Automatic Music Genre Classification Using Ensemble of Classifiers

Carlos N. Silla Jr., Celso A. A. Kaestner, Alessandro L. Koerich

IEEE International Conference on Systems, Man and Cybernetics
SMC 2007 Montreal, Canada October/2007
Outline

• Introduction
• Overview
• Feature Extraction
• Classification
• Ensemble Approach
• Experimental Results
• Conclusions
Introduction

- Music genres are categorical labels created by humans to assign to the music styles.

- These labels are related to the instrumentalization, rhythmic structure and harmonic content of the music.

- These labels are also related to the human perception and cultural aspects.
Introduction

• The music genre is an ambiguous descriptor

• It has been used to categorize and organize large collections of digital music

• Therefore, automatic music genre classification can be very helpful.
• Extract features from a single 30-second frame from the beginning of the music.

• Shortcoming: It can not deal with the time variation of the music signal
• Extract features from the whole music signal.
• Shortcoming: features are averaged. Local information is disregarded. Time becomes critical.
Introduction

• Classification is dependent on the time

• Tradeoff between local and global information extracted from the music signal

• Have a more robust classification
Overview

Sampled Audio Signal \( s(1) \ldots s(N) \)

Audio Sample \( s(t) \)

Audio Frame \( t_f \)

Frame Interval \( t_i \)

Feature Vector \( \mathbf{x}_1 \mathbf{x}_2 \ldots \mathbf{x}_d \)

Classifier

\( P_1(g|x)P \)

\( P_2(g|x)P \)

\( P_M(g|x)P \)
Overview

- Feature Vector: $x_1 x_2 \ldots x_D$
- Classifier
  - $P_1(g|x)$
- Feature Vector: $x_1 x_2 \ldots x_D$
- Classifier
  - $P_2(g|x)$
- Feature Vector: $x_1 x_2 \ldots x_D$
- Classifier
  - $P_M(g|x)$

Combination Rule

- $\hat{g}$
- musical genre
Overview

Sampled Audio Signal \( a(1), \ldots, a(N) \)

Audio Band \( \theta(t) \)

Audio Frame \( F \)

Frame Interval \( t_a \)

Feature Vector \( x_1, x_2, \ldots, x_n \)

Classifier

\( P_1(g|x) \)

Classifier

\( P_2(g|x') \)

Classifier

\( P_3(g|x') \)

Combination Rule

\( \hat{g} \)

musical genre
Feature Extraction

- MARSYAS framework [Tzanetakis & Cook]

- Three feature sets:
  - **Timbral Texture (19)**: means and variance of the spectral centroid, rolloff, flux, the time zero domain crossings, the first five MFCCs and low energy.
  - **Beat Related (6)**: relative amplitudes and the beat per minute.
  - **Pitch Related (5)**: maximum periods of the pitch peak in the pitch histograms.

- 30-dimensional feature vector
Classification

• Assuming a digital music signal as a sequence of $N$ samples

\[ S = \langle s(1), s(2), \ldots, s(N) \rangle = s_1^N \]

• Assuming a sequence of $M$ feature vectors extracted from the digital music signal

\[ X_t = \langle \bar{x}_D(1), \bar{x}_D(2), \ldots, \bar{x}_D(m), \ldots, \bar{x}_D(M) \rangle \]

where $m$ encompasses several digital music signal samples (audio frame)
Classification

Sampled Audio Signal $s(1) \ldots s(N)$

Audio Sample $s(i)$

Audio Frame $t_f$

Frame Interval $t_i$

Feature Vector $X_1 X_2 \ldots X_D$

$\overline{x_D}(1)$

$\overline{x_D}(2)$

$\ldots$

$\overline{x_D}(M)$
Classification

• To assign a class (musical genre) which better represents the music

• This problem can be framed from a statistical perspective where the goal is to find the musical genre \( g \) that is most likely, given the feature vector \( \bar{x}_D(.) \)

\[
\hat{g} = \arg \max_{g \in \mathcal{G}} P(g | \bar{x}_D(.) )
\]

