AUDIO MUSIC MOOD CLASSIFICATION USING SUPPORT VECTOR MACHINE

Cyril Laurier, Perfecto Herrera
Music Technology Group
Universitat Pompeu Fabra
{claurier, pherrera}@iua.upf.edu

ABSTRACT

The system submitted to the MIREX Audio Music Mood Classification task is described here. It uses a set of 133 descriptors and a Support Vector Machine classifier to predict the mood cluster. The features are spectral, temporal, tonal but also describe loudness and danceability. The features were selected previously according to experiments on our annotated databases. The SVM is optimized using a grid search algorithm.

1 INTRODUCTION

Mood classification is a new MIREX contest and it is quite an arduous challenge because of its subjectivity and the influence of social and cultural factors. In a paper presented at ISMIR [1], we explored how a content-based similarity measure can help to classify by mood a collection of music files. In the algorithm submitted, we train a SVM with many descriptors empirically selected.

2 OVERVIEW

From audio data (22kHz Mono 30 seconds excerpts), we extract an extensive set of features. To know which features are relevant, we have made a previous analysis using different selection methods on our own annotated databases. This algorithm extracts 133 features detailed in the next section. Then we normalize them and finally we train a SVM model. We try to optimize the parameters of the model using a grid search method detailed in section 2.2. The algorithm is implemented in C++ and compiled as a Win32 binary.

2.1 Feature Set

All the features used in this submission have been selected based on results obtained empirically with the exemplar set provided and our databases. Using several feature selection methods in WEKA (PrincipalComponent, InfoGainAttributeEval, CfsSubsetEval, SVMAttributeEval) we have sorted out 133 features of different kind.

2.1.1 Spectral Descriptors

In our experiments, spectral descriptors were particularly helpful to classify by mood the exemplar songs provided. We decided to use:

-Spectral centroid, crest, flux, rolloff, skewness
-HFC (High Frequency Content)
-Spectral Strong peak [2]
-MFCC
-Bark Bands
-Energy Band Ratio
-FlatnessDB [3]

2.1.2 Loudness Descriptors

Testing with our databases we discovered that loudness helps to discriminate between some mood categories. For instance, in the MIREX categories, songs from Cluster 5 are rather loud. Here is the list of descriptors computed:

-RMS
-Loudness from BarkBands
-Dynamic Complexity (fluctuation of the Loudness)
-Larm [4]
-Leq [5]

2.1.3 High Level musical Descriptors

Mood is a quite abstract and subjective categorization. For that purpose some high level descriptors were used, like Danceability [6]. Moreover tonal descriptors [7] like the mode (major or minor) and the key strength were profitable as one can expect.

2.1.4 Temporal Descriptors

Finally we also extract temporal descriptors like zero crossing rate, onset rate and BPM.

2.1.5 Statistics

Most of these features are extracted using windowing. Afterward we compute statistics of these values (min, max, mean, variance, derivative variance, second-derivative variance). The decision to keep or not each value is made using feature selection methods in WEKA as previously mentioned.

2.2 Classification

Once the features extracted and normalized, we train a Support Vector Machine model. We use the libsvm [8] library. According to preliminary tests, the best results were achieved by the C-SVC method with the RBF kernel (Radial Basis Function). Consequently we use this configuration in our algorithm. Then to decide which values to choose for the cost C and the γ of the kernel function, we implemented a grid search algorithm like one suggested in [9]. We keep the parameters that obtain the best accuracy using a 10-fold Cross Validation on the training set. Finally when the optimal parameters are found, we train a SVM model and use it to predict the mood categories.

3 ANALYSIS OF THE RESULTS

Our submission is ranked second in terms of accuracy. All the results are listed in Table 1.

3.1 Overall classification

Participant	Accuracy
IMIRSEL M2K knn	47.17%
IMIRSEL M2K svm	55.83%
Cyril Laurier, Perfecto Herrera	**60.50%**
Kyogu Lee 1	49.83%
Kyogu Lee 2	25.67%
Lidy, Rauber, Pertusa, Iñesta	59.67%
Michael Mandel, Dan Ellis	57.83%
Michael Mandel, Dan Ellis spectral	55.83%
George Tzanetakis	**61.50%**

Table 1. Raw Classification Accuracy Averaged Over Three Train/Test Folds

3.2 Confusion matrix

To better understand the strong and weak points of the algorithm, Table 2 describes the mood clusters, and Table 3 the confusion matrix.

