The scientific evaluation of music content analysis systems: Valid empirical foundations for future real-world impact

STURM, BLT; Maruri-Aguilar, H; Parker, B; Grossmann, H; International Conference on Machine Learning

The scientific evaluation of music content analysis systems: Valid empirical foundations for future real-world impact

Bob L. Sturm B.STURM@QMUL.AC.UK
School of Electronic Engineering and Computer Science, Queen Mary University of London, U.K.

Hugo Maruri-Aguilar, Ben Parker {H.MARURI-AGUILAR, B.PARKER}@QMUL.AC.UK
School of Mathematical Sciences, Queen Mary University of London, U.K.

Heiko Grossmann HEIKO.GROSSMANN@OVGU.DE
Institute for Mathematical Stochastics, Otto-von-Guericke University Magdeburg, Germany

Abstract

We discuss the problem of music content analysis within the formal framework of experimental design.

1. Preliminaries

By the formalism developed in Sturm et al. (2014), define a *music universe* Ω, a *music recording universe* $\mathcal{R}_\Omega$, *descriptor vocabularies* $\mathbb{F}$ (features) and $\mathcal{V}$ (tokens), and *Boolean semantic rules* $A' : f \to \{T, F\}$ and $A : s \to \{T, F\}$, where f and s are finite sequences of elements in $\mathbb{F}$ and $\mathcal{V}$, respectively. Define the *semantic universe*

$$\mathcal{S}_{\mathcal{V},A} := \{s \in \mathcal{V}^n | n \in \mathbb{N} \wedge A(s) = T\}. \quad (1)$$

The *semantic feature universe* $\mathcal{S}_{\mathbb{F},A'}$ is built similarly, using $\mathbb{F}$ and A'. A *recorded music description system* $\mathcal{S}$ is a map

$$\mathcal{S} : \mathcal{R}_\Omega \to \mathcal{S}_{\mathcal{V},A} \quad (2)$$

which is a composition of two maps: $\mathcal{E} : \mathcal{R}_\Omega \to \mathcal{S}_{\mathbb{F},A'}$ and $\mathcal{C} : \mathcal{S}_{\mathbb{F},A'} \to \mathcal{S}_{\mathcal{V},A}$. The map $\mathcal{E}$ is commonly known as a "feature extractor," and $\mathcal{C}$ as a "classifier." A *recorded music dataset* is an indexed sample

$$\mathcal{D} = ((r_i, s_i) : i \in \mathcal{I}) \subset \mathcal{R}_\Omega \times \mathcal{S}_{\mathcal{V},A} \quad (3)$$

where $\mathcal{I}$ indexes the samples. The sequence $(s_i)_{i\in\mathcal{I}}$ is called the *ground truth of* $\mathcal{D}$. Finally, *music content analysis research* encompasses all aspects above in order to connect "users" (people, organisations, etc.) with music and information about music.

2. Experimental design in general

Experimental design (Bailey, 2008) is necessary to plan, implement, analyse, and report valid tests of hypotheses while meeting constraints that are physical, economical, ethical, and so on. Fundamental concerns of experimental design are the specification of the *treatments*, the *experimental* and *observational units* (*plots*), the identification of *structures* in treatments and plots, the mapping of units to treatments, the *relevance* of the measurement, the *modelling* and *analysis* of measurements, and securing that an experiment can validly address given hypotheses.

An experimental design maps plots $\{\omega_1, \ldots, \omega_N\}$ to treatments $\mathcal{T} = \{1, \ldots, t\}$, as represented by a design matrix $\mathbf{X} := [\mathbf{u}_1, \cdots, \mathbf{u}_t]_{N\times t}$, of $N \times 1$ indicator vectors $\mathbf{u}_i$, where the jth row of $\mathbf{u}_i$ is 1 if plot ω_j receives treatment i, and 0 if not. Of interest are two subspaces: the *treatment subspace* $V_T := C(\mathbf{X})$, which is the column space of $\mathbf{X}$, and its orthogonal complement $V_T^\perp$, which contains all $N \times 1$ vectors orthogonal to $\mathbf{X}$. *Structure* in the treatments and/or plots may prompt other decompositions of the space $V = \mathbb{R}^N$ in order to test specific hypotheses.