* \( a \ posteriori \) probability
Classification

• It can be rewritten using Bayes’ rule:

\[ P(g|x_D(\cdot)) = \frac{P(x_D(\cdot)|g) P(g)}{P(x_D(\cdot))} \]

\[ \text{probability of data occurring} \]

\[ \text{a priori probability of } g \]

• Assuming that the genre \( g \) is in \( G \) and that the classifier computes the likelihoods of the entire set of musical genres in \( G \), then:

\[ \sum_{g \in G} P(g|x_D(\cdot)) = 1 \]
Classification

• In such a way, estimated \textit{a posteriori} probabilities can be used as confidence estimates

• Then, we obtain the posterior for the genre hypothesis as:

\[
P(g|\bar{x}_D(\cdot)) = \frac{P(\bar{x}_D(\cdot)|g)P(g)}{\sum_{g \in G} P(\bar{x}_D(\cdot)|g)P(g)}
\]
Classification

• Classifiers
  – Naïve Bayes
  – Support Vector Machines (SVM) with the pairwise classification decomposition strategy
  – Multilayer perceptron (MLP) neural network trained with the backpropagation momentum algorithm.

• These classifiers were chosen because they provide at the output, *a posteriori* estimates
Ensemble Approach

- **Majority Vote Rule**
  \[ \hat{g} = \max_{g \in G, m \in [1, \ldots, M]} \text{count} \, P_m(g | \bar{x}_D(m)) \]

- **Max Rule**
  \[ \hat{g} = \arg \max_{g \in G, m \in [1, \ldots, M]} P_m(g | \bar{x}_D(m)) \]

- **Sum Rule**
  \[ \hat{g} = \arg \max_{g \in G} \sum_{m=1}^{M} P_m(g | \bar{x}_D(m)) \]

- **Product Rule**
  \[ \hat{g} = \arg \max_{g \in G} \prod_{m=1}^{M} P_m(g | \bar{x}_D(m)) \]
Experimental Results

• Latin Music Database
  – 10 Latin music genres
  – 3,000 music samples (300 samples per class)

• Datasets
  – 50% for training (1,500)
  – 20% for validation (600)
  – 30% for testing (900)
Experimental Results

- Three 30-second segments from each music
  - Beginning (1st), Middle (2nd), End (3rd)
Experimental Results on individual segments

Classification Accuracy

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>1st segment</th>
<th>2nd segment</th>
<th>3rd segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>J4.8</td>
<td>39.60%</td>
<td>44.44%</td>
<td>38.80%</td>
</tr>
<tr>
<td>3-NN</td>
<td>45.83%</td>
<td>48.43%</td>
<td>56.26%</td>
</tr>
<tr>
<td>MLP</td>
<td>53.96%</td>
<td>56.40%</td>
<td>53.96%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>44.43%</td>
<td>47.76%</td>
<td>39.13%</td>
</tr>
<tr>
<td>SVM</td>
<td>57.43%</td>
<td>63.50%</td>
<td>54.60%</td>
</tr>
</tbody>
</table>

Classifiers: J4.8, 3-NN, MLP, Naïve Bayes, SVM
Experimental Results

Combination Rules

- MAX
- SUM
- PROD
- Maj Vote

Combination Rules

- MLP
- Naïve Bayes
- SVM

Classification Accuracy
Experimental Results

Classification Accuracy

- MLP: 56.40% vs. 62.50%
- Naïve Bayes: 47.76% vs. 46.66%
- SVM: 63.50% vs. 65.73%

Legend:
- Green: best individual segment
- Orange: best combination rule
Conclusions

• The ensemble approach provides a more accurate genre classification relative to the individual classifiers.

• The improvement in accuracy depends on the classifier and ranges from 1% to 7%.

• The results were achieved on large dataset composed by 3,000 music samples.
Questions ?