Content-Based Music Genre Classification Using Sparse Approximation Techniques

Kamelia Aryafar Trevor R. Adams Ali Shokoufandeh

Department of Computer Science, Drexel University, Philadelphia, PA, USA

Introduction

- This work proposes a novel method the for music genre classification problem(MGC [2]) into different **genre labels** in a public music data set.
- The main challenges in creating an automatic music classification system are:
  - The robust representation of audio signals in terms of low-level features or high-level audio keywords.
  - The construction of an automatic learning schema to classify these feature vectors into music genres.
- In this study, we first propose an empirical feature selection method. We then utilize the recently proposed ℓ_1-SVM [1] to perform genre classification.

Audio Feature Representation

- **Overview of the MGC:**
  - An automatic genre classification system is composed of two main components: feature representation and classifier.

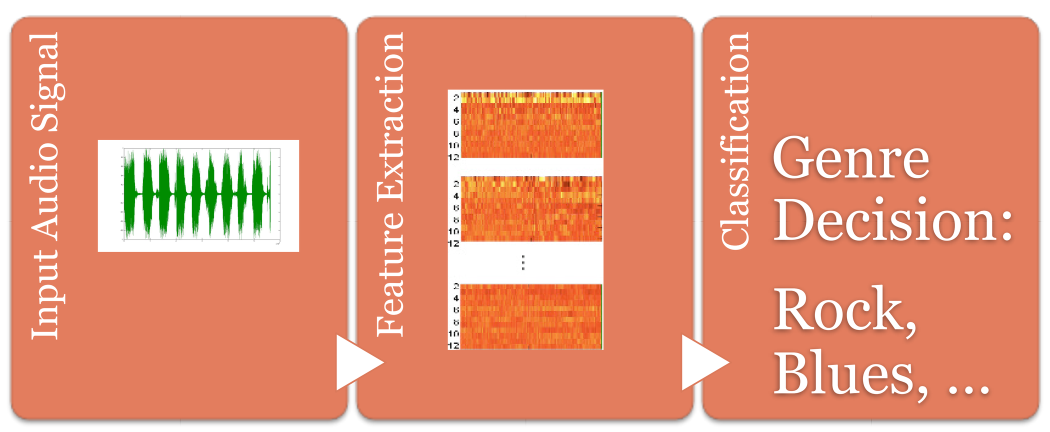
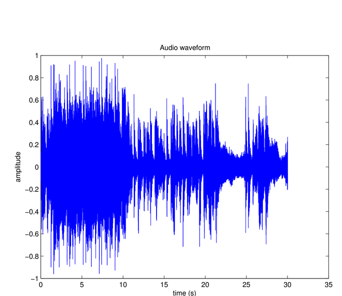
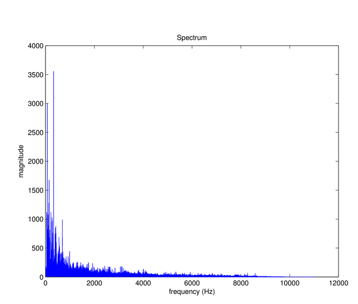


Figure 1: Robust feature extraction and classifier selection are the two main challenges for automatic music genre classification.

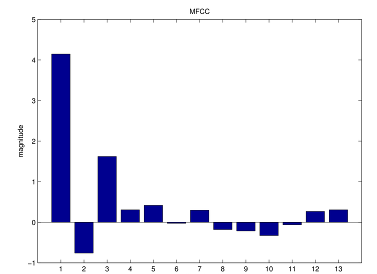
- **Content-based feature representation:**
  - Several features have been proposed in the literature of the MIR community to represent short-time or long-time audio characteristics.
  - Performance of these features for music genre classification vary by the choice of learning method and the feature representation of the audio signals.
  - The selected audio features include both short-time and long-time audio features:
    - Mel frequency cepstral coefficients (MFCCs) and chroma features are extracted using a sliding texture window.
    - Spectral centroid, entropy, spectral irregularity, brightness, roll off, spread, skewness, kurtosis and flatness are also extracted as signal level representations of long-time audio characteristics.



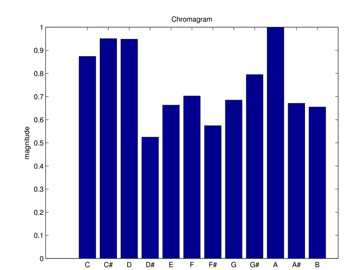
(a) A sample song



(b) Spectrum of a sample song



(c) MFCCs for a sample song



(d) Chroma features histogram of a sample song

Figure 2: A sample song is represented in the time and frequency domain in 2(a) and 2(b) respectively . 2(c) Shows MFCCs of the sample song while the chroma features histogram is illustrated in 2(d).

