Stability of Symbolic Feature Group Importance in the Context of Multi-Modal Music Classification

Igor Vatolkin[1] and Cory McKay[2]

[1]Department of Computer Science, TU Dortmund University
[2]Department of Liberal and Creative Arts, Marianopolis College
[1]igor.vatolkin@udo.edu;[2]cory.mckay@mail.mcgill.ga

Overview

- Multi-modal music classification creates supervised models trained on features from different modalities
- Six modalities in this study: audio signal, model-based predictions, symbolic, album covers, playlists, lyrics
- Multi-group feature importance measures individual relevance of a feature group under investigation
- Features from other groups are allowed, but their proportion should be as small as possible
- Focus on eight symbolic feature groups: pitch, melodic, chords, rhythm, tempo, instrument presence / prevalence, texture
- Research hypothesis: multi-group feature importance may vary for different classification methods and evaluation measures
- Measurement of stability of importance for three classifiers and four measures

Reasons for Multi-Modal Music Classification

- Complementary information in different modalities typically leads to a better classification performance [1]–[4]
- Better understanding of how and why we define musical categories
- Despite possible redundancies between modalities, some information may be hard to extract (e.g., pitch statistics based on audio instead of the score)
- Analysis of relationships between modalities

Evaluation of Feature Groups: Concept

- Random initialization of a population with different feature sets: all modalities are allowed
- Evolutionary multi-objective feature selection switches individual feature dimensions on/off
- Optimization of two criteria:
  - $g_k \uparrow$: proportion of the features from the group k under investigation
  - $m_e \downarrow$: error measure
- Non-dominated front contains trade-off feature sets which are not dominated by any other set (see below Def. 1), with two boundary sets:
  - the highest g_k and the highest m_e: the best error which is achieved with a feature set as "pure" as possible, i.e., with only or almost only features from the group k
  - the lowest g_k and the lowest m_e: the lowest error at all, achieved with a feature set which contains features from other groups
- Dominated hypervolume: all theoretically possible feature sets which are worse with respect to both criteria
- Shaded area in Figure 1: difference between hypervolume of the ideal point and hypervolumes of feature sets in the non-dominated front
  - Small value: m_e can be only slightly improved by adding features from groups beyond k, i.e. the feature group k is more important
  - Large value: m_e can be greatly improved by adding features from groups beyond k, i.e., the feature group k is less important

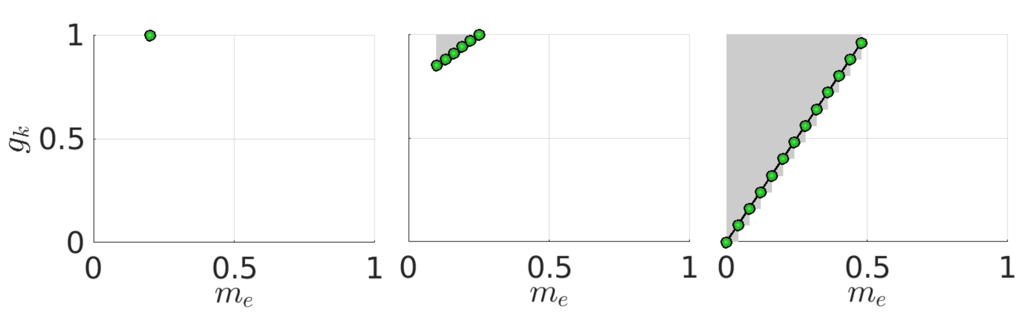

Figure 1: *Theoretically possible non-dominated fronts of feature sets for the minimization of m_e and maximization of g_k (partly taken from [4]).*

Evaluation of Feature Groups: Formal Details

- Let $m_1, ..., m_O$ be evaluation measures to minimize. Feature set $\mathbf{q}_1$ DOMINATES another set $\mathbf{q}_2$ when:

$$\forall i \in \{1, ..., O\} : m_i(\mathbf{q}_1) \leq m_i(\mathbf{q}_2) \text{ and}$$
$$\exists j \in \{1, ..., O\} : m_j(\mathbf{q}_1) < m_j(\mathbf{q}_2) \tag{1}$$

