TWO ALGORITHMS FOR TIMBRE- AND RHYTHM-BASED MULTI-RESOLUTION AUDIO CLASSIFICATION

James Bergstra
Université de Montréal
Montreal, Quebec, Canada
H3T IJ4
bergstrj@iro.umontreal.ca

Norman Casagrande
Université de Montréal
Montreal, Quebec, Canada
H3T IJ4
casagran@iro.umontreal.ca

Douglas Eck
Université de Montréal
Montreal, Quebec, Canada
H3T IJ4
eckdoug@iro.umontreal.ca

Keywords: genre classification, artist recognition, audio features, multiclass boosting, mixture of experts

1 ALGORITHM

This section sketches out our multiresolution algorithm for the classification of music audio signals. We do not strictly adhere to the standard approach of feature-extraction followed by supervised machine-learning; instead, this idiom is encapsulated within an algorithm that compensates for the large discrepancy in temporal scale between feature extraction (47 milliseconds) and song classification (3-5 minutes). Our solution is to aggregate and classify features at an intermediate scale (we chose 13.9 seconds). We partition an input song into contiguous, non-overlapping segments of 13.9 seconds, and compute the mean and variance in standard timbre features over each segment.

We calculated a relatively large number of frame-level timbre features:

1. 256 RCEPS
2. 64 MFCC
3. 32 Linear predictive coefficients
4. 32 Low-frequency Fourier magnitudes
5. 16 Rolloff
6. 1 Linear prediction error
7. 1 Zero-crossing rate

We extracted these from 47ms frames of single-channel audio at 22050Hz using our own software Bergstra (2005). This resulted in 402 frame-level features, so that our meta-feature vector had 804 dimensions.

For such a large meta-feature vector, we relied on an extension of ADABOOST Schapire (1997), called ADABOOST.MH, described by Freund and Schapire (1995). We used ADABOOST.MH to boost decision stumps (algorithm 1) and 2-level trees (algorithm 2). We classified each of these meta-features independently with ADABOOST, and average the outputs across the set of meta-features to get a class confidence vector for the song. We label the song with the label that is most confident.

2 RESULTS

The two variants ranked first among all entries in case of the Magnature dataset and obtained an accuracy of 77.75% (compared to an accuracy of 71.96% that was obtained by the third ranked entryMandel and Ellis (2005)). In case of the USPOP dataset, the tree-based learner edged out the single-threshold learner; the accuracies are, respectively, 86.92% and 86.29% (compared to an accuracy of 85.65% that was obtained by the third ranked entry-Mandel and Ellis (2005)).

The single threshold weak learner ranked first in artist-recognition as well with 77.26% on the Magnatune dataset, whereas the three level tree weak learner ranked third with 74.45%. The second entry'sMandel and Ellis (2005) performance was 76.60%. On the USPOP dataset, Mandel and Ellis (2005) obtained 68.30%, whereas our algorithms ranked a distant second and third, with 59.88% (single-threshold) and 58.96% (three-level tree), respectively.

3 DISCUSSION

Since the tree weak-learner outperformed the stump weak-learner, we conjecture that the genres are not well separated in our the feature-space. It is interesting that the algorithm of West and Cox (2005) is outperformed by our model, because they used a similar segmentation strategy. We believe, the large number of features we used provides important additional information, and that ADABOOST.MH over small decision trees is a more robust classifier than a single regression tree, especially in a high-dimensional space like the one we created.

The method of Mandel and Ellis (2005) is simple and elegant, and although our algorithms were close overall in performance, we wonder if theirs would not be more effective if applied to several shorter segments of music.

We would like to point out that although ADABOOST.MH can take a long time to train, it is actually very quick to evaluate. Although our running times were generally high, we estimate that the classification of the test set examples took time on the order of 30 seconds. Certainly, classification is quick in comparison with the extraction of frame-level features.

ACKNOWLEDGEMENTS

Thanks to Alexandre Lacoste for many interesting ideas in conversation.

References

James Bergstra. The montreal scientific library, 2005. http://savannah.nongnu.org/cvs/?group=libmsl.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, pages 23–37, 1995.

Michael Mandel and Dan Ellis. Song-level features and SVMs for music classification. Extended Abstract, 2005. MIREX 2005.

Robert E. Schapire. Using output codes to boost multiclass learning problems. In *Proceedings of the 14th International Conference on Machine Learning*, pages 313–321. Morgan Kaufmann, 1997.

Kris West and Stephen Cox. Finding an optimal segmentation for audio genre classification. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, 2005.