Empirical Feature Selection

- Table 1 illustrates the dimensionality of each feature.
- Short-time features are represented using a mean feature vector across all texture windows.
- Figure 3 illustrates the classification accuracy rate using various feature vectors on the GTZAN data set with a GMM classifier.

Audio Feature	Dimensionality
MFCCs	13
spectral entroid	1
entropy	1
spectral irregularity	1
brightness	1
roll off	1
spread	1
skewness	1
kurtosis	1
flatness	1
chroma	12

Table 1: Selected audio features and dimensionality of the feature space is shown above. The short-time audio features are represented using the mean feature vector across all texture windows.

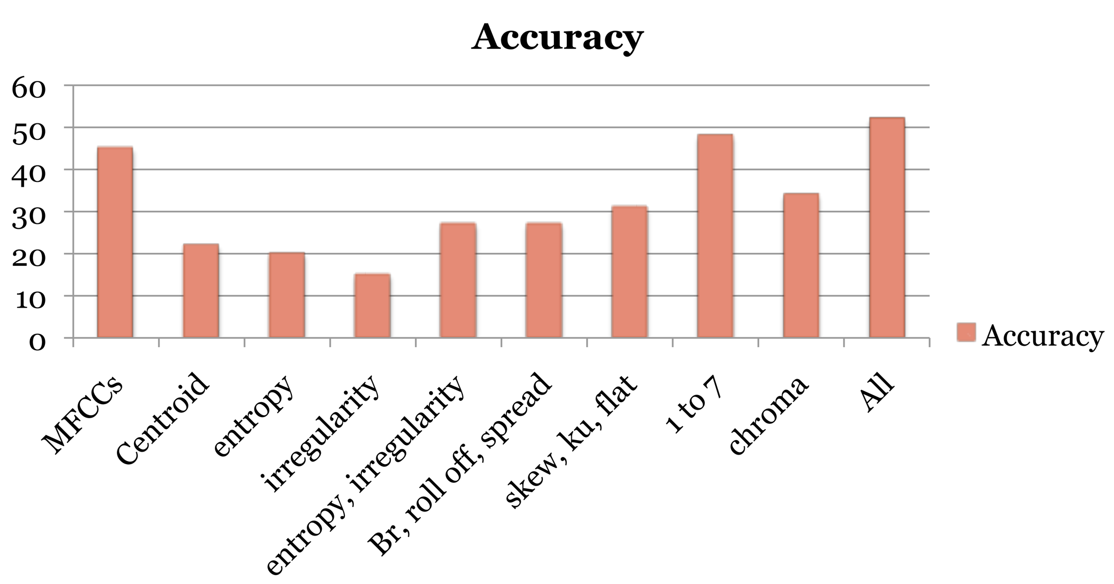


Figure 3: Features performance: average classification accuracy is reported using various feature vectors on GTZAN data set. Each experiment is repeated independently 3 times. The last column corresponds to the concatenated feature vector which outperforms the single feature vectors.

- MFCCs can be shown to be the most effective single feature vectors for music classification among selected features.
- The concatenation of all features outperforms singular feature vectors for genre classification.

Sparsity-eager SVM

- The sparsity-eager support vector machine classifier [1], i.e., the ℓ_1-SVM classifier combines the ideas of classical SVM with sparse approximation techniques.
  - higher generalization accuracy on new (test) samples
  - increased robustness against over-fitting to the training examples
  - provides scalability in terms of the classification complexity
- Given a set $\langle (x_1, y_1), \cdots, (x_M, y_M) \rangle$ of M training examples, we aim to find a vector $\alpha \in \mathbb{R}^M$ such that α is sufficiently sparse and yields a classifier $w \doteq \sum_{i=1}^{M} \alpha_i y_i x_i$ which has low empirical loss. Therefore the classfier has an adequately large separating margin.
-

$$\text{minimize} \quad \|\alpha\|_0 + \frac{C}{M}\sum_{i=1}^{M} \xi_i$$

$$\text{subject to } 1 - y_i \sum_{j=1}^{M} \alpha_j y_j x_j^\top x_i \le \xi_i,$$