- Let $\mathbf{r} \in \mathbb{R}^O$ be the REFERENCE POINT (a worst possible feature set) and Λ_d the volume of a set in $\mathbb{R}^O$. For $\mathbf{q}_1, \ldots, \mathbf{q}_\phi$, which are not dominated by any other feature sets (NON-DOMINATED FRONT), the DOMINATED HYPERVOLUME is:

$$H(\mathbf{q}_1, \ldots, \mathbf{q}_\phi; \mathbf{r}) = \Lambda_d \left(\bigcup_{i=1}^{\phi} [\mathbf{q}_i, \mathbf{r}] \right) \tag{2}$$

- Let $\mathbf{q}_{ID} \in \mathbb{R}^O$ be the IDEAL POINT with the best individual values of $m_1, ..., m_O$ from all non-dominated feature sets

$$h = H(\mathbf{q}_{ID}; \mathbf{r}) - H(\mathbf{q}_1, \ldots, \mathbf{q}_\phi; \mathbf{r}) \tag{3}$$

- MULTI-GROUP FEATURE IMPORTANCE is defined as:

$$i_h = 1 - h(m_e \downarrow, g_k \uparrow) \tag{4}$$

- NORMALIZED MULTI-GROUP FEATURE IMPORTANCE is defined as:

$$I_h = max \left\{ \frac{i_h - 0.75}{0.25}, 0 \right\} \tag{5}$$

Experiments: Datasets and Algorithms

- Datasets (features available at `https://zenodo.org/record/5651429`)
  - LMD-aligned [6]: 1575 tracks selected for a balanced genre distribution, genre annotations
  - SLAC [1]: all 250 tracks used, genre and sub-genre annotations
- Classifiers
  - Random forest (RF)
  - k-nearest neighbors (kNN)
  - Support vector machines (SVM)
- Measures
  - Let TP be true positives, TN true negatives, FP false positives, and FN false negatives
  - Balanced relative error:

$$m_{BRE} = \frac{1}{2} \left(\frac{FN}{TP + FN} + \frac{FP}{TN + FP} \right) \tag{6}$$

  - Recall:

$$m_{REC} = \frac{TP}{TP + FN} \tag{7}$$

  - Specificity:

$$m_{SPEC} = \frac{TN}{TN + FP} \tag{8}$$

  - F1-measure:

$$m_{F1} = \frac{2 \cdot m_{PREC} \cdot m_{REC}}{m_{PREC} + m_{REC}}, \tag{9}$$

  where

$$m_{PREC} = \frac{TP}{TP + FP} \tag{10}$$

Experiments: Symbolic Feature Groups

Group	Feature Examples
Pitch	First pitch, last pitch, major or minor, pitch class histogram, pitch variability, range
Melodic	Amount of arpeggiation, direction of melodic motion, melodic intervals, repeated notes
Chords	Chord type histogram, dominant seventh chords, variability of number of simultaneous pitches
Rhythm	Initial time signature, metrical diversity, note density per quarter note, prevalence of dotted notes
Tempo	Initial tempo, mean tempo, minimum and maximum note duration, note density and its variation
Instrument presence	Note prevalences of pitched and unpitched instruments, pitched instruments present
Instrument prevalence	Prevalences of individual instruments/instrument groups: acoustic guitar, string ensemble, etc.
Texture	Average number of independent voices, parallel fifths and octaves, voice overlap

Table 1: *Sample jSymbolic [5] features grouped into eight semantically meaningful groups.*

