Chord Based Feature Extraction for Genre Classification in Popular Brazilian Music

Bruna Wundervald [1*], Rogério Hultmann Filho[2], Walmes Marques Zeviani[2]

[1]PhD Candidate, Maynooth University *e-mail: brunadaviesw@gmail.com

[2] Federal University of Paraná

Introduction

- Music: a cultural element intrinsec to society.
- Music data: has many forms of representation, each one carrying different levels of information (sheet music, chords, lyrics, MIDI,...).
- Music genres: there is no exact definition for each class.

The main goals of this work are to

1. Propose an efficient method to extract music chords using the software R.
2. Extract the chords for a big set of brazilian popular music and perform feature engineering.
3. Find the most important features for classifying the songs in its respective genre.

Methods and Materials

Methods

- Exploratory & graphical data analysis (available at **brunaw.com/slides/braziliangenres**).
- Feature engineering: transforming the variables in features that represent the adjacent problem in a better way. Example:

$chord$	$major$	$contain\,7th$	$contain\,6th$
C	1	0	0
$Gm7$	0	1	0

$\underbrace{\qquad\qquad\qquad}_{extracted\,features}$

- Use of the **Spotify API** for obtaining the year and popularity of each song.
- Classification - Random Forests, a combination of trees:
  - Partitions of the space of variables in retangular regions and fitting of a simple model in each one of them.
  - The prediction is the most common genre in each region.
  - Minimizes the Gini impurity criteria:

$$Gini = 1 - \sum_{i=1}^{r} p_i^2 \quad (1)$$

where p_i is the proportion of each class in the training set.

  - The algorithm does not need too many computational efforts.

Materials

- Selected genres: MPB, Bossa Nova, Samba, Reggae, Pop, Rock, Forró and Sertanejo.
- Data: extracted from the CifraClub website (https://www.cifraclub.com.br/) via webscraping.
  - Resulting R package: `chorrrds`.
- Data analysis & modelling: software R - packages `ggplot2`, `dplyr` and `randomForest`.

Results

- Total of songs: 8261.
- The 22 extracted features were divided in four thematic groups: triads (6), tetrads (6), common transitions (3) and miscellany (7):
  - Summarized by song with percentages and means (for example percentage of chords with the sixth note in the song).
- Four models were fitted in the train set (70 %), in a 'nested' way. In order, the models contained the variables:
  1. triads
  2. triads + tetrads
  3. triads + tetrads + common transitions
  4. triads + tetrads + common transitions + miscellany
- The general accuracy, obtained with the test set (30 %), increased with the addition of variables.
- *Non Information Rate* (proportion of the most common genre) of **34 %**.
- The general accuracy reached a maximum of **62 %**, almost twice the *NIR*.
- The better predicted genres were 'Sertanejo' **(Acc. of 89 %)** and 'Samba' **(Acc. of 66 %)**.

The R-Music Blog

This work motivated the creation of an entire blog about music data analysis in R: **https://r-music.rbind.io/**

- Goal: to have a serie of posts about the extraction of music data and how to analyze it.
- Members: PhDs, PhD candidates, university teachers. data scientists and researchers from different areas interested in MIR.

Figure. The R-Music blog hexagon.

Conclusions

Conclusions

- It is possible to predict brazilian music genres from its harmonic structure.
- The most important features for the classification are the ones extracted using music chords.
- Being able to interpret the features is of great value.

Bibliography

[1] Hastie, Trevor, Tibshirani, Robert, Friedman. *The Elements of Statistical Learning The Elements of Statistical Learning*, 2009

[2] Caldas, Waldenyr. *Iniciação à Música Popular Brasileira*.Brasil, 2010.

[3] Real, E.C. *Feature extraction and sufficient statistics in detection and classification*.1996.