Music Mood Classification
- an SVM based approach

Sebastian Napiorkowski
Topics on Computer Music (Seminar Report)
HPAC - RWTH - SS2015
Contents

1. Motivation

2. Quantification and Definition of Mood

3. How mood classification is done

4. Example: Mood and Theme Classification based on an Support Vector Machine approach
Motivation

• Imagine you could search songs based on the mood

• Create Playlists that follow a mood

• Create Playlists that follow a theme (e.g. party time)

• Users are already trying [1]:

  - music related searches
    - mood related: 15%
    - theme related: 30%
Contents

1. Motivation

2. Quantification and Definition of Mood
   1. Perception and Definition
   2. MIREX mood clusters
   3. Russell/Thayer’s Valence-Arousal model

3. How mood classification is done

4. Example: Mood and Theme Classification based on an Support Vector Machine approach
Perception and Definition

- Emotions can be [2]
  - **expressed** by music – feelings that are “intrinsic” to a given track
  - **induced** by music – feelings that the listener associates with a given track

- Music can have a [4]
  - **Mood** – the state and/or quality of a particular feeling associated to the track (e.g. happy, sad, aggressive)
  - **Theme** – refers to context or situations which fit best when listening to the track (e.g. party time, christmas, at the beach)
Perception and Definition

- Emotions can be [2]
  - **expressed** by music – feelings that are “intrinsic” to a given track
  - **induced** by music – feelings that the listener associates with a given track

- Music can have a [4]
  - **Mood** – the state and/or quality of a particular feeling associated to the track (e.g. happy, sad, aggressive)
  - **Theme** – refers to context or situations which fit best when listening to the track (e.g. party time, christmas, at the beach)
MIREX mood clusters

- MIREX (Music Information Retrieval Evaluation eXchange) (first mood task 2007)
- mutual exclusive clusters

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>Cluster 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>passionate, rousing, confident, boisterous, rowdy</td>
<td>rollicking, cheerful, fun, sweet, amiable/ good natured</td>
<td>literate, poignant, wistful, bittersweet, autumnal, brooding</td>
<td>humorous, silly, campy, quirky, whimsical, witty, wry</td>
<td>aggressive, fiery, tense/ anxious, intense, quirky, visceral</td>
</tr>
</tbody>
</table>
Russell/Thayer’s Valence-Arousal model

- most noted dimensional model [3]
- emotion exist on a plane along independent axes
- high to low - arousal (intensity)
- positive to negative - valence (appraisal of polarity)
Contents

1. Motivation

2. Quantification and Definition of Mood

3. How mood classification is done
   1. Content-based Audio Analysis

4. Example: Mood and Theme Classification based on an Support Vector Machine approach
How mood classification is done (or tried at least) [3]

• Contextual Text Information
  • mining web documents
  • social tags
  • Emotion recognition from lyrics
• Content-based Audio Analysis
• Hybrid Approaches
How mood classification is done (or tried at least) [3]

- Contextual Text Information
  - mining web documents
  - social tags
  - Emotion recognition from lyrics
  - Content-based Audio Analysis
  - Hybrid Approaches
Content-based Audio Analysis

- much prior work in Music-IR: **audio features**
- overview of most common used acoustic features used for mood recognition:
  - “blackbox toolset for audio classification”

<table>
<thead>
<tr>
<th>Type</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamics</td>
<td>RMS energy</td>
</tr>
<tr>
<td>Timbre</td>
<td>Mel-frequency cepstral coefficients (MFCCs), spectral shape, spectral contract</td>
</tr>
<tr>
<td>Harmony</td>
<td>Roughness, harmonic changes, key clarity, maharanis</td>
</tr>
<tr>
<td>Register</td>
<td>Chromagram, chroma centroid and deviation</td>
</tr>
<tr>
<td>Rhythm</td>
<td>rhythm strength, regularity, tempo, beat histograms</td>
</tr>
<tr>
<td>Articulation</td>
<td>Event density, attack slope, attack time</td>
</tr>
</tbody>
</table>
Content-based Audio Analysis

“more or less AC power”

“tune combination pleasant for the ear”

spectrum is projected onto 12 bins forming one octave

“time a tune gets to it’s loudest part”