Results: Non-Dominated Fronts

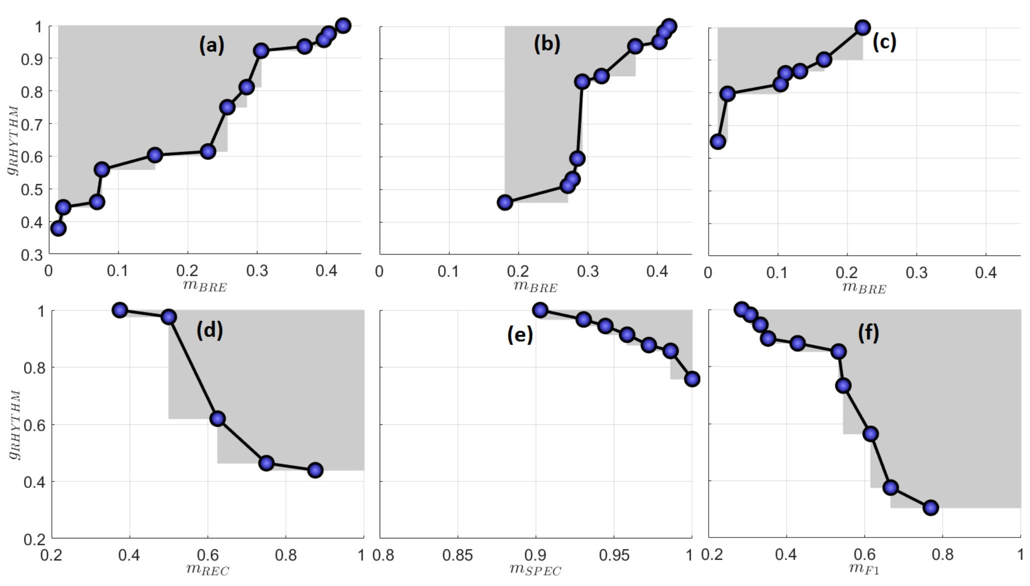

Figure 2: *Non-dominated fronts after multi-objective feature selection for the identification of Traditional Blues in the SLAC dataset [1], with rhythmic descriptors the feature group being focused on. Top row: the share of rhythmic descriptors g_{RHYTHM} is maximized and the balanced relative error m_{BRE} is minimized using random forest (a), k-nearest neighbors (b), or support vector machine (c) classifiers. Bottom row: random forest classifiers are used to maximize both g_{RHYTHM} and recall m_{REC} (d), specificity m_{SPEC} (e), or F1-measure m_{F1} (f).*

