CROSS-COLLECTION EVALUATION FOR MUSIC CLASSIFICATION TASKS

Dmitry Bogdanov, Alastair Porter, Perfecto Herrera, Xavier Serra
Music Technology Group, Universitat Pompeu Fabra
name.surname@upf.edu

ABSTRACT

Many studies in music classification are concerned with obtaining the highest possible cross-validation result. However, some studies have noted that cross-validation may be prone to biases and that additional evaluations based on independent out-of-sample data are desirable. In this paper we present a methodology and software tools for cross-collection evaluation for music classification tasks. The tools allow users to conduct large-scale evaluations of classifier models trained within the AcousticBrainz platform, given an independent source of ground-truth annotations, and its mapping with the classes used for model training. To demonstrate the application of this methodology we evaluate five models trained on genre datasets commonly used by researchers for genre classification, and use collaborative tags from Last.fm as an independent source of ground truth. We study a number of evaluation strategies using our tools on validation sets from 240,000 to 1,740,000 music recordings and discuss the results.

1. INTRODUCTION

Music classification is a common and challenging Music Information Retrieval (MIR) task, which provides practical means for the automatic annotation of music with semantic labels including genres, moods, instrumentation, and acoustic qualities of music. However, many researchers limit their evaluations to cross-validation on small-sized datasets available within the MIR community. This leaves the question of the practical value of these classifier models for annotation, if the goal is to apply a label to any unknown musical input.

In the case of genre classification, Sturm counts that 42% of studies with experimental evaluation use publicly available datasets, including the famous GTZAN music collection (23%, or more than 100 works) and the ISMIR-2004 genre collection (17.4%), both of which contain no more than 1000 tracks. There is some evidence that such collections sizes are insufficient [7]. The MIREX annual evaluation includes a genre classification task, [1] which is currently run on a collection of 7000 tracks annotated with 10 genres, which is not publicly available to researchers. Some studies employ larger datasets, annotating genre using web-mined sources [11, 17, 18]. It has been shown that some datasets have flaws, including inconsistent and controversial labeling, absence of artist filters, and presence of duplicate examples (for example, GTZAN [20]).

Cross-validation is routinely used as a method for approximating a classifier's performance in real-world conditions, but such estimation is not free from some pitfalls. The ability of trained classifier models to generalize remains under question when following this method [14]. As some studies have noted, typical k-fold cross-validation on small-sized collections is prone to biases and additional evaluations based on independent out-of-sample data are desirable in order to avoid them [2, 8, 12]. Cross-collection validation is also suggested in other domains [1, 3–5, 13].

In order to address these problems and be able to better assess the performance of music classifier models, we propose a cross-collection evaluation process, that is, an evaluation of models on independent sets of music tracks annotated with an independent ground-truth source (which we call *validation sets*). In this paper we present a methodology and software tools for such evaluation for music classification tasks. We use AcousticBrainz, [2] a community-based platform for gathering music information from audio [16]. It contains MIR-related music features for over 3 million recordings including duplicates (we use the term *recording* to refer to a single analysis of music track). It provides the functionality to create datasets consisting of recordings and associated ground truth, training classifier models, and applying them to recordings present in AcousticBrainz. A number of models trained on genre datasets used within MIR are already included. Our tools allow the AcousticBrainz community to conduct cross-collection evaluations of classifier models trained on the AcousticBrainz website given any independent source of ground-truth annotations for recordings and a mapping between a model's classes and the classes within that ground truth.

In order to demonstrate the proposed methodology and tools, we evaluate five genre classifier models trained on MIR genre datasets. We build a genre ground truth for recordings in AcousticBrainz using collaborative tags from

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

[2] https://acousticbrainz.org

Last.fm [3] and consider various evaluation strategies for mapping the classifier models' outputs to the ground-truth classes. We use our tools on validation sets from 240,000 to 1,740,000 recordings and discuss the obtained results. Finally, we publish our genre ground-truth dataset on our website.

2. CROSS-COLLECTION EVALUATION METHODOLOGY

We define cross-collection evaluation to be an evaluation of a classifier model on tracks in a validation set annotated with an independent ground-truth source. We propose that the validation set is obtained or collected from a different source to the data used for training. This is distinct from holdout validation where a part of a dataset is used to verify the accuracy of the model. Because of the data's different origin it provides an alternative view on the final performance of a system. We develop tools based on this methodology for use in AcousticBrainz, because of the commodity of its infrastructure for building and evaluating classifiers, and the large amount of music recordings which it has analyzed.

