of Coimbra, Pólo II, Coimbra, Portugal

Abstract

We propose a five regression models' system to classify music emotion. To this end, a dataset similar to MIREX contest dataset was used. Songs from each cluster are separated in five sets and labeled as 1. A similar number of songs from other clusters are then added to each set and labeled 0, training regression models to output a value representing how much a song is related to the specific cluster. The five outputs are combined and the highest score used as classification. An F-measure of 68.9% was obtained. Results were validated with 10-fold cross-validation and feature selection was tested.

1. Background

In the last decades we have seen a tremendous growth in the music industry, increasing the interest of research in areas like music information retrieval (MIR). However, mood emotion recognition (MER) is still a complex and vastly unexplored field, in part due to the subjectivity associated with emotions and the ambiguity regarding its description. Additionally, it is not yet well-understood how and why music elements create specific emotional responses in listeners (Yang et al., 2008).

To the best of our knowledge, the first paper on MER was published in 2003, by Feng et al. (2003). The majority of following research works (e.g. Yang et al., 2008; Yang et al., 2004; Lu et al., 2006), proposed similar classification approaches, experimenting with different sets of features, mood taxonomies, number of classes and datasets.

One of the main issues in the field is the lack of a standard dataset with audio clips and emotional information. Due to this fact, each author presents results based on his own dataset, making it impossible to compare results obtained between different studies.

In order to address this problem, an annual evaluation campaign for MIR where researchers can compare their algorithms was created (MIREX), using a collection of 600 audio clips divided in five clusters. However, the dataset is exclusive to the MIREX contests, thus unavailable to anyone. In latest edition, the best audio MER algorithm achieved an accuracy of 64%[1] (Want et al., 2010), highlighting once more the complexity and room for improvement. This result was one of the

2. Methodology

In this work, we propose a strategy where one classification model for each cluster is trained (a total of five), outputting a percentage. The percentages are then combined and the cluster with the highest score is used as the final result. Tests are conducted recurring to a previously gathered dataset, created with a similar configuration to the MIREX dataset.

2.1 Audio Feature Extraction

A dataset of 903 WAV clips organized in five clusters, similarly to the MIREX was used. This dataset and user annotated clusters were gathered from the Allmusic[2] database, relatively balanced between clusters.

The feature extraction process was done resorting to three different audio frameworks: PsySound (11), MIR Toolbox (177) and Marsyas (65), extracting a total of 253 features, normalized to the [0, 1] interval. PsySound and MIR Toolbox are MATLAB toolboxes, while Marsyas is a fast C++ audio analysis framework specific for MIR applications. Although some authors have already studied relevant musical attributes for mood analysis (Friberg, 2008), many are still to be implemented. This is due to the difficulty to extract them or not being fully understood and requiring further studies.

2.2 Classification System

A combination of five different regression models was used to predict a song's cluster. To this end, songs were divided by cluster in five distinct datasets and labeled 1. To each dataset, an equal number of songs from other clusters were added and labeled 0. As an example, dataset

In *5th International Workshop on Machine Learning and Music*, Edinburgh, Scotland, UK, 2012.

[1] http://nema.lis.illinois.edu/nema_out/9b11a5c8-9fcf-4029-95eb-51ed561cfb5f/results/evaluation/

[2] http://www.allmusic.com/

three consisted of 215 songs from the third cluster, labeled as 1, and 215 from the other clusters, labeled as 0. This dataset is then used to train a regression model, outputting a percentage of how related a test song is to the third cluster. The five outputs are combined, selecting the highest score as the classification (top1). The accuracy of the system using the best two scores (top2) was also measured.

To ensure the validity of the experience, 20 repetitions of 10-fold cross validation were used. To this end, each of the five datasets is divided in 10 folds, using nine folds to train the respective regression model. The remaining folds are combined in a global test data set, and used to test the output of the complete system. In order to reduce the number of features, the RReliefF algorithm was tested, selecting a smaller set of features for each regression model.

3. Results

The used dataset was built based on the known characteristics of the MIREX dataset. Still, they are different and results must be analyzed with this in mind. We believe that this dataset is harder than the MIREX data set, based in the Marsyas in MIREX 2010 results and testing the same features in the collected dataset, which were 10% lower. However, to confirm this, the same exact system must be tested in a future MIREX contest.

In terms of classification, the proposed system obtained an F-measure of 68.9%, with 69.2% precision and 68.8% recall. These results are a considerable improvement when compared with our tests using a single classification model (F-measure of 47.2%). It also surpasses the best results for mood classification in MIREX 2010. Although the data sets are different, it may indicate the possibility of an interesting result in a future edition. As for feature selection, a highly reduced set of features was obtained. Still, the results would also drop by 12%, when compared to a bigger set of features.

Table 1. Confusion matrix obtained with the best settings.

		PREDICTED				
		C1	C2	C3	C4	C5
ANNOTATED	C1	**72.4%**	8.9%	3.9%	7.1%	16.3%
	C2	7.1%	**66.1%**	6.0%	9.8%	6.0%
	C3	4.6%	9.5%	**77.3%**	10.2%	7.2%
	C4	8.4%	11.8%	10.0%	**68.3%**	8.6%
	C5	7.5%	3.7%	2.8%	4.6%	**61.9%**

Finally, precision results using the top2 clusters rose to almost 85%. This was done in order to test a previously identified semantic and acoustic overlap between clusters 1-5 and 2-4 (Laurier, 2007) in the MIREX dataset. Although not clearly visible, the confusion matrix presented in Table 1 indicates this, with 16.3% of songs from cluster 1 being identified as cluster 5.

4. Conclusions

When comparing with previous studies and MIREX mood contest results[3], this approach obtained comparable or even better results of 68.9%. Still, due to dataset differences, especially in the annotation process, results cannot be trivially compared. In order to solve this issue and to get an acceptable idea of the similarity between both datasets and how the system will perform, participation in the next MIREX mood classification contest is being considered.

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.

References

Yang, Y.-H., Lin, Y.-C., Su, Y.-F., and Chen, H. H. A Regression Approach to Music Emotion Recognition. IEEE Trans. on Audio, Speech, and Language Processing, 16(2):448-457, 2008.

Feng, Y., Zhuang, Y., and Pan, Y. Popular Music Retrieval by Detecting Mood. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2(2), pp. 375-376, New York, USA, 2003. ACM Press.

Yang, D., and Lee, W. Disambiguating Music Emotion Using Software Agents. Proceedings 5th International Conference on Music Information Retrieval, pp. 52–58, Barcelona, Spain. 2004

Lu, L., Liu, D., and Zhang, H.-J. Automatic Mood Detection and Tracking of Music Audio Signals. IEEE Transactions on Audio, Speech and Language Processing, 14(1), pp. 5-18, 2006

Wang, J., Lo, H., and Jeng, S. MIREX 2010: Audio Classification Using Semantic Transformation And Classifier Ensemble. Proceedings of the 6th International WOCMAT & New Media Conference (WOCMAT 2010), pp. 2-5, 2010

Friberg, A. Digital Audio Emotions - An Overview of Computer Analysis and Synthesis of Emotional Expression in Music. Proceedings 11th International Conference on Digital Audio Effects, pp. 1-6. Espoo, Finland. 2008

Laurier, C. Audio music mood classification using support vector machine. MIREX task on Audio Mood Classification, pp. 2-4. 2007

[3] http://www.music-ir.org/mirex/wiki/2010:MIREX2010_Results