Results: Tables

		Pitch	Melodic	Chords	Rhythm
Country	RF $\|m_{BRE}$	0.796±0.05	0.675±0.07	0.798±0.02	0.820±0.04
	kNN $\|m_{BRE}$	0.924±0.08	0.704±0.07	0.832±0.03	0.865±0.04
	SVM $\|m_{BRE}$	0.756±0.03	0.688±0.04	0.785±0.04	0.730±0.03
	RF $\|m_{F1}$	0.610±0.09	0.439±0.03	0.576±0.10	0.654±0.05
	RF $\|m_{REC}$	0.849±0.01	0.722±0.06	0.823±0.03	0.882±0.02
	RF $\|m_{SPEC}$	0.760±0.03	0.678±0.06	0.768±0.04	0.785±0.04
Electronic	RF $\|m_{BRE}$	0.825±0.04	0.774±0.02	0.815±0.01	0.871±0.02
	kNN $\|m_{BRE}$	0.867±0.07	0.748±0.11	0.861±0.06	0.914±0.06
	SVM $\|m_{BRE}$	0.824±0.04	0.773±0.01	0.770±0.05	0.836±0.05
	RF $\|m_{F1}$	0.695±0.05	0.583±0.05	0.681±0.02	0.771±0.02
	RF $\|m_{REC}$	0.877±0.03	0.751±0.08	0.830±0.05	0.911±0.03
	RF $\|m_{SPEC}$	0.757±0.09	0.738±0.05	0.748±0.02	0.835±0.02
Rock	RF $\|m_{BRE}$	0.933±0.00	0.887±0.03	0.911±0.07	0.917±0.03
	kNN $\|m_{BRE}$	0.791±0.15	0.716±0.09	0.753±0.08	0.802±0.13
	SVM $\|m_{BRE}$	0.808±0.02	0.827±0.09	0.816±0.09	0.797±0.04
	RF $\|m_{F1}$	0.799±0.03	0.705±0.06	0.843±0.05	0.804±0.02
	RF $\|m_{REC}$	0.960±0.03	0.971±0.03	0.982±0.03	0.969±0.03
	RF $\|m_{SPEC}$	0.956±0.03	0.878±0.01	0.955±0.01	0.947±0.01
RockAltern	RF $\|m_{BRE}$	0.787±0.06	0.768±0.10	0.815±0.07	0.891±0.02
	kNN $\|m_{BRE}$	0.658±0.14	0.678±0.07	0.639±0.10	0.830±0.14
	SVM $\|m_{BRE}$	0.670±0.12	0.679±0.13	0.797±0.08	0.804±0.14
	RF $\|m_{F1}$	0.535±0.20	0.458±0.05	0.764±0.14	0.740±0.05
	RF $\|m_{REC}$	0.578±0.19	0.794±0.19	0.855±0.04	0.884±0.03
	RF $\|m_{SPEC}$	0.977±0.03	0.918±0.06	0.987±0.02	0.980±0.03
RockMetal	RF $\|m_{BRE}$	0.914±0.03	0.840±0.06	0.873±0.09	0.847±0.11
	kNN $\|m_{BRE}$	0.986±0.02	0.797±0.12	0.797±0.18	0.835±0.04
	SVM $\|m_{BRE}$	0.758±0.25	0.788±0.05	0.915±0.04	0.826±0.03
	RF $\|m_{F1}$	0.943±0.04	0.605±0.14	0.742±0.15	0.739±0.18
	RF $\|m_{REC}$	0.965±0.03	0.703±0.23	0.795±0.26	0.757±0.16
	RF $\|m_{SPEC}$	0.995±0.01	0.972±0.02	0.985±0.00	0.992±0.01

Table 2: *Normalized multi-group feature importances, aggregated over three folds. The first two genres are taken from LMD-aligned, the last three from SLAC. The group with the highest importance is marked in deep red, and with the lowest importance in deep blue.*