$$0 \le \alpha_i \le \frac{C}{M}, \xi_i \ge 0, \; i \in \{1, \ldots, M\}.$$

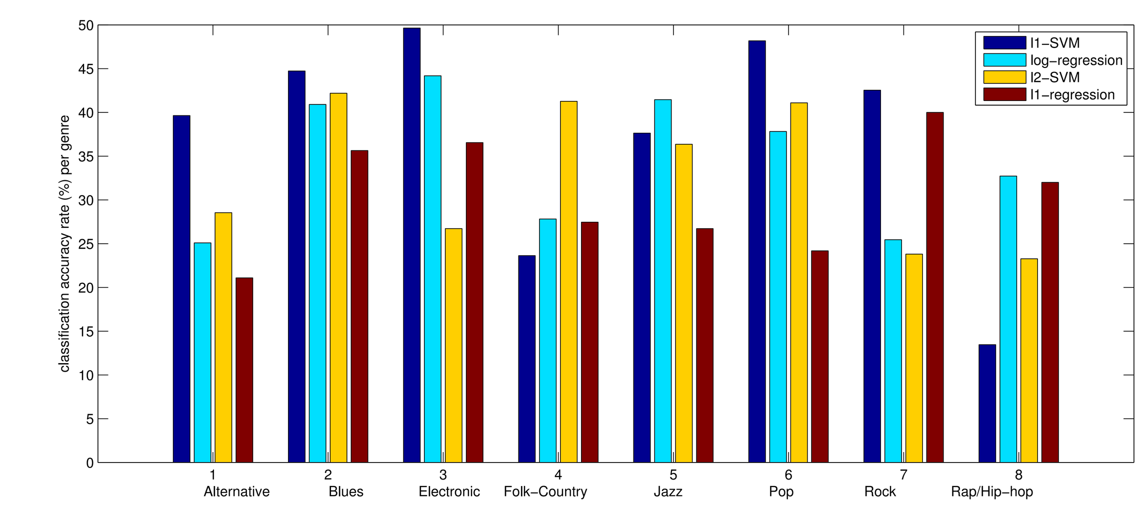
- the classification decision for a new sample x will be based on $\hat{y} \doteq \mathbf{Sign}\left(\sum_{i:\alpha_i \neq 0} \alpha_i y_i x_i^\top x\right)$ [1]

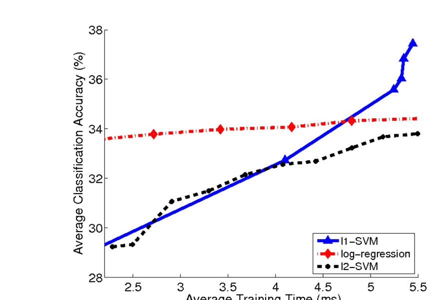
Experimental Results

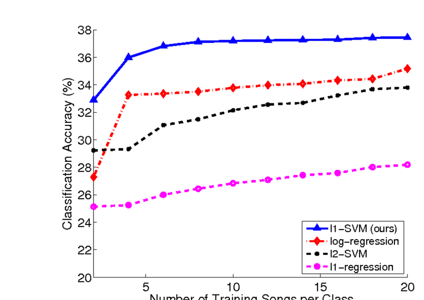
- **Data set:**
  - We used the publicly available benchmark dataset for audio classification and clustering.
  - The dataset contains samples of 1886 songs obtained from the Garageband site.
  - The data set includes 9 different genre samples of various sizes.

Genre	Samples
alternative	145
blues	120
electronic	113
folk-country	222
funk soul/R&B	47
jazz	319
pop	116
rap/hip-hop	300
rock	504

- **Experimental setup**
  - Validation method: 10-fold cross validation
  - Performance measure: classification accuracy rate







- The ℓ_1-SVM method outperforms the ℓ_1-regression, logistic regression, and SVM optimization using only MFCCs.

Classification method	Average accuracy rate
ℓ_1-SVM	37.43%
log-regression	34.43%
ℓ_2-SVM	32.90%
ℓ_1 regression	30.45%

Table 2: Average classification accuracy rate for music genre classification on the Homburg data set [3] is illustrated using MFCC features only. Each experiment is repeated independently 50 times and the average accuracy rate is reported [1].

Future Work

- Incorporate other audio features
  - Bag of audio keywords
  - Textual metadata
- Music artist identification in specific genre

Literature

[1] Kamelia Aryafar, Sina Jafarpour, and Ali Shokoufandeh.
Automatic musical genre classification using sparsity-eager support vector machines.
In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1526–1529. IEEE, 2012.

[2] Kamelia Aryafar and Ali Shokoufandeh.
Music genre classification using explicit semantic analysis.
In *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, MIRUM '11, pages 33–38, New York, NY, USA, 2011. ACM.

[3] Helge Homburg, Ingo Mierswa, Bülent Möller, Katharina Morik, and Michael Wurst.
A benchmark dataset for audio classification and clustering.
In *ISMIR*, pages 528–531, 2005.