An experiment results in measurements on treated units, producing N *responses* $\mathbf{y} \in V$. The effect of the treatments on $\mathbf{y}$ is often described by a linear model

$$\mathbf{Y} = \boldsymbol{\tau} + \mathbf{Z} \quad (4)$$

where $\boldsymbol{\tau} = \sum \tau_i \mathbf{u}_i \in V_T$ and $\tau_1, \ldots, \tau_t$ are the *treatment parameters* (unknown constants). Moreover, $\mathbf{Z}$ is a random vector which can be modelled further to reflect structure in the plots and random error. Testing hypotheses about $\{\tau_i\}$ relies upon assumptions about $\mathbf{Z}$. The *simple textbook model* for *unstructured* plots and a *completely randomised design* assumes that $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$. Structure in the plots necessitates the use of a different model for $\mathbf{Z}$. Hypothesis testing by the method of *analysis of variance* de-

composes $\mathbf{y}$ over orthogonal subspaces of V, such as V_T and $V_T^{\perp}$, in accordance with the hypotheses and the plot structure, and then assesses their contributions.

3. Experimental design for MCA systems

Of principal interest is a comparison of music content analysis (MCA) systems $\{\mathcal{S}_1, \mathcal{S}_2, \ldots\}$, as in (2), which are the treatments $\mathcal{T}$. A common experiment toward this proceeds:

1. Build or choose a $\mathcal{D}$, indexed by $\mathcal{I}$
2. Partition $\mathcal{I}$ into non-overlapping sets $\mathcal{I}_{\text{train}}, \mathcal{I}_{\text{test}}$
3. Build $\{\mathcal{S}_1, \mathcal{S}_2, \ldots\}$ from tuples of $\mathcal{D}$ indexed by $\mathcal{I}_{\text{train}}$
4. Treat recordings of $\mathcal{D}$ indexed by $\mathcal{I}_{\text{test}}$ and compare results to ground truth, i.e., record successes and failures.

This is often done using K-fold cross validation (KfCV), or repeated random partitioning, in which case steps 2-4 are repeated. To summarise the results of this experiment, a figure of merit (FoM) is computed, e.g., accuracy, which is then used to make comparisons between treatments.

This process is typified by the experiment of Tzanetakis & Cook (2002), which is one of the most cited articles in MCA. They use a $\mathcal{D}$ with 1000 recordings and a ground truth with 100 occurrences of each of the 10 elements (genre) in $\mathcal{S}_{\mathcal{V},A}$. They perform the experiment above with several systems using the same instance of $\mathcal{E}$ but different classifiers built from different methods. For each kind of system $\mathcal{S}_k$, they perform 100 repetitions of 10fCV in $\mathcal{D}$. Each repetition uses a fresh random partition of the 100 recordings per genre in $\mathcal{D}$ into 10 mutually exclusive subsets $A_1^{(j)}, \ldots, A_{10}^{(j)}$ of size 10, where $j = 1, \ldots, 10$ identifies the genre. For $i = 1, \ldots, 10$ the ith fold is then defined as the set-theoretic union of $A_i^{(1)}, \ldots, A_i^{(10)}$. Each 10fCV uses for each $i \in \{1, \ldots, 10\}$ the ith fold as $\mathcal{I}_{\text{test}}$ and the union of the remaining nine folds as $\mathcal{I}_{\text{train}}$. For each $(r_i, s_i) \in \mathcal{D}$ with $i \in \mathcal{I}_{\text{test}}$ the response is 1 if the system output matches the ground truth, that is if $\mathcal{S}_k(r_i) = s_i$, and 0 otherwise. This gives 1000 responses per repetition and a corresponding proportion of matches. Tzanetakis and Cook compute the mean rate of success (accuracy) and the standard deviation of those proportions over the 100 repetitions. The table below shows some of their results.

Classifier Kind	FoM (accuracy) (mean $\pm$ std. dev.)
GS	0.59 ± 0.04
GMM(3)	0.61 ± 0.04
KNN(3)	0.60 ± 0.04

Three conclusions they make from their experiment are:

A. Since $\mathcal{D}$ is a "representative" sample from $\mathcal{R}_\Omega \times \mathcal{S}_{\mathcal{V},A}$, the FoM (based on $\mathcal{D}$) of GS, GMM(3) and KNN(3) in Table 3 are "indicative" of the FoM that would be obtained if these kinds of systems were applied to the entire universe $\mathcal{R}_\Omega$.