2.1 Evaluation strategies

We consider various evaluation strategies concerning the comparison of a classifier's estimates and the ground-truth annotations in a validation set. A direct mapping of classes between the ground truth and the classifier is not always possible due to differences in their number, names, and actual meaning. Therefore, it is necessary to define a mapping between classes. Class names may imply broad categories causing difficulties in determining their actual coverage and meaning, and therefore inspection of the contents of the classes is advised. The following cases may occur when designing a mapping: a) a class in the classifier can be matched directly to a class in the validation set; b) several classes in the classifier can map to one in the validation set; c) one class in the classifier can map to many in the validation set; d) a class in the validation set cannot be mapped to any class in the classifier; e) a class in the classifier cannot be mapped to any class in the validation set. The case d) represents the subset of the validation set which is "out of reach" for the classifier in terms of its coverage, while e) represents the opposite, where the model is able to recognize categories unknown by the ground truth. We show an example of such a mapping in Section 3.3. The design of the mapping will affect evaluation results.

Validation sets may vary in their size and coverage and may contain a wider range of annotations than the classifier being evaluated. We consider the following strategies:

- **S1**: Use only recordings from the validation set whose ground truth has a matching class in the classifier. For example, if a recording is only annotated with the class electronic, and this class does not appear in the classifier, we discard it.
- **S2**: Use all recordings in the validation set and treat recordings from classes that do not exist in the classifier as an incorrect classification.

The validation set may have multiple class annotations per recording (e.g., in case of genre annotations, both pop and rock could be assigned to the same recording). Where the validation set has more than one ground-truth class for a recording we consider different methods of matching these classes to classifiers' estimates:

- **ONLY**: Only use recordings that have one ground-truth class, and discard the rest of the recordings when computing evaluation metrics.
- **ALL**: When a recording has more than one ground-truth class, accept an estimate as correct if it matches any of them, even though for the rest of the classes it would be considered a misclassification.

There may be duplicate recording representing the same music track (as is the case for AcousticBrainz, for which inconsistent classifier estimates have been observed). We consider two ways of dealing with them:

- **D1**: Remove all recordings that have duplicates from the evaluation.
- **D2**: Treat all recordings independently.

2.2 Evaluation metrics

Using class mappings one can compute confusion matrices for a classifier model for all combinations of S1/S2 with ONLY/ALL and D1/D2. The confusion matrix counts the percentage of correct classifier class estimates for each ground-truth class in the validation set. When a recording has more than one ground-truth class in method ALL, the recording is counted in all associated classes. Results are combined in the case when a class in the model is mapped to more than one class in the validation set. We estimate *accuracy*, the percentage of correctly recognized recordings. This value can be skewed due to difference in the sizes of each ground-truth class, and therefore we also compute *normalized accuracy* by scaling match counts according to the number of recordings within each class.

2.3 Tools for cross-collection evaluation

We have developed a set of tools as part of AcousticBrainz which let users train and evaluate classifier models. [4] Our tools let a user evaluate the quality of this model using an independent validation set. They can conduct any of the evaluation strategies mentioned above for any classifier model trained using AcousticBrainz.

To use our tools, a user first creates two datasets in AcousticBrainz. They define one dataset to be used to train a model, and the other to be used as the validation set. To ensure reliability of the accuracy results, the user can perform artist filtering [6, 15] during both the training and the

[3] http://www.last.fm

[4] We use the existing model training process, which selects best SVM parameters in a grid search using cross-validation and trains a model using all the data [10]. More details on the model training process are provided at https://acousticbrainz.org/datasets

cross-collection evaluation process. With artist filtering, during training we randomly select only one recording per artist for the model. On cross-collection evaluation, only recordings by artists different from the ones present in the training data are used. To get the artist of a recording we use the MusicBrainz API to request artist information for a recording using its MBID. Classes are randomly truncated to the same size, with a lower limit of 200 instances per class.

An interface lets the user map classes from the classifier dataset to classes in the validation set. In the case that there are no suitable matches, a class can be discarded from consideration during evaluation. The tools generate a report including statistics on the employed dataset and ground truth, accuracy values and confusion matrices. The results can be exported as HTML, LaTeX, or other machine-readable formats for further use. The tool is integrated into the AcousticBrainz server [5] to make this cross-collection evaluation process available for all AcousticBrainz users.

3. EVALUATION OF GENRE CLASSIFIER MODELS

We applied our evaluation methodology to assess five genre classifiers, three of which were already available within AcousticBrainz. We built two more classifiers using the dataset creator on the AcousticBrainz website with two previously published sources for genre annotation. We built an independent ground truth by mining folksonomy tags from Last.fm.

3.1 Music collections and classifier models

- **GTZAN Genre Collection (GTZAN)**. [6] A collection of 1000 tracks for 10 music genres (100 per genre) [23], including blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. The GTZAN collection has been commonly used as a benchmark dataset for genre classification, due to it being the first dataset of this kind made publicly available [22].
- **Music Audio Benchmark Data Set (MABDS)**. [7] A collection of 1886 tracks retrieved from an online music community and classified into 9 music genres (46–490 per genre) [9] including alternative, electronic, funk/soul/rnb, pop, rock, blues, folk/country, jazz, and rap/hiphop.
- **AcousticBrainz Rosamerica Collection (ROS)**. A collection of 400 tracks for 8 genres (50 tracks per genre) including classical, dance, hip-hop, jazz, pop, rhythm & blues, rock, and speech (which is more a type of audio than a musical genre). The collection was created by a professional musicologist [7].
- **MSD Allmusic Genre Dataset (MAGD)**. A collection of genre annotations of the Million Song Dataset, derived from AllMusic [17]. We mapped MAGD to the AcousticBrainz collection, [8] which reduced the size of the dataset from 406,427 to 142,969 recordings and applied an artist filter (keeping only one recording per artist). The resulting dataset used for training contained 11 genres (206 tracks per genre) including pop/rock, electronic, rap, jazz, rnb, country, international, latin, reggae, vocal, and blues. [9]
- **Tagtraum genre annotations (TAG)**. A collection of genre annotations for Million Song Dataset, derived from Last.fm and beaTunes [18] (CD2C variation). As with MAGD, we mapped the dataset (reducing the size from 191,408 to 148,960 recordings) and applied an artist filter, resulting in 13 genres including rock, electronic, pop, rap, jazz, rnb, country, reggae, metal, blues, folk, world and latin (228 tracks per genre). [10]

Classifier	Accuracy	Normalized accuracy	Random baseline	Size	Number of classes
GTZAN	75.52	75.65	10	1000	10
MABDS	60.25	43.5	11.1	1886	9
ROS	87.56	87.58	12.5	400	8
MAGD	47.75	47.75	9.09	2266	11
TAG	47.87	47.87	7.69	2964	13

Table 1: Cross-validation accuracies (%) for all classifier models.

The models for GTZAN, MABDS, and ROS are provided in AcousticBrainz as baselines for genre classification but these collections were trained without artist filtering, as the recordings in these datasets are not associated with MBIDs. [11] We inspected available metadata for these collections and gathered non-complete artist lists for artist filtering in our cross-collection evaluation. We were able to identify 245 artists for 912 of 1000 recordings for GTZAN [19], 1239 artists for all 1886 recordings for MABDS, and 337 artists for 365 of 400 recordings for ROS.

Table 1 presents accuracies and normalized accuracies for the considered models obtained from cross-validation on training. The accuracies for GTZAN, MABDS, and ROS models are reported on the AcousticBrainz website. In general we observe medium to high accuracy compared to the random baseline. The results obtained for GTZAN are consistent with 78–83% accuracy observed for a number of other state-of-the-art methods on this collection [21]. Confusion matrices [12] do not reveal any significant misclassifications except for MABDS, for which alternative, funk/soul/rnb, and pop classes were frequently misclassified as rock (more than 35% of cases).

