ON EVALUATION IN MUSIC AUTOTAGGING: A NOTION OF VALIDITY

B. L. Sturm
Aalborg University Copenhagen
bst@create.aau.dk

F. Gouyon, J. Oliveira, N. Hespanhol
INESC Porto
fgouyon@inescporto.pt

T. Langlois
Lisbon University
tl@di.fc.ul.pt

ABSTRACT

An evaluation approach in music autotagging research must be valid in order to draw any reasonable conclusion about how well a system performs in pairing music signals with tags in a "meaningful" way. Since the concept of validity has yet to be defined within autotagging evaluation, we formalize a notion of it here.

1. INTRODUCTION

The importance of validity in the design of experiments cannot be understated: validity determines the scope of what one can logically conclude from an experiment regardless of its outcome [1, 10]. Since evaluation in music information retrieval (MIR) research is nothing more than experiments, it is essential to consider their validity [8, 10]. The meaningfulness of many more published evaluation results hangs in the balance. For instance, what can be logically concluded from the results of a decade of evaluation at MIREX?

As a more specific example, consider an area of MIR research that has attracted much work in recent years, "music autotagging" [2, 4, 9]: the automatic and accurate assignment of a variety of labels (tags) that are meaningful to some users to elements of a music collection. Music autotagging can facilitate the search and retrieval of music in the collection by using simple textual queries. Many publications in music autotagging research report for systems figures of merit (FoM) computed from tests in benchmark datasets, such as CAL500 [9]. However, is such an evaluation approach valid for concluding how accurately an autotagging system can assign tags in the real world? Is such an evaluation approach valid for concluding that one of several autotagging systems is more accurate than others for assigning tags in the real world?

Published autotagging research typically report FoM that are significantly better than that expected of a random system [4, 9]. This result is surprising for a few reasons. First, the tags present in benchmark datasets are composed of many vague concepts, some of which are extrinsic to a music recording. For instance, most tags include terms that appear indicative of genre, "locale" and moods [2]. How can an autotagging system thus appear to perform much better than that expected of a random system? Second, it seems reasonable that the measured performance of a music genre recognition system would indicate the best performance of an autotagging system that predicts genre *and* emotion *and* "locale" *and* any number of other kinds of tags. However, recent work points to the conclusion that "improvement" to genre recognition and emotion recognition systems has been an illusion due to experiments that are not valid with respect to such conclusions [8]. A music genre recognition system can have a high FoM from an evaluation in spite of having no capacity to recognize genre in music. A serious question thus arises: are the evaluation approaches commonly used in music autotagging research valid for concluding that an autotagging system will be successful in the real world, that one autotagging system will be more successful than another system in the real world, or that progress in developing and improving autotagging systems has even been made?

In this paper, we focus on evaluation in music autotagging research, and attempt to clarify its objectives. After we briefly review approaches to autotagging evaluation, we discuss and formalize the notion of validity in autotagging evaluation with respect to its objectives.

2. BRIEF REVIEW OF MUSIC AUTOTAGGING

Given the appearance of book chapters (e.g., [2]), several journal articles (e.g., [9]) and conference papers (e.g, [6]), PhD theses (e.g., [7]), tutorials (ISMIR 2013), as well as six years of the MIREX "Audio Tag Classification" task (ATC), [1] music autotagging is an established problem in MIR. Turnbull et al. [9] discuss music autotagging as multi-label classification; and Seyerlehner et al. [6] describe it as "transform[ing] an audio feature space into a semantic space, where music is described by words." We adopt the same notions here by defining a music autotagging system as one that "meaningfully" (defined by a user [5]) *annotates*, i.e., assigns words, to recorded music.

A standard approach to music autotagging evaluation is having a system annotate a set of test signals, and then comparing the resulting tags to the "ground truth." More specifically, the experimentalist counts the number of true positives, false positives, true negatives, and false negatives of each tag. From these, several FoM are computed and reported, such as "Average Tag Recall," "Average Tag

[1] http://www.music-ir.org/mirex/wiki/MIREX_HOME

Precision," "Average Tag F-Measure." Between 2008 – 2012, ATC employed this approach to systematically and rigorously evaluate using standardized datasets about 60 music autotagging systems. This evaluation procedure also appears in many other works, e.g., [9]. A fundamental aspect of evaluation is test data. A variety of benchmark datasets appear in autotagging research, e.g., CAL500, MagnaTagatune, and the Million Song Dataset, MajorMiner and USPOP. In the music autotagging literature, systems are most often publicized and compared according to FoM (e.g., accuracy, precision, F-score) in some of these benchmark datasets.

The standard evaluation in music autotagging is believed to provide a meaningful picture of how well a system has "learned" to identify a set of concepts inherent to a dataset, or an idea of how well it will perform in the real world. Hence, a critical question to answer is whether the evaluation approach common to music autotagging is valid for such conclusions.