		Tempo	Instr. Pres.	Instr. Prev.	Texture
Country	RF $\|m_{BRE}$	0.656±0.03	0.945±0.01	0.614±0.02	0.700±0.03
	kNN $\|m_{BRE}$	0.861±0.06	0.952±0.03	0.645±0.02	0.779±0.03
	SVM $\|m_{BRE}$	0.651±0.05	0.861±0.03	0.633±0.01	0.665±0.02
	RF $\|m_{F1}$	0.413±0.11	0.889±0.01	0.285±0.12	0.381±0.08
	RF $\|m_{REC}$	0.755±0.08	0.965±0.01	0.668±0.06	0.733±0.08
	RF $\|m_{SPEC}$	0.618±0.06	0.927±0.02	0.604±0.09	0.674±0.03
Electronic	RF $\|m_{BRE}$	0.754±0.07	0.951±0.02	0.697±0.03	0.746±0.02
	kNN $\|m_{BRE}$	0.954±0.01	0.964±0.02	0.700±0.04	0.708±0.05
	SVM $\|m_{BRE}$	0.768±0.00	0.916±0.01	0.683±0.03	0.758±0.03
	RF $\|m_{F1}$	0.542±0.08	0.934±0.01	0.450±0.01	0.553±0.02
	RF $\|m_{REC}$	0.738±0.12	0.944±0.01	0.718±0.04	0.759±0.04
	RF $\|m_{SPEC}$	0.742±0.04	0.956±0.01	0.660±0.03	0.733±0.01
Rock	RF $\|m_{BRE}$	0.833±0.05	0.995±0.00	0.915±0.04	0.905±0.02
	kNN $\|m_{BRE}$	0.596±0.32	0.995±0.00	0.804±0.01	0.702±0.11
	SVM $\|m_{BRE}$	0.758±0.03	0.982±0.00	0.866±0.06	0.749±0.06
	RF $\|m_{F1}$	0.628±0.05	0.984±0.01	0.787±0.04	0.743±0.03
	RF $\|m_{REC}$	0.910±0.07	1.000±0.00	1.000±0.00	0.974±0.02
	RF $\|m_{SPEC}$	0.870±0.02	0.994±0.00	0.928±0.02	0.941±0.06
RockAltern	RF $\|m_{BRE}$	0.824±0.08	0.971±0.01	0.728±0.09	0.681±0.05
	kNN $\|m_{BRE}$	0.751±0.08	0.952±0.05	0.423±0.24	0.593±0.19
	SVM $\|m_{BRE}$	0.751±0.11	0.903±0.07	0.695±0.10	0.592±0.04
	RF $\|m_{F1}$	0.597±0.13	0.911±0.07	0.442±0.15	0.498±0.24
	RF $\|m_{REC}$	0.847±0.22	0.942±0.04	0.698±0.08	0.765±0.10
	RF $\|m_{SPEC}$	0.907±0.11	1.000±0.00	0.963±0.04	0.934±0.01
RockMetal	RF $\|m_{BRE}$	0.840±0.04	0.988±0.00	0.950±0.03	0.840±0.09
	kNN $\|m_{BRE}$	0.720±0.30	0.985±0.01	0.795±0.17	0.601±0.17
	SVM $\|m_{BRE}$	0.750±0.08	0.981±0.02	0.946±0.04	0.810±0.03
	RF $\|m_{F1}$	0.630±0.15	0.985±0.01	0.802±0.04	0.775±0.04
	RF $\|m_{REC}$	0.823±0.05	0.991±0.01	1.000±0.00	0.807±0.16
	RF $\|m_{SPEC}$	0.965±0.03	1.000±0.00	0.992±0.01	0.990±0.01

Table 3: *Normalized multi-group feature importances, aggregated over three folds. The first two genres are taken from LMD-aligned, the last three from SLAC. The group with the highest importance is marked in deep red, and with the lowest importance in deep blue.*

Conclusions

- Proposed framework helps to investigate a deeper analysis of multiple musical modalities
- Importance of feature groups varies by change of a classifier and measure
- However: far away from a random behavior
- Future work: measurement of the impact of other settings (evaluation measures, classifier hyper-parameters)

References

[1] C. McKay, J. A. Burgoyne, J. Hockman, J. B. L. Smith, G. Vigliensoni, and I. Fujinaga: Evaluating the Genre Classification Performance of Lyrical Features Relative to Audio, Symbolic and Cultural Features. Proc. ISMIR, 213–218 (2010)

[2] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva: Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis. Proc. CMMR, 570–582 (2013)

[3] S. Oramas, F. Barbieri, O. Nieto, and X. Serra: Multimodal Deep Learning for Music Genre Classification. Trans. Int'l Society for Music Information Retrieval, 1(1):4–21 (2018)

[4] I. Vatolkin and C. McKay: Multi-Objective Investigation of Six Feature Source Types for Multi-Modal Music Classification. Trans. Int'l Society for Music Information Retrieval, 5(1):1–19 (2022)

[5] C. McKay, J. Cumming, and I. Fujinaga: jSymbolic 2.2: Extracting Features from Symbolic Music for use in Musicological and MIR Research. Proc. ISMIR, 348–354 (2018)

[6] C. Raffel: Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching. PhD thesis, Graduate School of Arts and Sciences, Columbia University (2016)

ISMIR 2022 23rd International Society for Music Information Retrieval Conference, Bengaluru, India, December 4-8, 2022