3.2 Preparing an independent ground truth

Last.fm contains tag annotations for a large number of music tracks by a community of music enthusiasts. While these tags are freeform, they tend to include commonly

[5] https://github.com/metabrainz/acousticbrainz-server

[6] http://marsyasweb.appspot.com/download/data_sets

[7] http://www-ai.cs.uni-dortmund.de/audio.html

[8] http://labs.acousticbrainz.org/million-song-dataset-mapping

[9] New age and folk categories were discarded due to insufficient number of instances after artist filtering keeping equal number of recordings per class.

[10] Similarly, new age and punk categories were removed.

[11] https://acousticbrainz.org/datasets/accuracy

[12] Available online: http://labs.acousticbrainz.org/ismir2016

recognized genres. We used the Last.fm API to get tag names and counts for all recordings in the AcousticBrainz database. Tag counts for a track are weighted, where the most applied tag for the track has a weight of 100. As tags are collaborative between users, we expect them to represent the "wisdom of the crowds". We obtained tags for 1,031,145 unique music tracks. Data includes 23,641,136 tag-weight pairs using 781,898 unique tags.

Genre tags as entered by users in Last.fm vary in their specificity (e.g., "rock", "progressive rock"). Meanwhile, the classifiers that we evaluate estimate broad genre categories (according to their class names). Therefore, we matched Last.fm tags to genres, and grouped these genres to broader "top-level" genres which we then matched to the output classes of our classifiers. We used a tree of genres [13] from beets, [14] a popular music tagger which uses data from MusicBrainz, Last.fm, and other sources to organize personal music collections. This genre tree is constructed based on a list of genres on Wikipedia, [15] further moderated by the community to add a wider coverage of music styles. [16] The tree includes 16 top-level genres: african, asian, avant-garde, blues, caribbean and latin american, classical, country, easy listening, electronic, folk, hip hop, jazz, pop, rhythm & blues, rock, and ska. [17] The taxonomy provided by the genre tree may not always be grounded on acoustical/musical similarity. For example, the asian top-level category includes both traditional music and western-influenced styles such as j-pop; jazz includes bebop, free jazz, and ragtime; electronic includes both electroacoustic music and techno. Similar inconsistencies in the contents of classes are not uncommon in genre datasets [18], and have been observed in GTZAN by [20], and also in our informal reviews of other datasets.

We matched tags directly to genres or subgenres in the tree and then mapped them to their top-level genre. Weights were combined for multiple tags mapped to the same top-level genre. Unmatched tags were discarded. We removed tracks where the weight of the top tag was less than 30, normalized the tags so that the top tag had a weight of 100 and again discarded tags with a weight of less than 30. After the cleaning process, we gathered genre annotations for 778,964 unique MBIDs, corresponding to 1,743,674 AcousticBrainz recordings (including duplicates).

3.3 Genre mappings

We created mappings between the classes of the classifier models and the validation set (Table 2). We created no mapping for 'disco' on GTZAN, 'international' and 'vocal' on MAGD, and 'world' on TAG as there were no clear matches. For ROS we did not map 'speech' as it did not represent any musical genre. Recordings estimated by classifiers as these classes were ignored during evaluation.

[13] https://github.com/beetbox/beets/blob/0c7823b4/beetsplug/lastgenre/genres-tree.yaml

[14] http://beets.io

[15] https://en.wikipedia.org/wiki/List_of_popular_music_genres

[16] The list from Wikipedia contains only modern popular music genres

[17] We discarded the comedy and other genres

3.4 Results

Using our tools, we performed an evaluation using sets from 240,000 to 1,740,000 recordings. Table 3 presents the accuracies and number of recordings used for the evaluation of each classifier model. We bring attention to the S2-ONLY-D1 strategy as we feel that it reflects a real world evaluation on a variety of genres, while being relatively conservative in what it accepts as a correct result (the ONLY-D1 variation). Also of note is the S1-ONLY-D1 strategy as it evaluates the classifiers on a dataset which reflects their capabilities in terms of coverage. We present confusion matrices for the S2-ONLY-D1 strategy in Table 4 for MAGD and TAG models (confusion matrices differ little across all evaluation strategies according to our inspection). [18]

Inspection of confusion matrices revealed a few surprising genre confusions. The MABDS model confuses all ground-truth genres with electronic (e.g., blues 64%, folk 62% of recordings misclassified). This tendency is consistent with inspection of this model's estimates in the AcousticBrainz database (81% estimated as electronic). No pattern of such misclassification was present in the confusion matrix during the training stage. Although this model was trained on unbalanced data, electronic was among the smallest sized classes (only 6% of the MABDS collection). Similarly, the GTZAN model tends to estimate all music as jazz (>73% of recordings of all genres are misclassified), which is again consistent with genre estimates in AcousticBrainz (90% estimated as jazz), with no such problems found during training.

The ROS model does not misclassify genres as harshly, confusing pop with rhythm & blues (26%), jazz with classical (21%), electronic with hip hop and rhythm & blues, jazz with rhythm & blues, and rhythm & blues with pop (<20% of cases for all confusions). For the MAGD model we see misclassifications of pop with rhythm & blues (21%), pop with carribean & latin and country, and rhythm & blues with blues (<15%). The TAG model performed better than MAGD, with no genre being misclassified for another more than 15% of the time, though we see a moderate amount of blues, electronic, folk, and pop instances being confused with rock as well as rhythm & blues with blues. The confusions for all three models make sense from musicological and computational points of view, evidencing how controversial genre-tagging can be, and that the computed features may not be specific enough to differentiate between genre labels.

3.5 Discussion

In general, considering exclusion/inclusion of the duplicate recordings in the evaluation (D1/D2), we observed that the differences in accuracy values are less than 4 percentage points for all models. We conclude that duplicates do not create any strong bias in any of our evaluation strategies even though the sizes of D1/D2 testing sets vary a lot.

[18] Complete results for all classifier models are available at: http://labs.acousticbrainz.org/ismir2016

Ground truth	GTZAN	MABDS	ROS	MAGD	TAG
african	-	-	-	-	-
asian	-	-	-	-	-
avant-garde	-	-	-	-	-
blues	blues	blues	-	blues	blues
caribbean and latin american	reggae	-	-	latin, reggae	latin, reggae
classical	classical	-	classical	-	-
country	country	folk/country	-	country	country
easy listening	-	-	-	-	-
electronic	-	electronic	dance	electronic	electronic
folk	-	folk/country	-	-	folk
hip hop	hip-hop	rap/hiphop	hip-hop	rap	rap
jazz	jazz	jazz	jazz	jazz	jazz
pop	pop	pop	pop	pop/rock	pop
rhythm and blues	-	funk/soul/rnb	rnb	rnb	rnb
rock	rock, metal	rock, alternative	rock	pop/rock	metal, rock
ska	-	-	-	-	-

Table 2: Mapping between classifier classes and ground-truth genres.

Model	Strategy	Recordings	Accuracy	Normalized accuracy
GTZAN	S1-ALL-D1	274895	13.68	12.78
	S1-ALL-D2	1235692	10.72	12.73
	S1-ONLY-D1	242346	13.27	12.95
	S1-ONLY-D2	1053670	10.35	12.91
	S2-ALL-D1	373886	10.06	6.39
	S2-ALL-D2	1623809	8.16	6.37
	S2-ONLY-D1	**292840**	**8.99**	**6.42**
	S2-ONLY-D2	1214253	6.93	6.37
MABDS	S1-ALL-D1	361043	31.76	17.52
	S1-ALL-D2	1660333	31.13	16.75
	S1-ONLY-D1	292220	28.39	18.44
	S1-ONLY-D2	1277695	27.44	17.96
	S2-ALL-D1	386945	29.63	9.86
	S2-ALL-D2	1743034	29.65	9.42
	S2-ONLY-D1	**302448**	**26.41**	**10.40**
	S2-ONLY-D2	1302343	25.82	10.15
ROS	S1-ALL-D1	320398	51.73	47.36
	S1-ALL-D2	1518024	50.12	45.32
	S1-ONLY-D1	269820	50.46	52.52
	S1-ONLY-D2	1229136	48.82	51.72
	S2-ALL-D1	379302	43.70	20.72
	S2-ALL-D2	1683696	45.19	19.83
	S2-ONLY-D1	**296112**	**43.12**	**23.58**
	S2-ONLY-D2	1252289	45.19	23.24
MAGD	S1-ALL-D1	323438	59.35	42.13
	S1-ALL-D2	1505105	59.91	40.41
	S1-ONLY-D1	265890	59.56	48.34
	S1-ONLY-D2	1184476	60.83	48.40
	S2-ALL-D1	347978	55.92	23.70
	S2-ALL-D2	1590395	57.36	22.73
	S2-ONLY-D1	**272426**	**56.36**	**27.35**
	S2-ONLY-D2	1187287	58.94	27.56
TAG	S1-ALL-D1	327825	59.85	44.46
	S1-ALL-D2	1482123	60.58	42.12
	S1-ONLY-D1	265280	59.51	52.97
	S1-ONLY-D2	1139297	60.49	52.77
	S2-ALL-D1	342544	57.35	27.79
	S2-ALL-D2	1532129	58.67	26.32
	S2-ONLY-D1	**268543**	**56.94**	**33.13**
	S2-ONLY-D2	1147231	58.62	33.10

Table 3: Cross-collection evaluation accuracies (%) for all classifier models.

Similar conclusions can be made with respect to the inclusion of recordings with conflicting genre ground truth (ONLY/ALL). Conflicting cases of genre annotations only account for 21% of our validation set. In the case of our Last.fm annotations, multiple labeling of ground truth does not affect our results, still, one should explore both strategies to ensure that the same holds for other ground truths.

The only large difference in accuracy values is observed when comparing S1 and S2 strategies—S1 yields higher accuracies as all additional recordings in S2 are considered incorrect no matter what the classifier selects. Normalized accuracy allows us to assess the performance of a classifier given a hypothetical validation set with an equal number of instances per class. In our S2 strategy, many validation set classes not matched to a classifier class, and therefore considered incorrect, contained a small number of recordings (e.g., african, asian, avant-garde, and easy listening; see Table 4). Because of this we observe a larger difference in normalized accuracies between S1 and S2.

Based on the results we conclude that the models for ROS, MAGD and TAG perform the best. Their normalized accuracies are two times better than other classifiers under any condition. Interestingly, the ROS model is trained on the smallest collection (400 tracks, compared to 1000 tracks for GTZAN and 1886 tracks for MABDS, and over 2000 for MAGD an TAG), while we expected that it would suffer from insufficient training size.

What can be the reason for such a differing performances of models? MAGD uses as its source genre annotations made by experts from Allmusic, while the ROS collection was created by a musicologist specifically for the task of content-based music classification, which may be the reason for their better performance. The annotations in TAG were taken from two different sources, and were only used when both sources agreed on the genre [18].

4. CONCLUSIONS

The majority of studies on music classification rely on estimating cross-validation accuracy on a single ground truth, while it has been criticized as being shortsighted to shed light on the real capacity of a system to recognize music categories [21]. In our study we go beyond this approach and show an additional way to ascertain the capacity of classifiers by evaluating across collections. We believe that cross-collection generalization is an interesting metric to

Ground-truth genre	size (%)	Estimated genre: blues (blues)	carribean & latin (latin, reggae)	country (country)	electronic (electronic)	hip hop (rap)	jazz (jazz)	pop, rock (pop/rock)	rhythm & blues (rnb)
blues	2.7	**48.30**	8.54	11.27	2.88	1.31	8.94	13.16	5.61
carribean & latin	1.9	6.98	**57.11**	4.70	5.64	6.88	7.44	3.04	8.20
country	4.3	14.09	8.28	**61.89**	1.00	0.43	2.24	8.38	3.67
electronic	20.1	2.01	4.57	0.90	**59.28**	6.22	4.35	16.22	6.45
hip hop	2.3	1.22	10.39	0.14	8.77	**65.63**	1.08	3.80	8.97
jazz	7.7	9.39	7.00	3.68	2.87	1.02	**67.22**	4.60	4.22
pop	5.8	4.63	18.65	18.23	7.12	3.47	2.87	**23.98**	21.05
rhythm & blues	3.4	16.36	13.87	11.88	3.66	5.93	5.00	10.74	**32.56**
rock	43.9	7.60	6.26	6.10	7.14	0.82	2.48	**67.38**	2.23
african	0.3	13.09	33.54	10.22	8.38	6.13	9.20	4.50	14.93
asian	1.2	2.13	16.85	7.72	8.35	2.58	3.66	35.34	23.36
avant-garde	0.2	14.33	8.26	5.51	12.40	5.92	20.11	30.72	2.75
classical	2.2	4.85	4.05	5.73	4.30	0.45	66.05	13.86	0.70
easy listening	0.4	9.52	13.79	27.59	6.62	2.21	18.07	14.06	8.14
folk	3.1	15.55	14.80	28.92	5.75	0.89	10.71	20.14	3.24
ska	0.7	4.80	39.28	5.71	13.90	9.21	2.16	19.66	5.28