<table>
<thead>
<tr>
<th>Type</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamics</td>
<td>RMS energy</td>
</tr>
<tr>
<td>Timbre (tone color)</td>
<td>Mel-frequency cepstral coefficients (MFCCs), spectral shape, spectral contract</td>
</tr>
<tr>
<td>Harmony</td>
<td>Roughness, harmonic changes, key clarity, maharanis</td>
</tr>
<tr>
<td>Register</td>
<td>Chromagram, chroma centroid and deviation</td>
</tr>
<tr>
<td>Rhythm</td>
<td>Rhythm strength, regularity, tempo, beat histograms</td>
</tr>
<tr>
<td>Articulation</td>
<td>Event density, attack slope, attack time</td>
</tr>
</tbody>
</table>
Content-based Audio Analysis

<table>
<thead>
<tr>
<th>Type</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamics</td>
<td>RMS energy</td>
</tr>
<tr>
<td>Timbre</td>
<td>Mel-frequency cepstral coefficients (MFCCs), spectral shape, spectral contract</td>
</tr>
<tr>
<td>Harmony</td>
<td>Roughness, harmonic changes, key clarity, maharanis</td>
</tr>
<tr>
<td>Register</td>
<td>Chromagram, chroma centroid and deviation</td>
</tr>
<tr>
<td>Rhythm</td>
<td>rhythm strength, regularity, tempo, beat histograms</td>
</tr>
<tr>
<td>Articulation</td>
<td>Event density, attack slope, attack time</td>
</tr>
</tbody>
</table>

“like JPEG for sound”

figure taken from [http://www.pampalk.at/ma/documentation.html](http://www.pampalk.at/ma/documentation.html)
Contents

1. Motivation

2. Quantification and Definition of Mood

3. How mood classification is done

4. Example: Mood and Theme Classification based on an Support Vector Machine approach
   
   1. Datasets
   
   2. Audio Feature - SV-Machine learning
   
   3. Social Tags - Naive Bayes classifier
4. Example: Mood and Theme Classification based on an Support Vector Machine approach

based on:

“Music Mood and Theme Classification - a hybrid approach”

Kerstin Bischoff, Claudiu S. Firan, Raluca Paiu, Wolfgang Nejdl

Cyril Laurier, Mohamed Sordo

L3S Research Center
Appelstr. 4, Hannover, Germany

Music Technology Group
Universitat Pompeu Fabra
4. Example: Mood and Theme Classification based on an Support Vector Machine approach

based on:

“Music Mood and Theme Classification - a hybrid approach”

worked on MIREX mood clusters [5]

Kerstin Bischoff, Claudiu S. Firan, Raluca Paiu, Wolfgang Nejdl

Cyril Laurier, Mohamed Sordo

L3S Research Center
Appelstr. 4, Hannover, Germany

Music Technology Group
Universitat Pompeu Fabra
Datasets: The truth, the whole truth, and nothing but the truth

- Find a ground truth dataset for training
- "ground truth" refers to the accuracy of the training set
- **AllMusic.com** (1995), Data gets created by music experts therefore good ground truth corpus:
  - Found 178 different moods and 73 Themes
  - 5,770 Tracks with moods assigned
  - 8,158 track-mood assignments (avg. 1.73 moods, max. 12)
  - 1,218 track-theme assignments (avg. 1.21 themes, max. 6)
Dataset: Social Tags

• **Last.fm** (2002) popular UK-based Internet radio and music community website

• Obtain tags for tracks from **AllMusic.com**

• Not all 5,770 Tracks have user tags

• Dataset is reduced to 4,737 Tracks
Dataset: Social Tags

- **Last.fm (2002)** popular UK-based Internet radio and music community website

- Obtain tags for tracks from [AllMusic.com](http://AllMusic.com)

- Not all 5,770 Tracks have user tags

- Dataset is reduced to 4,737 Tracks
Dataset: Prepare for multiclass classifier (1/2)

• We use the MIREX mood clusters

• five to seven AllMusic.com mood labels define together a MIREX mood cluster

• as mood clusters are mutual exclusive we restrict our dataset to tracks with 1-to-1 mood-track relations

• therefore dataset is reduced to 1192 distinct tracks
Dataset: Prepare for multiclass classifier (1/2)

- We use the MIREX mood clusters
- five to seven AllMusic.com mood labels define together a MIREX mood cluster
- as mood clusters are mutual exclusive we restrict our dataset to tracks with 1-to-1 mood-track relations
- therefore dataset is reduced to 1192 distinct tracks
To get an equal training set for the classifier, the cluster size is reduced to 200 per cluster.

- 5 Clusters means

- 1000 tracks for machine learning
Dataset: Prepare for multiclass classifier (1/2)

- To get an equal training set for the classifier, the cluster size is reduced to 200 per cluster
- 5 Clusters means
- 1000 tracks for machine learning
Dataset: Prepare for multiclass classifier (1/2)

- To get an equal training set for the classifier, the cluster size is reduced to 200 per cluster
- 5 Clusters means
- 1000 tracks for machine learning
Support Vector Machine Learning Dataset
1000 Tracks
classify 200ms frame-based extracted features

• timbral
• tonal
• rhythmic including MFCCs, BPM
• chroma features
• spectral centroid
• ...

23
assign mood from ground truth set
calculate support vectors
Radial Basis Function (RBF) kernel performed best
Results and Evaluation

- audio features were classified by a SVM
- also social tags were used to classify a track
  - with a Naive Bayes classifier (calculating Likelihoods)
- Algorithm is the same as in an other paper submitted to MIREX, but the results differ as they obtained 60.5 % accuracy and here we obtained only…

<table>
<thead>
<tr>
<th>Mood MIREX</th>
<th>Classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (audio)</td>
<td>0.450</td>
<td></td>
</tr>
<tr>
<td>NB (tags)</td>
<td>0.565</td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>0.575</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mood THAYER</th>
<th>Classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (audio)</td>
<td>0.517</td>
<td></td>
</tr>
<tr>
<td>NB (tags)</td>
<td>0.539</td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>0.596</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Themes clustered</th>
<th>Classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (audio)</td>
<td>0.527</td>
<td></td>
</tr>
<tr>
<td>NB (tags)</td>
<td>0.595</td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>0.625</td>
<td></td>
</tr>
</tbody>
</table>
### Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Mood MIREX</th>
<th>Moody MIREX</th>
<th>Themes clustered</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classifier</strong></td>
<td><strong>Accuracy</strong></td>
<td><strong>classifier (audio)</strong></td>
<td><strong>Accuracy</strong></td>
</tr>
<tr>
<td>SVM (audio)</td>
<td>0.450</td>
<td>SVM (audio)</td>
<td>0.517</td>
</tr>
<tr>
<td>NB (tags)</td>
<td>0.565</td>
<td>NB (tags)</td>
<td>0.539</td>
</tr>
<tr>
<td>Combined</td>
<td><strong>0.575</strong></td>
<td>Combined</td>
<td><strong>0.596</strong></td>
</tr>
</tbody>
</table>

- classifier relying only on audio features perform worse than pure tag based
- but combined: improve overall results
- The used ground-truth set was not that good as expected
- possible improvements:
  - filter training and test instances using listeners (that focus on audio only)
Conclusion

- Emotions are fuzzy and it’s not trivial to define them
- Machine learning highly depends on quality of training data
- It is hard to find a high quality ground truth dataset that is large enough
- since 2007 the results seem disillusioning: **mood classification is “hard to do”**

<table>
<thead>
<tr>
<th>MIREX year</th>
<th>Best Mood Classification Accuracy [6]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>0.6633</td>
</tr>
<tr>
<td>2013</td>
<td>0.6833</td>
</tr>
<tr>
<td>2012</td>
<td>0.6783</td>
</tr>
<tr>
<td>2011</td>
<td>0.6950</td>
</tr>
<tr>
<td>2010</td>
<td>0.6417</td>
</tr>
<tr>
<td>2009</td>
<td>0.6567</td>
</tr>
<tr>
<td>2008</td>
<td>0.6367</td>
</tr>
<tr>
<td>2007</td>
<td>0.6150</td>
</tr>
</tbody>
</table>
References