Cluster 1	passionate, rousing, confident boisterous, rowdy
Cluster 2	rollicking, cheerful, fun sweet, amiable/good natured
Cluster 3	literate, poignant, wistful bittersweet, autumnal, brooding
Cluster 4	humorous, silly, campy quirky, whimsical, witty, wry
Cluster 5	aggressive, fiery, tense/anxious intense, volatile, visceral

Table 2. Description of the mood clusters

We notice that the best predictable categories are cluster 3 and 5, which correspond roughly to sad and aggressive. The other clusters were more difficult to predict as

Truth/Predicted	1	2	3	4	5
Cluster 1	**45.8**	11.7	5.0	17.5	20.0
Cluster 2	10.8	**50.0**	11.7	27.5	0.0
Cluster 3	1.7	11.7	**82.5**	4.1	0.0
Cluster 4	10.0	31.7	4.2	**53.3**	0.8
Cluster 5	18.3	1.7	2.5	6.7	**70.8**

Table 3. Confusion Matrix, horizontally the distribution of the prediction for a given Cluster

one can expect listening to the examples. Consequently all the algorithms perform better with this two clusters. The category with the worse accuracy is cluster 1 often predicted as cluster 5. This makes sense as there are some acoustic similarities. Both are energetic, loud and many of both use electric guitar. Looking at the other submissions the same confusion appears. Moreover there is a clear confusion between cluster 2 and 4. Looking at the mood adjectives of this clusters, we can notice a possible semantic overlap. For example, using Wordnet [1] , we find that fun (cluster 2) and humorous (cluster 4) share the synonym : amusing. Besides humorous is a synonym of funny. We can observe this confusion also in the other algorithms results. To sum up we can argue that there are three main points :

1. Cluster 3 and 5 are the most predictable
2. There is a problem to predict Cluster 1 because it is close to Cluster 5 (acoustic similarities)
3. There is a confusion between Cluster 2 and 4 (possible semantic overlap)

3.3 Runtime

The submission is far from being optimized in terms of time. It runs more than 60 times slower than the most accurate algorithm. There are several reasons for that. The first is the grid search algorithm. It tries all the possibilities of C and γ mentioned in 2.2 (that is 360 combinations), and evaluates each pair with several 10-folds Cross Validations (two in this version). Finally a runtime error forced us to disable the compiler optimization, reducing a lot the speed of the executable. We believe that by narrowing the range of the grid search, doing only one CV and optimizing the build, we could reduce the computational cost to an acceptable value with the same accuracy.

4 FUTURE WORK

Many things can be tested and improved. The first refinement would be to pre-train a SVM model with the exemplar set. Afterwards, combining the probabilities from the SVM model trained with the training data and the probabilities from the pre-trained model, we should increase the accuracy. Then, If we still stick to this kind of approach, we can imagine an online feature selection instead

[1] http://wordnet.princeton.edu/

of the offline pre-selection. This would probably increase the accuracy of the overall system and allow us to add more descriptors without doing again the manual empirical analysis. Finally, we can think to try different classifiers (online or offline), and above all to add descriptors designed according to knowledge about mood perception and the related musical attributes.

5 ACKNOWLEDGEMENTS

This research has been partially supported by the PHAROS project (IST FP6-045035) Platform for searcH of Audiovisual Resources across Online Spaces. We want to thank Enric Guaus for his suggestions and help with libsvm, Emilia Gómez and Joan Serrà for their ideas, Nicolas Wack and Thomas Aussenac for their code and support.

6 REFERENCES

[1] Sordo, M. Laurier, C. Celma, O. "Annotating Music Collections: How content-based similarity helps to propagate labels" *Proceedings of 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

[2] Gouyon, F. and Herrera, P. "Exploration of techniques for automatic labeling of audio drum tracks instruments" *Proceedings of MOSART Workshop on Current Research Directions in Computer Music*, Barcelona, Spain 2001.

[3] Peeters, G. "A large set of audio features for sound description (similarity and classification) in the CUIDADO project" *CUIDADO I.S.T. Project Report* , 2004.

[4] Skovenborg, E. Nielsen, S.H. "Evaluation of Different Loudness Models with Music and Speech Material" *AES 117th Convention* , San Francisco, CA, USA, 2004.

[5] Soulodre, G. "Evaluation of Objective Loudness Meters" *AES 116 Convention* , Berlin, Germany, 2004.

[6] Streich, S. "Music Complexity: a multi-faceted description of audio content" *Ph.D. Dissertation*, UPF, Barcelona, 2007.

[7] Gómez, E. "Tonal Description of Music Audio Signals" *Ph.D. Dissertation*, UPF, Barcelona, 2006.

[8] C.-C. Chang, C.-J. Lin. "LIBSVM: a library for support vector machines" *Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm/*, 2001.

[9] C.-W. Hsu, C.-C. Chang, C.-J. Lin. "A practical guide to support vector classification" *http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf*