Support Vector Musicality: The Methods of Genre Definition

INTRODUCTION

Although no artistic movement can be fully characterized by a single categorical label, it is reasonable to claim that every musical genre is anchored by some motifs, beliefs, techniques, and styles which qualify its constituent songs and artists. The primary interest of this project is to generate 'definitions' of music genres by recognizing aggregated patterns and trends in their low- and mid-level audio features; we rest our faith in the premise that a song's timbre, pitch, rhythm, and harmony contribute substantially towards its genre characterizations. Thus the most valuable part of the project is not necessarily the models' training and testing errors, but rather the 'meaning' of the model's mathematical formulation and what the fitted parameters can tell us about the data.

Music Information Retrieval is a very actively field with prolific publications in applications of novel features and improvements in classification performance. This project does not attempt the same objectives, but rather seeks to interpret the results produced by the Black Box and question the definition of 'genre' itself and how music is perceived and understood.

The Million Song Dataset (MSD)

This project is largely motivated by the existence of the Million Song Dataset. With support from the National Science Foundation, the collaboration between LabROSA of Columbia University's Electrical Engineering department and The Echo Nest published a paper, along with the dataset, in the Proceedings of the 12th International Society for Music Information Retrieval (ISMIR) Conference in 2011 *[1]*. Prior to this, the Music Information Retrieval community had been actively conducting experiments on a few datasets ranging from 698 to 3,227 songs with genre labels*[3]*. Now the public availability of metadata for 1,000,000 songs introduced vast new opportunities for collaboration and exploration in many branches of MIR research.

A surprising yet defining quality of this dataset is the absence of any actual audio soundtracks, due to copyright laws as well as technical convenience. Rather, The Echo Nest had performed analyses on the audio waveforms and derived mid-level acoustic features, most notably pitch and timbre, then annotated with song metadata such as title, artist information, album name, and other algorithmically computed metrics. Relevant for this project, the MSD Allmusic Genre Dataset (MAGD) provides unoptimized expert annotated genre label ground truth for 422,714 songs.

Literature Review

At the following ISMIR conference in 2012, Schindler et al. from Vienna University of Technology published a paper presenting experimental genre classification results of various features and models on the MSD with the MASD labels as ground truth *[2]*. The top performing combination yielded 27.41% accuracy by applying a Support Vector Machine classifier on Statistical Spectrum Descriptor (SSD) features *(described in later section)*. On these features, the k-Nearest Neighbors clustering algorithm produced comparable accuracy of 27.07%. The other tested models––naïve Bayes, decision tree, and random forest––all yielded approximately 20% or less accuracy. Besides SSD, other features were also tested but all yielded approximately 15% or less accuracy, with one notable exception. The Mel-Frequency Cepstral Coefficients (MFCC) feature *(also discussed in later section)*, performed overall better than the other features, with 24.13% by k-NN clustering.

In the same year, these same researchers published another paper at the 10th International Workshop on Adaptive Multimedia Retrieval *[3]*. Still addressing genre classification on the MSD, this paper demonstrates classification improvements with the incorporation of the temporal domain in its audio features. One other significant difference compared to the previous paper is the choice of genre label ground truths. This paper instead evaluated classification performance on the four de facto test collections before the advent of the MSD: GTZAN, ISMIR Genre, ISMIR Rhythm,

and Latin Music Database. It's worth noting that the size of each of these datasets is less than 1% that of the MASD; additionally, the types of genres comprised by each dataset diverge significantly. The paper details the formulation of its top performing feature set: a combination of MFCC, Chroma features, loudness, and statistical moments calculated for all segments. Averaging across the four datasets, these features yielded improved accuracies of 76.1% with SVM and 68.1% with k-NN.

Objective

Genre classification is one of the most researched topics in MIR; there already exists a substantial body of work benchmarking many features and models and producing impressive classification metrics, and surely there are *many* possibilities for approaches to improve performance, especially with the granularity of features and the sheer quantity provided by the MSD. However, this project chooses to *not* pursue a new formulation of features or exploration of unexpected models for classification performance improvement for two main reasons.

The first is a practical one. Due to the limited time frame (7 weeks from proposal to poster presentation) and more importantly the lack of access to hardware and software infrastructures for parallelizing large-scale data processing, the volume of experiments necessary for such a task seem infeasible, and the probability of yielding statistically meaningful performance improvement, if any, is basically zero. Secondly, an ideological disinclination towards the 'Black Box paradigm' seems perhaps a more defensible motivation to instead search for some unintelligible meaning beyond performance metrics. If I ask the Machine a question, and the Machine replies with a *reasonable* answer, I'm rather more interested in discovering how the Machine arrived at that solution than interrogating for a better answer. Additionally, it surely appears counterproductive to strive for better performance via brute force without appropriate understanding of the underlying mechanisms.

APPROACH

First, we construct a rudimentary approximation of a known solution. As long as the model meets some baseline accuracy, it's reasonable to proceed on the assumption that the model's fitted parameters are semantically meaningful.