3. VALIDITY IN AUTOTAGGING EVALUATION

Denote a music autotagging system S, and a set of tags $\mathcal{T}$. It is important to note that S is already built, and ready to run by a user wishing to apply $\mathcal{T}$ to their music. Denote by $\Gamma_S(t)$ the "degree of understanding" of a $t \in \mathcal{T}$ by S. This describes how well S is expected to perform in using t (or not) to annotate music in the real world. One wishes of course for $\Gamma_S(t)$ to be good for all $t \in \mathcal{T}$.

3.1 What is desired from an evaluation?

A major problem comes when $\Gamma_S(t)$ is not observable, in which case it must be inferred from something observable. What is desired from the evaluation of a music autotagging system S is a reliable estimate of its degrees of understandings of the set of tags it has learned, i.e., $\{\Gamma_S(t)|t \in \mathcal{T}\}$. With this knowledge, one can judge whether a system can meet the requirements of a use case, whether one system is better than another, and finally whether progress is being made in solving the problems addressed by research in music autotagging.

Music autotagging research addresses the unobservability of $\Gamma_S(t)$ by estimating it from FoM produced from an evaluation of S applied to a test dataset of recorded music signals and their tags. Denote a test dataset as a set of tuples $\Phi = \{(\mathbf{x}_i, \mathcal{T}_i)\}_{i\in\mathcal{I}}$, where $\mathbf{x}_i$ is a signal, $\mathcal{T}_i \subseteq \mathcal{T}$ are its "ground truth" tags, and $\mathcal{I}$ indexes Φ. By comparing the output of S to the "ground truth" of Φ, one computes a FoM $\widehat{\Gamma}_S(t;\Phi)$. The implicit assumption of standard evaluation is thus that it is a valid indicator of $\Gamma_S(t)$. In other words, that when $\widehat{\Gamma}_S(t;\Phi)$ is high, or better than that of another system (perhaps tested using a formal statistical test), then this means S is "working" or "working better."

3.2 Valid indicator of performance

We define a music autotagging evaluation to be a *valid indicator of performance* when for any S

$$[\widehat{\Gamma}_S(t;\Phi) \text{ good}] \Leftrightarrow [\Gamma_S(t) \text{ high}] \quad (1)$$

and when, for any two systems S_1, S_2

$$[\widehat{\Gamma}_{S_1}(t;\Phi) \text{ better than } \widehat{\Gamma}_{S_2}(t;\Phi)] \Leftrightarrow [\Gamma_{S_1}(t) \text{ higher than } \Gamma_{S_2}(t)] \quad (2)$$

where $\Leftrightarrow$ is logical equivalence. In other words, (1) says a valid evaluation produces a good FoM of any S on t if and only if that S has a high "degree of understanding" of t; and (2) says a valid evaluation produces a better FoM for any S_1 than for any other S_2 on t if and only if S_1 has a higher "degree of understanding" than S_2 of t. The notions of "good" and "better" rely on the quality of the estimate $\widehat{\Gamma}_S(t;\Phi)$ [3]. If (1) and (2) do not hold for a $t \in \mathcal{T}$, then that evaluation is not a valid indicator of performance of any music autotagging system on that concept — no matter the FoM that results. The fundamental question is now no longer, "How good is $\widehat{\Gamma}_S(t;\Phi)$?", or, "Is $\widehat{\Gamma}_{S_1}(t;\Phi)$ significantly better than $\widehat{\Gamma}_{S_2}(t;\Phi)$?", but now, "Is this evaluation a valid indicator of performance?"

4. REFERENCES

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

[2] T. Bertin-Mahieux, D. Eck, and M. Mandel. Automatic tagging of audio: The state-of-the-art. In W. Wang, editor, *Machine Audition: Principles, Algorithms and Systems*. IGI Publishing, 2010.

[3] E. R. Dougherty and L. A. Dalton. Scientific knowledge is possible with small-sample classification. *EURASIP J. Bioinformatics and Systems Biology*, 2013:10, 2013.

[4] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green. Automatic generation of social tags for music recommendation. In *Proc. Neural Info. Process. Syst. Conf.*, 2008.

[5] M. Schedl, A. Flexer, and J. Urbano. The neglected user in music information retrieval research. *J. Intell. Info. Systems*, 2013 (in press).

[6] K. Seyerlehner, G. Widmer, M. Schedl, and P. Knees. Automatic music tag classification based on block-level features. In *Sound and Music Computing Conf.*, 2010.

[7] M. Sordo. *Semantic Annotation of Music Collections: A Computational Approach*. PhD thesis, Universitat Pompeu Fabra, 2012.

[8] B. L. Sturm. 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.

[9] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Trans. Audio, Speech, Lang. Process.*, 16, 2008.

[10] J. Urbano, M. Schedl, and X. Serra. Evaluation in music information retrieval. *J. Intell. Info. Systems*, 41(3):345–369, Dec. 2013.