B. Since the FoM of GS, GMM(3) and KNN(3) using the proposed $\mathcal{E}$ are better than randomly mapping $\mathcal{R}_\Omega$ to $\mathcal{S}_{\mathcal{V},A}$ (in which case the FoM is expected to be 0.10), the set $\mathcal{S}_{\mathbb{F},A'}$ is informative for $\mathcal{S}_{\mathcal{V},A}$ in the whole universe $\mathcal{R}_\Omega$.

C. Since the FoM in Table 3 are better than randomly mapping $\mathcal{R}_\Omega$ to $\mathcal{S}_{\mathcal{V},A}$, the systems are recognising $\mathcal{S}_{\mathcal{V},A}$ in $\mathcal{R}_\Omega$.

The *validity* of each of these conclusions relies on two strong assumptions. Conclusion A assumes that $\mathcal{D}$ is a random sample from $\mathcal{R}_\Omega \times \mathcal{S}_{\mathcal{V},A}$, a set that is never explicitly defined. Since there is clear evidence that $\mathcal{D}$ is not a random sample (Sturm, 2014), the experiment may or may not result in FoM that reflect the performance of a system applied to $\mathcal{R}_\Omega$. The other two conclusions relate to the behaviour expected of a system operating randomly, and assume that an MCA system's ability to reproduce the ground truth of $\mathcal{D}$ is either due to chance or *caused* by "musical content" relating $\mathcal{R}_\Omega$ to $\mathcal{S}_{\mathcal{V},A}$. This assumption is not true: another way to reproduce the ground truth of $\mathcal{D}$ is by exploiting "non-musical content" (characteristics of subsets of $\mathcal{R}_\Omega$ unrelated to $\mathcal{S}_{\mathcal{V},A}$), e.g., a *confounding* introduced by the construction of $\mathcal{D}$. The existence of such confounding in the dataset used in Tzanetakis & Cook (2002), as well as others, has been clearly demonstrated (Pampalk et al., 2005; Sturm, 2014). Because of its lack of control over such a possibility, the experiment above has no validity for conclusions B and C.

A possible measurement model for the experiment in line with Bailey (2008) would recognise the individual cross validations as the observational units and "repetitions" as an additional plot factor. A system of each kind k is applied as a treatment to 100 whole repetitions (each consisting of 10 units). The response for each unit u is the proportion of correct classifications of the 100 pairs (r_i, s_i) with $i \in \mathcal{I}_{\text{test}}(u, v)$, where $\mathcal{I}_{\text{test}}(u, v)$ denotes the test set for the validation u in repetition v. Denote by p_{kuv} the probability that a system of type k gives a correct classification for u and v. A model for this situation is

$$p_{kuv} = \beta_0 + \beta_k + \gamma_v + \epsilon_{uv}, \tag{5}$$

where, using an appropriate parameterisation, β_0 is the probability of success of a random system and β_k is the additional contribution of the system to type k. Further, γ_v is a contribution due to the random partitioning in repetition v, and ϵ_{uv} models all other "errors" independent of the system type due to cross validation instances u within v. For this interpretation of β_0 and β_k to be valid, the experiment would also need to include a random system.

To provide evidence for the conclusions B and C of Tzanetakis & Cook (2002) one would need to test the hypothesis $\beta_k = 0$ in (5). This test would then be performed in the stratum for repetitions. For three kinds of systems, as in the table above, and an additional random system (each with 100 repetitions and 10 validations) and assuming normality, the corresponding F test has 396 denominator degrees of freedom. Assuming all individual cross validations are independent can result in an incorrect test overstating the significance of the results. The applicability of this is suspect, however, since γ_v is known to depend on the system type (Pampalk et al., 2005; Sturm, 2014).

Bailey, R. A. *Design of comparative experiments*. Cambridge University Press, 2008.

Pampalk, E., Flexer, A., and Widmer, G. Improvements of audio-based music similarity and genre classification. In *Proc. ISMIR*, pp. 628–233, Sep. 2005.

Sturm, B. L. The state of the art ten years after a state of the art: Future research in music information retrieval. *J. New Music Research*, 43(2):147–172, 2014.

Sturm, B. L., Bardeli, R., Langlois, T., and Emiya, V. Formalizing the problem of music description. In *ISMIR*, pp. 89–94, 2014.

Tzanetakis, G. and Cook, P. Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process.*, 10(5):293–302, July 2002.