Features

We choose Statistical Spectrum Descriptors (SSD), timbre, and Chroma features because existing literature suggests they perform better than almost all other features, with the exception of calculating the SSD for each segment. We forego the incorporation of the temporal domain because it *substantially* inflates the data dimensionality, especially taking into account that the SSD already has 168 columns multiplied by at least thousands of training or testing data.

Statistical Spectrum Descriptor (SSD)

Music can be fundamentally represented as soundwaves, which are characterized by amplitude and frequency. Given amplitudes at frequencies ranging from 0 to 10 Hz, we discretize them into 60 bins. Then, according to theories of psychoacoustics, these 60 bins of fundamental frequencies are grouped into 24 'critical bands' which nonlinearly corresponds to pitch perceived by the human ear. The Statistical Spectrum Descriptor refers to the statistical moments (mean, median, variance, skewness, min- and max-value) of all amplitudes within *each critical band*. Thus, each SSD has dimension $24 \times 7 = 168$ ($n_{critical\_bands} \times n_{moments}$).

Timbre

The timbre of a note refers to its tone color or tone quality. More specifically, timbre describes the *textures* of two notes with the *same pitch and loudness* but produced by different sources. For example, the same note played on the piano sounds different than if it were played on a guitar or sung by a human.

Chroma

Chroma corresponds to the 12 pitch classes: {C, C#, D, D#, E, F, F#, G, G#, A, A#, B} and does not distinguish between notes separated by octaves. Chroma describes the harmonic and melodic characteristics well while being robust against variations in timbre. Intuitively, a song and a cover version of the same song would have different timbres, but similar Chroma.

Model

We choose the Support Vector Machine with polynomial kernel and one-vs-rest approach to multi-class classification. In general, SVM yields robust performance in many fields of applications, and existing MIR literature confirm this is also true for genre classification. Its mathematical formulation conveniently allows for kernelization, making it particularly well suited for very high dimensional and non-linearly separable data.

RESULTS

Training & Testing Datasets

292 columns: 168 for Statistical Spectrum Descriptors, 124 for timbre and Chroma features.

Tested with 2,000 samples per class. Train sample distribution as follows:

Pop Rock	235,214	RnB	12,304	Blues	4,796
Electronic	38,540	International	12,181	Vocal	4,179
Rap	18,855	Country	9,686	Folk	3,784
Jazz	15,741	Religious	6,775	New Age	1,992
Latin	15,475	Reggae	4,898	Comedy Spoken	56

Classifier Performance Metrics

Train accuracy	**92.45%**	**Test accuracy**	**35.80%**

Training Errors

First, we examine the training errors along with its confusion matrix. As expected, the diagonal demonstrates that the majority of training samples are correctly classified. Though predicting on training data almost always suffers from overfitting, it lends some credibility to perhaps some sort of 'rationale' behind the misclassified instances. All instances shaded red indicate 1.5% to 2.5% error rate, while all instances shaded blue indicate that only zero or one samples were misclassified.

By far, 'International' is the most misclassified class; this makes sense because 'International' simply means 'Not American' and it's perfectly reasonable to confuse some samples with other more musically distinct genres. Next, two noteworthy patterns immediately command our attention. 'Rock' was heavily confused with 'Blues', 'Country', and 'Folk' while 'Country' was often confused with 'Folk', 'Rock', 'Religious', and 'Vocal.' Leveraging some abridged historical domain knowledge, one might recognize some correlation in that country music originated in the 1920s in southern US with roots in southeastern genres such as folk and blues, while 'Rock and Roll' first originated as a genre in the US in the 1950s and drew heavily on blues, RnB, and country influences as well as folk, jazz, and others. On the other hand, some genres were rarely confused as anything else. Namely, 'Spoken Comedy' prescribes predominantly vocal speech with rather limited instrumental sounds while 'Rap' is often specifically characterized extremely rhythmic speech not found in any other genre.

Confusion Matrix for Predictions on Training Data

	blues	comedy	country	electro	folk	intnl.	jazz	latin	new age	rock	rap	reggae	religious	r n b	vocal
blues	1869	5	4	3	14	23	20	18	9	5	3	6	5	4	12
comedy	15	1883	14	6	14	14	3	12	1	1	0	4	7	10	16
country	19	12	1839	2	25	19	10	17	8	2	1	3	18	8	17
electronic	10	2	7	1871	9	32	6	15	4	4	6	13	5	13	3
folk	30	7	45	3	1780	35	8	28	15	2	0	6	14	6	21
International	18	3	16	4	31	1821	11	24	13	3	11	16	5	6	18
jazz	22	3	10	5	12	31	1814	27	21	3	0	12	8	16	16
latin	25	2	23	4	28	29	10	1821	6	4	5	9	6	14	14
new age	13	1	14	12	22	50	19	22	1821	2	0	0	7	14	3
pop rock	32	3	33	16	35	45	14	34	8	1719	3	12	16	22	8
rap	4	7	6	8	4	24	4	14	1	2	1869	38	2	16	1
reggae	14	1	6	2	2	21	7	25	0	1	7	1902	2	8	2
religious	20	5	36	1	25	38	7	37	7	6	7	8	1767	24	12
r n b	29	1	20	6	23	26	10	26	9	4	6	20	8	1799	13
vocal	32	4	35	3	27	30	35	18	10	3	2	5	8	18	1770
misclassified	283	56	269	75	271	417	164	317	112	42	51	152	111	179	156

