Musical Genre Tag Classification With Curated and Crowdsourced Datasets

Omar Diab, Anthony Mainero, Reid Watson
Stanford University, Computer Science
{odiab, amainero, rawatson}@stanford.edu

Abstract

Analyzing music audio files based on genres and other qualitative tags is an active field of research in machine learning. When paired with particular classification algorithms, most notably support vector machines (SVMs) and k-nearest-neighbor classifiers (KNNs), certain features, including Mel-Frequency Cepstral Coefficients (MFCCs), Chroma attributes and other spectral properties, have been shown to be effective features for classifying music by genre. In this paper we apply these methods and features across two datasets (GTZAN and the Million Song Dataset) with four different tag sources (GTZAN, The Echo Nest, MusicBrainz, and Last.fm).

Two of these tag sources are professionally curated (GTZAN and MusicBrainz) while the other two are crowdsourced—that is, unmonitored users create the tags for each track. Two of the datasets had features on a track-by-track basis (GTZAN and Last.fm) while the other two are classified by artist. By exploring the cross-validation balanced accuracy across these different datasets, we find that classifications are made significantly more accurately on curated tags in comparison to crowdsourced tags, but that tagging by artist as opposed to by song creates a considerably smaller difference in effect. We found, however, that crowdsourced tags can be effective when done in large enough quantites, as seen in the Last.fm dataset.

1. Introduction

Musical genre classification is an active field of machine learning research. Some argue that machine learning for genre classification is intractable for normal use because genres are not clearly defined. On the other hand, people using online music services are very likely to search for music by genre, so understanding how to automatically classify music by genre would be useful. At a minimum, overarching genres like rock or disco likely exhibit enough distinction for computers to effectively distinguish between them. Hence many scientists have attempted—and largely succeeded—in producing quality classifiers for determining genres.

In that effort, several papers have explored the efficacy of learning algorithms to predict genres. In his paper George Tzanetakis effectively classified genres on live radio broadcasts using a Gaussian classifier [1]. Mandel et. al. used SVMs on artist- and album-level features to make similar classifications as well [2]. Another study explored mixtures of Gaussians and k-nearest-neighbors for the same task [3]. Each of these studies used similar features—Mel-Frequency Cepstral Coefficients and chroma properties of the audio waveform, for instance—to make these classifications.

With prior work in mind we decided to focus on KNNs, SVMs, and other classifiers and explore their relative performance. In particular, we

are interested in determining how the nature of each dataset affects each classifier's accuracy.

2. Datasets

	Curated	Crowdsourced
Tag by Artist	*MusicBrainz*	*The Echo Nest*
Tag by Song	*GTZAN*	*Last.fm*

Table 1. Properties of the datasets used

We gathered songs from two sources: the Million Song Subset and GTZAN. The Million Song Subset is a selection of the metadata pertaining to 10,000 songs. In particular, it contains tags aggregated from The Echo Nest, MusicBrainz, and Last.fm [4]. Thus, we have four tag databases in total: MusicBrainz, The Echo Nest, Last.fm, and GTZAN.

Two of these databases, MusicBrainz and GTZAN, are curated; that is, humans assign their tags selectively with accuracy in mind for academic purpose. The other two databases, EchoNest and Last.fm, are crowdsourced: users apply tags with no moderator oversight. Thus, curated tag datasets are expected to be more accurate overall than crowdsources ones.

Additionally, two of these sources, The Echo Nest and MusicBrainz, assign tags by artist; Last.fm and GTZAN tags, on the other hand, apply to individual tracks.

GTZAN is a database of music created by George Tzanetakis specifically for machine learning analysis of genre classification problems. The selected music is classified into ten genres: blues, classical, country, disco, hip hop, jazz, metal, pop, reggae, and rock. Because the Million Song Subset did not contain enough classical and disco songs, those genres were ignored in our analysis.

We created an intersection of 1,359 songs that were present and tagged in each of our tag datasets. For each, we acquired 30-second previews of the audio tracks from the Million Song Subset (scraped from the 7Digital.com public API) and the GTZAN dataset. We used MARSYAS [5] to extract a number of relevant features from the raw audio files. The features we extracted were:

1. Mel-Frequency Cepstral Coefficients – short term spectral-based features which model amplitude across the spectrum.
2. Zero Crossings – the number of times the waveform crosses 0.
3. Spectral Centroids – where the spectral "center of mass" of a sound is.
4. Chroma Properties – discretizes the spectrum into chromatic keys, and represents the presence of each key.
5. Spectral Rolloff – the frequency at which high frequencies decline to 0, typically computed when the waveform hits 85% energy.

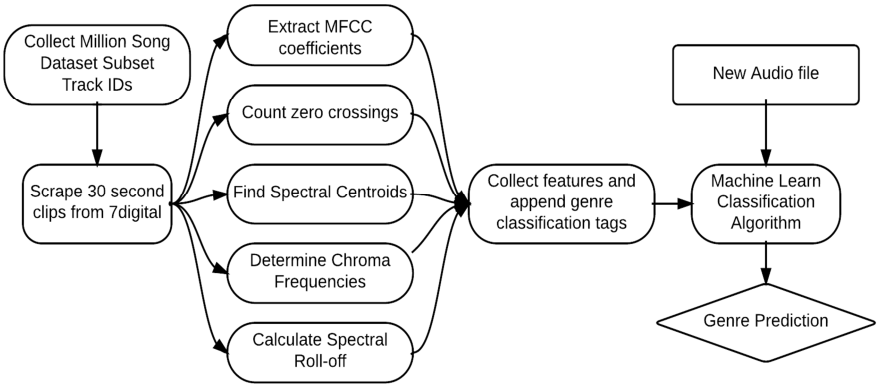


Figure 1. Proposed method of collecting data

When taken together, these features denote a wide range of sonic characteristics of the music, including instrumentation, tonal variation, timbre texture, and production attributes. Together, they constitute the "musical surface" of the song [1].

To sanity test our tag data, we made note of the relation between the frequency of songs of a certain genre appearing in online music datasets and the number records sold in that genre in 2011 [6]. Through this data, we found that blues, metal and reggae are overrepresented in online music

datasets, while country and pop are underrepresented. This information, while tangential, is a potential avenue for future research.

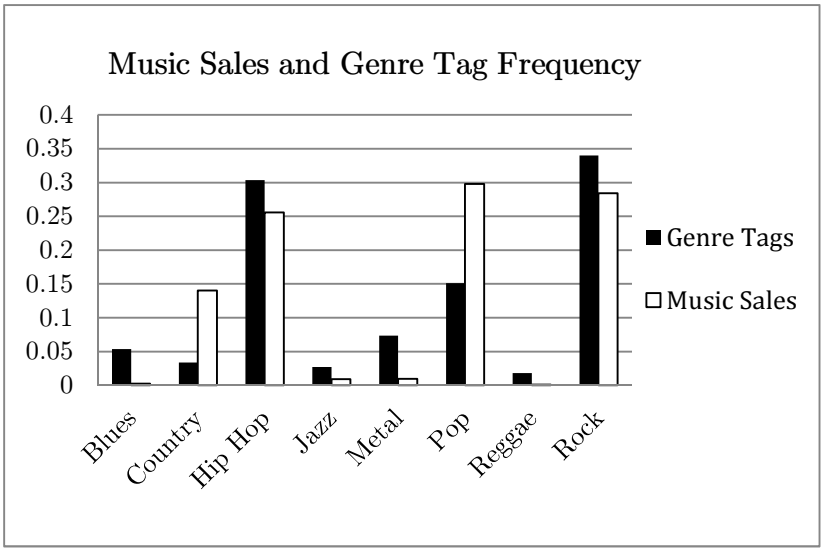


Figure 2. Music sales versus aggregated genre tag frequency in our datasets

3. Methodology

For each genre, we train our dataset on a classifier and use cross-validation with 5 folds to assess its overall accuracy. Logistic regression and Naïve Bayes require no specific model selection beforehand; with *k*-nearest-neighbors we used $k = \sqrt{n}$, where n is the number of training examples.

For SVMs, we use a modified version of LibSVM's `grid.py` module to produce an optimal choice of parameters for the SVM. In a SVM with a radial basis kernel, the objective function is to minimize:

$$J = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$$

and the kernel is:

$$K(x_i, x_j) = \exp\left(-\gamma \left| x_i - x_j \right|^2\right)$$

By varying our choices of C and γ, we can change the way the SVM decides on a dividing hyperplane. C represents the cost of a misclassification, so higher values of C encourage fewer misclassifications; and γ represents the relative importance of a single data point.

`grid.py` trains several SVMs with multiple choices of C and γ and compares them. By default, the comparator is cross-validation accuracy; however, our data sets are unbalanced—for instance, there are far more non-blues examples than blues examples, allowing the blues classifier to achieve a high cross-validation accuracy by just classifying everything negatively. Therefore, we selected balanced accuracy (BAC) as our metric, which is defined as follows:

$$\text{BAC} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$$

where

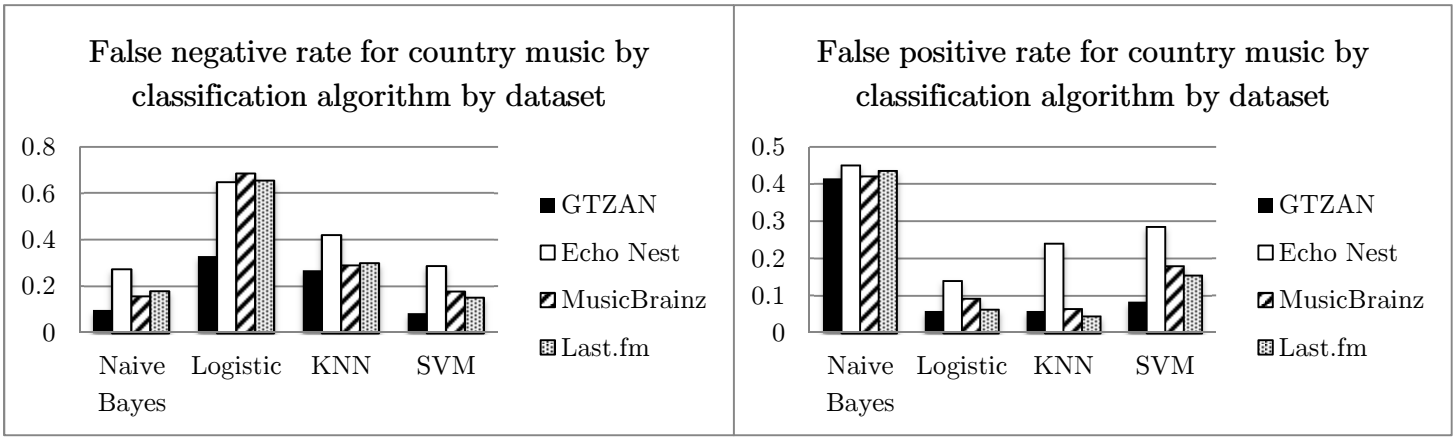
$$\text{Sensistivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{TP} = \text{true positives},$$
$$\text{FN} = \text{false negatives, etc.}$$

By using this metric to measure the efficacy of the SVM, we mitigate the aforementioned issues.

Additionally, the classifier was run with weights on positive and negative datapoints. The weights denote how much a training example of each class affects the objective function: for instance, if there is a weight of z on positive test examples, then misclassifications of positive test examples are z times as costly. To further offset unbalanced data sets, positive test examples were weighted by the number of negative test examples, and vice versa.



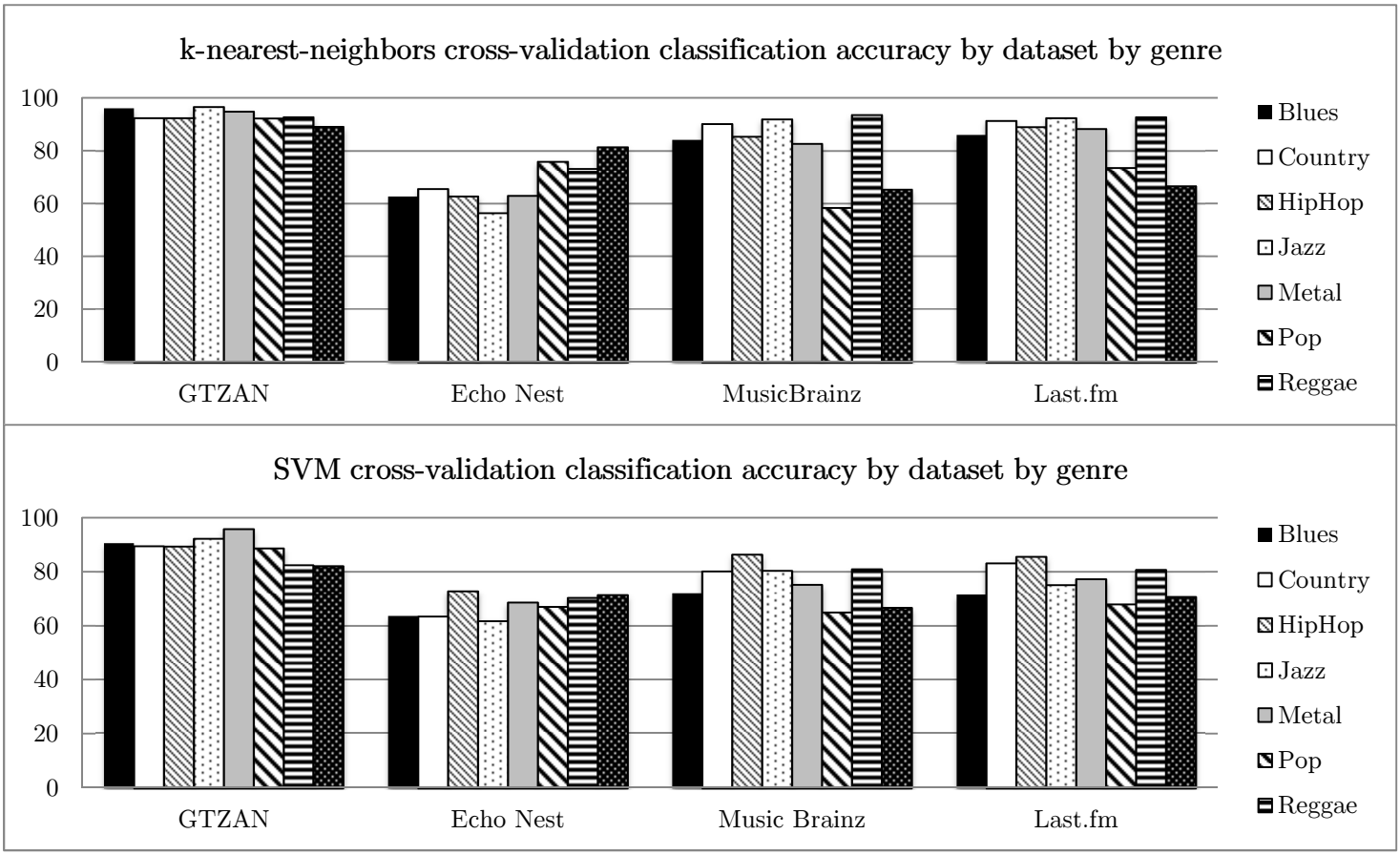
4. Results

a. Logistic Regression and Naïve Bayes

Logistic Regression and Naïve Bayes were relatively ineffective. Naïve Bayes exhibited large false positive rates and low false negative rates for across genres in all the datasets, while logistic regression produced high false negative and low false positive rates. Hence, both of them produced relatively low cross-validation accuracies and are not particularly useful classifiers for this task.

b. k-*nearest neighbors and SVMs*

Both KNNs and SVMs produced strong results. They exhibited both low false positive and false negative rates, and generally high cross-validation accuracy. Across all datasets and genres, KNNs and SVMs produced 83% and 77% cross-validation accuracies, respectively.



5. Discussion

We dismiss logistic regression and Naïve Bayes as being comparatively ineffective genre classifiers with a spectral feature set. Qualitatively, spectral features are likely not independent. On the other hand, SVMs and KNNs do not rest on independence assumptions, explaining why they performed relatively well. These results corroborate past studies in machine learning genre classifiers.

Comparing cross-validation accuracy across datasets, we find that each algorithm performs better on the GTZAN, MusicBrainz, and Last.fm datasets than on The Echo Nest dataset. Using KNNs and SVMs on the GTZAN dataset produces 92% and 88% average cross-validation accuracies across all genres, respectively. On the other hand, with the same algorithms on the Echo Nest dataset, we get 68% and 67%, respectively.

We attribute the improvement in classification accuracy between the Echo Nest and the other datasets to the fact that the latter are either curated or filtered. While our Last.fm data is crowdsourced, we narrow down our tags by frequency, so that only tags that are at least 1/3 as popular as the most popular tags for a song are chosen. Coupled with the fact that genre tags like rock and country are more common than other tags, we achieve decent accuracy with crowdsourced data so long as genre tags are applied to a song with high frequency. The Last.fm dataset actually performs about equally well as MusicBrainz, though this is likely attributed to our far larger set of tags on Last.fm songs than on MusicBrainz songs, giving us more training examples for Last.fm.

Additionally, we found that the effect of curated versus crowdsourced datasets is far more significant than the difference between artist-level and track-level tagging. This is likely due to the fact that while many artists exhibit wide variation in the types of music they make, they are less likely to stray between overarching genres like rock and disco. Hence, while studies suggest that music from specific artists or albums tend to have similar spectral qualities, creating an "album-effect", these effects are less significant for genre classification than the nature of the tagging source.

Overall, we find that a large number of examples and curated tags (GTZAN) give us the most machine-predictable dataset.

6. Citations

[1] Tzanetakis, George, Georg Essl, and Perry Cook. "Automatic Musical Genre Classification Of Audio Signals." *The International Society for Music Information Retrieval*. 2001. http://ismir2001.ismir.net/ pdf/tzanetakis.pdf

[2] Mandel, Michael I. and Daniel P.W. Ellis. "Song-level features and support vector machines for music classification". *The International Society for Music Information Retrieval*. Columbia University, 2005. http://www.ee.columbia.edu/~dpwe/pubs/ismir05-svm.pdf

[3] Li, Tao, Mitsunori Ogihara, and Qi Li. "A Comparative Study on Content-Based Music Genre Classification". *Special Interest Group on Information Retrieval*. 2003, p. 282. http://users.cis.fiu.edu/~taoli/pub/sigir03-p282-li.pdf

[4] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.

[5] Tzanetakis, George and Perry Cook. MARSYAS: a framework for audio analysis. Organised Sound 2000. 4(3) http://www.marsyas.info

[6] "The Nielsen Company & Billboard's 2011 Music Industry Report". *Business Wire*, 2011. http://www.businesswire.com/news/home/20120105005547/en/Nielsen-Company-Billboard%E2%80%99s-2011-Music-Industry-Report