Combining Features Extracted From Audio, Symbolic and Cultural Sources

Cory McKay and Ichiro Fujinaga

Overview

This research experimentally investigated the classification utility of combining features extracted from audio, symbolic and cultural sources of musical information. This was done by comparing the results of a series of genre classification experiments performed using all seven possible combinations and subsets of the three feature types.

The jMIR Software Suite

jMIR is a set of free and open-source software tools developed for use in automatic music classification research. The following jMIR components were used to perform the experiments described in this research:

jAudio: An audio feature extractor that includes implementations of 26 core features, as well as implementations of *metafeatures* and *aggregators* that can be used to automatically generate many more features.

jSymbolic: A symbolic feature extractor for processing MIDI files. jSymbolic is packaged with 111 mostly original features.

jWebMiner: A cultural feature extractor that extracts features from the Internet based on search engine co-occurrence page counts. Many user options are available to improve results, including search synonyms, filter strings and site weightings.

ACE: A metalearning classification system that automatically experiments with a variety of machine learning and dimensionality reduction algorithms in order to evaluate which are best suited to particular problems. ACE can also be used as a simple automatic classification system.

These and other jMIR components are described in much more detail in earlier publications and on jmir.sourceforge.net.

Overview

The SAC (Symbolic, Audio and Cultural) dataset was assembled to provide matching symbolic recordings, audio recordings and cultural metadata from which features could be extracted. SAC consists of 250 MIDI files, 250 MP3 recordings of the same music and matching metadata that can be parsed by jWebMiner for use in extracting cultural features from the Internet. The MIDI and audio files were acquired from separate sources in order to truly study the feature types independently.

SAC is divided into 10 different genres, with 25 pieces per genre. These 10 genres consist of 5 pairs of similar genres, as shown next.

SAC's Genre Taxonomy

Blues: Modern Blues *and* Traditional Blues
Classical: Baroque *and* Romantic
Jazz: Bop *and* Swing
Rap: Hardcore Rap *and* Pop Rap
Rock: Alternative Rock *and* Metal

Basic Methodology

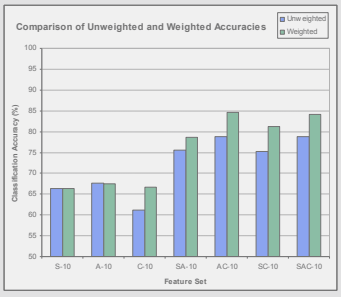
jAudio, jSymbolic and jWebMiner were used to extract features from the audio, symbolic and cultural metadata components of the SAC dataset, respectively. ACE was then used to perform 10-fold cross-validation genre classification experiments on each of the 7 possible subset combinations of these three feature types, once for the 5-genre taxonomy, and once for the 10-genre taxonomy. This resulted in a total of 14 classification average success rates that could be compared.

Weighted Success Rates

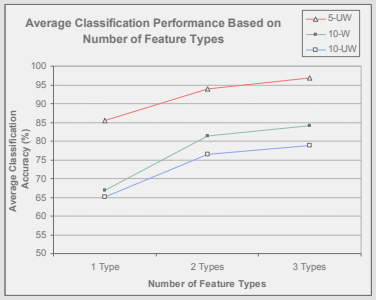
Normalized weighted classification success rates were calculated for each of the 10-genre experiments in order to compare the relative seriousness of misclassifications. This rate was calculated by scoring a misclassification within a genre pair (e.g., Baroque instead of Romantic) as 0.5 of an error, and by scoring a misclassification outside of a pair as 1.5 of an error.

Results: 5 Genres

Results: 10 Genres



Results: Combined Averages



The above figure shows the average classification accuracy rates for all experiments employing just one type of feature (S, A and C), two types of features (SA, AC and SC) or all three types of features (SAC). The three trend lines refer to the 5-genre taxonomy (5-UW), the unweighted 10-genre taxonomy (10-UW) and the weighted 10-genre taxonomy (10-W).

Conclusions

Multiple feature types vs. one: All feature groups consisting of two or more feature types outperformed all single feature groups. This improvement was statistically significant according to a Wilcoxon signed-rank test with a significance level of 0.125. The group consisting of all three feature types achieved average gains over the single feature type groups of 11.3% on the 5-genre taxonomy and 13.7% on the 10-genre taxonomy.

Three feature types vs. two: Combining all three feature types resulted, on average, in small increases in performance compared to the groups consisting of two feature types. These increases in performance were not statistically significant, however, as they were only 2.3% for the 5-genre taxonomy and 2.7% for the 10-genre taxonomy.

Seriousness of misclassifications: It was found that erroneous classifications tended to be to classes that were closer to the model classes in the experiments involving multiple feature types, particularly when cultural features were involved. The normalized weighted classification success rates were higher than the unweighted rates by an average of 5.7% when cultural features were present, compared to 0.9% when cultural features were not present, a result that is statistically significant with a significance level of 0.005 according to Student's paired t-test.

Genre classification feasibility: The classification accuracy rates of experiments involving multiple feature types were encouragingly high compared to earlier MIREX genre classification evaluations. This may be an indication that any ultimate ceiling on genre classification performance may not be as low as some have worried.

Social Sciences and Humanities Research Council of Canada
McGill
Schulich School of Music / École de musique Schulich
CIRMMT Centre for Interdisciplinary Research in Music Media and Technology
DDMAL