Testing Errors

	blues	comedy	country	electro	folk	intnl.	jazz	latin	new age	rock	rap	reggae	religious	r n b	vocal
Precision	4.92	0.26	9.51	48.18	3.62	4.32	21.13	9.34	3.88	94.40	39.81	12.08	3.96	13.04	7.58
Recall	34.11	71.43	29.65	32.44	25.40	9.67	31.22	20.36	37.55	10.35	49.68	61.13	15.73	24.22	46.66
F1 score	8.59	0.52	14.40	38.77	6.33	5.98	25.19	12.81	7.03	18.65	44.21	20.18	6.33	16.93	13.02
Support	4796	56	9686	38540	3484	12181	15741	15475	1992	235213	18855	4898	6775	12304	4179

While the overall testing accuracy is 35.80%, it may not be the most informative metric for multi-class classification, since it requires for each sample that each label set be correctly predicted. To supplement our evaluation of the classifier performance, we also consider its precision and recall. Precision is the ratio *tp* / (*tp* + *fp*) where *tp* is the number of true positives and *fp* the number of false positives; intuitively, precision is the ability of the classifier *not* to label a negative sample as positive. Recall is the ratio *tp* / (*tp* + *fn*) where *fn* is the number of false negatives; intuitively, recall is the ability of the classifier to find all positive samples. And the F1 score is the harmonic mean of precision and recall. Traditionally, all three of these metrics range from 0 to 1; in the table above, they've been scaled to 0 to 100 for readability.

For most classes, recall performs significantly better than precision, indicating that the classifier was much better at recognizing that genre and was not likely to label anything else as that genre. 'Jazz', 'Rap', 'RnB' are exceptions with recall only slightly higher than precision, so the classifier recognized characteristics of these genres in many other samples as well. 'Electronic' and 'Rock' stand out as two notable outliers, both with higher precision than recall. For Electronic, both metrics are 'moderate' and close in range while for Rock, precision reaches a stunning 94.4% with recall lagging much behind at 10.35%. It seems that while the classifier did not mistake anything that's not Rock as Rock, it did decide a rather large portion of Rock would be better described as something else.

FINDINGS & INTERPRETATIONS

The ability to recognize patterns in the training dataset confusion matrix is exciting because it confirms the motivating assumption of this project––that musical genres are substantially defined by acoustic features such as rhythm and timbre. However, it is also sobering, because one might start to question the utility of posing genre recognition as a multi-class classification problem. Supervised classification problems optimize mathematical models to 'learn' ground truths and evaluate the models based on the predictions' compliance to these presupposed truths. The very cursory analysis above points out some flaws in this setup. The model 'misclassifies' a song labeled 'RnB as 'Blues' and we designate that as an error when in fact the model has recognized the song's melodic or textural qualities which, when identified by the discerning ear of a music critic, might champion the song as 'nostalgic' or 'soulful' or 'nuanced.'

That machine learning might be capable of 'critiquing' the validity of its ground truth labels may indicate that it's capable of much more than empirical risk minimization and may demand more complex approaches to machine learning problems than maximizing primitively defined metrics. It's certainly possible that sometimes ground truth labels aren't always 'true' and there's perhaps more potential for machine learning models to 'detect errors' than is currently being exploited.

More specifically, meaningful interpretation of the dual coefficients in SVMs yet remains to be solved. If the orthogonality to decision boundaries can be deciphered to reveal weight ratios of features which contribute towards correct predictions, not only would classification performance improve, there could be a completely new approach to feature/column selection than techniques available today.

CONCLUSION

Though this project decidedly takes a stance on the approach to genre classification, crafting more robust feature spaces and experimenting with other models are certainly not mutually exclusive to gaining insight about semantic interpretations of the data. In fact, the success of the latter is contingent upon continued investment in the former. For future progress, I hope to delve into the source code of *libsvm* (or the source code of any other SVM solver) to dissect the composition and manipulation of its *alpha* matrix in order to extract the underlying mathematical mechanisms responsible for the yet inexplicable proficiency of SVMs.

References

[1] D. P. W. E. Thierry Bertin-Mahieux and P. L. Brian Whitman, "The million song dataset," in Proceedings of the 12th International Conference on Music Information Retrieval, 2011.

[2] R. M. Alexander Schindler and Andreas Rauber, "Facilitating comprehensive benchmarking experiments on the million song dataset," in Proceedings of the 13th International Society for Music Information Retrieval Conference, 2012.

[3] A. R. Alexander Schindler, "Capturing the temporal domain in Echonest Features for improved classification effectiveness," in Proceedings of the 10th International Workshop on Adaptive Multimedia Retrieval, 2012.