(a) MAGD. Columns for pop and rock are summed together as they match the same model class

Ground-truth genre	size (%)	Estimated genre: blues (blues)	carribean & latin (latin, reggae)	country (country)	electronic (electronic)	folk (folk)	hip hop (rap)	jazz (jazz)	pop (pop)	rhythm & blues (rnb)	rock (metal, rock)
blues	2.7	**48.09**	4.70	6.57	1.08	7.17	0.63	9.55	3.45	6.72	12.04
carribean & latin	1.9	3.47	**59.11**	3.26	2.93	4.37	5.56	6.96	4.47	6.77	3.12
country	4.3	11.58	5.29	**51.90**	0.37	11.15	0.15	2.55	6.13	3.18	7.70
electronic	20.1	0.82	4.53	0.81	**53.66**	5.45	5.82	2.62	9.33	3.33	13.63
folk	3.1	6.03	4.46	11.23	3.20	**47.22**	0.35	5.30	7.53	2.59	12.09
hip hop	2.3	0.83	9.14	0.11	7.03	0.31	**67.00**	1.52	2.30	8.75	3.00
jazz	7.7	7.68	5.19	2.15	1.43	4.93	0.71	**65.56**	3.28	6.05	3.02
pop	5.8	3.76	9.24	11.43	4.02	8.56	2.40	4.10	**35.78**	8.38	12.32
rhythm & blues	3.4	12.59	11.40	8.37	1.96	3.73	4.45	5.29	11.82	**31.94**	8.46
rock	43.9	4.19	2.65	3.61	2.95	5.15	0.57	1.59	7.54	1.99	**69.76**
african	0.3	12.43	29.94	5.37	5.08	14.12	5.65	7.34	7.34	8.19	4.52
asian	1.2	1.63	6.64	3.47	4.52	3.94	2.67	1.98	52.76	6.12	16.26
avant-garde	0.2	8.38	5.67	2.47	9.25	14.43	4.19	16.40	4.07	4.93	30.21
classical	2.2	5.78	1.64	4.76	1.73	30.75	0.16	41.08	5.38	2.19	6.53
easy listening	0.4	6.81	7.07	17.38	2.97	21.05	0.79	14.76	15.20	5.94	8.03
ska	0.7	3.00	42.24	4.00	9.71	0.75	9.66	1.15	5.36	3.55	20.57

(b) TAG

Table 4: Confusion matrices for S2-ONLY-D1 strategy. Original class names for the classifiers are listed in parentheses. Misclassifications >10% are shaded light gray and >20% dark gray.

take into account for validating the robustness of classifier models. We propose a methodology and software tools for such an evaluation. The tools let researchers conduct large-scale evaluations of classifier models trained within AcousticBrainz, given an independent source of ground-truth annotations and a mapping between the classes.

We applied our methodology and evaluated the performance of five genre classifier models trained on MIR genre collections. We applied these models on the AcousticBrainz dataset using between 240,000 and 1,740,000 music recordings in our validation sets and automatically annotated these recordings by genre using Last.fm tags. We demonstrated that good cross-validation results obtained on datasets frequently reported in existing research may not generalize well. Using any of the better-performing models on AcousticBrainz, we can only expect a 43–58% accuracy according to our Last.fm ground truth when presented with any recording on AcousticBrainz. We feel that this is still not a good result, and highlights how blurred the concept of genres can be, and that these classifiers may be "blind" with respect to some important musical aspects defining some of the genres. More research effort is required in designing musically meaningful descriptors and making them error-resistant, as well as understanding the relationships between different genre taxonomies.

Importantly, the application of the proposed methodology is not limited to genres and can be extended to other classification tasks. In addition to the proposed methodology and tools, we release a public dataset of genre annotations used in this study. [19] In our future work we plan to investigate and publish more independent sources of ground-truth annotations, including annotations by genre and mood, that will allow researchers to conduct more thorough evaluations of their models within AcousticBrainz.

5. ACKNOWLEDGEMENTS

This research has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 688382. We also thank Gabriel Viglensoni for help with mining Last.fm tags.

[19] https://labs.acousticbrainz.org/lastfm-genre-annotations

6. REFERENCES

[1] J. Bekios-Calfa, J. M Buenaposada, and L. Baumela. Revisiting linear discriminant techniques in gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):858–864, 2011.

[2] D. Bogdanov, J. Serrà, N. Wack, P. Herrera, and X. Serra. Unifying low-level and high-level music similarity measures. *IEEE Transactions on Multimedia*, 13(4):687–701, 2011.

[3] V. Chudáček, G Georgoulas, L. Lhotská, C Stylios, M. Petrík, and M Čepek. Examining cross-database global training to evaluate five different methods for ventricular beat classification. *Physiological measurement*, 30(7):661–677, 2009.

[4] N. Erdogmus, M. Vanoni, and S. Marcel. Within-and cross-database evaluations for gender classification via befit protocols. In *International Workshop on Multimedia Signal Processing (MMSP'14)*, pages 1–6, 2014.

[5] C. Fernández, I. Huerta, and A. Prati. A comparative evaluation of regression learning algorithms for facial age estimation. In *Face and Facial Expression Recognition from Real World Videos*, pages 133–144. Springer, 2015.

[6] A. Flexer and D. Schnitzer. Effects of album and artist filters in audio similarity computed for very large music databases. *Computer Music Journal*, 34(3):20–28, 2010.

[7] E. Guaus. *Audio content processing for automatic music genre classification: descriptors, databases, and classifiers*. PhD thesis, Universitat Pompeu Fabra, 2009.

[8] P. Herrera, A. Dehamel, and F. Gouyon. Automatic labeling of unpitched percussion sounds. In *AES 114th Convention*, 2003.

[9] H. Homburg, I. Mierswa, B. Möller, K. Morik, and M. Wurst. A benchmark dataset for audio classification and clustering. In *International Conference on Music Information Retrieval (ISMIR'05)*, pages 528–531, 2005.

[10] C. W. Hsu, C. C. Chang, and C. J. Lin. A practical guide to support vector classification. 2003.

[11] D. Liang, H. Gu, and B. O'Connor. Music genre classification with the Million Song Dataset. Technical report, Carnegie Mellon University, 2011.

[12] A. Livshin and X. Rodet. The importance of cross database evaluation in musical instrument sound classification: A critical approach. In *International Conference on Music Information Retrieval (ISMIR'03)*, 2003.

[13] M. Llamedo, A. Khawaja, and J. P. Martínez. Cross-database evaluation of a multilead heartbeat classifier. *IEEE Transactions on Information Technology in Biomedicine*, 16(4):658–664, 2012.

[14] A. Y Ng. Preventing" overfitting" of cross-validation data. In *International Conference on Machine Learning (ICML'97)*, pages 245–253, 1997.

[15] E. Pampalk, A. Flexer, and G. Widmer. Improvements of audio-based music similarity and genre classificaton. In *International Conference on Music Information Retrieval (ISMIR'05)*, pages 628–633, 2005.

[16] A. Porter, D. Bogdanov, R. Kaye, R. Tsukanov, and X. Serra. AcousticBrainz: a community platform for gathering music information obtained from audio. In *International Society for Music Information Retrieval Conference (ISMIR'15)*, 2015.

[17] A. Schindler, R. Mayer, and A. Rauber. Facilitating comprehensive benchmarking experiments on the million song dataset. In *International Society for Music Information Retrieval Conference (ISMIR'12)*, pages 469–474, 2012.

[18] H. Schreiber. Improving genre annotations for the million song dataset. In *International Society for Music Information Retrieval Conference (ISMIR'15)*, 2015.

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

[20] B.L. Sturm. An analysis of the GTZAN music genre dataset. In *International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies (MIRUM'12)*, pages 7–12, 2012.

[21] B.L. Sturm. Classification accuracy is not enough. *Journal of Intelligent Information Systems*, 41(3):371–406, 2013.

[22] B.L. Sturm. A survey of evaluation in music genre recognition. In *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation*, pages 29–66. Springer, 2014.

[23] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.