

Predicting genre labels for artists using FreeDB*

James Bergstra, Alexandre Lacoste, and Douglas Eck

Dept. of Computer Science
Université de Montréal
CP 6128 succ Centre-Ville
Montreal, QC
H3C 3J7, Canada

bergstrj, lacostea, eckdoug@iro.umontreal.ca

Abstract

This paper explores the value of FreeDB as a source of genre and music similarity information. FreeDB is a public, dynamic, uncurated database for identifying and labeling CDs with album, song, artist and genre information. One quality of FreeDB is that there is high variance in, e.g., the genre labels assigned to a particular disc. We investigate here the ability to use these genre labels to predict a more constrained set of “canonical” genres as decided by the curated but private database AllMusic (i.e. multi-class learning). This work is relevant for study in music similarity: we present an automatic, data-driven method for embedding artists in a continuous space that corresponds to genre similarity judgments over a large population of music fans. At the same time, we observe that FreeDB is a valuable resource to researchers developing music classification algorithms; it serves as a reference for what music is popular over a large population, and provides relevant targets for supervised learning algorithms.

Keywords: Music Similarity, Music Classification, Genre Recognition, FreeDB

1. Introduction

The importance of musical genre in music information retrieval is reflected in recent research efforts to create objective and canonical genre hierarchies. [1] introduced a genre hierarchy that uses instrumentation and historical descriptors. Other researchers such as [2] have defined relatively complete taxonomies for the purposes of comparing the performance of different classifiers. Genre has also been the focus of MIR computing contests, the most recent being the MIREX 2005 [3] Symbolic Genre Classification and Audio Genre Classification contests.

Several commercial websites offer genre labels as part of their music descriptions. AllMusic¹ is perhaps the most popular, offering a wide range of information about bands, such as “Genre”, “Style”, “Similar Artists” and “Followers”. GraceNote (formerly CDDDB)², and FreeDB³ provide music meta-data such as album title, song titles, group and genre, indexed by a unique compact-disc identifier.

These approaches treat genre as both canonical (i.e. generated by well-established rules) and tied to the musical quality of a song. Unfortunately in reality this is not always the case. Traditionally, the music industry has used the concept of genre to orient consumers in record stores and internet sites. This leads to situations in which genre is decided not by the type of music found on a CD but rather by market pressures [4]. For example, a traditional “Blues” disc might be placed in the “Rhythm & Blues” if that is where it is likely to generate more sales.

More generally, [1] outline a number of problems associated with musical genre as commonly used in the music industry; [5] summarize:

1. They are designed for albums, not tracks.
2. There is considerable disagreement among different taxonomies on how to classify individual albums.
3. Within these taxonomies, taxons do not bear fixed semantics, leading to ambiguity and redundancy.
4. They are sometimes culture-specific, and often not related to actual musical content.

1.1. Two views of genre

This study compares two approaches to genre. The first is the traditional canonical view of genre, described above. As an example of this approach we use the AllMusic service which offers a clean and authoritative set of genre for every album and artist in its database. AllMusic is curated and relatively complete, and has shown itself to be very useful as evidenced by its immense popularity. However in addition to suffering from the general weaknesses outlined above

*Submitted to ISMIR 2006; Draft version; Not for citation.

¹ <http://allmusic.com>

² <http://gracenote.com>

³ <http://freedb.com>

concerning canonical genre systems, AllMusic is also prohibitively expensive to license for unlimited access, and is thus of limited value to the MIR research community.

The second approach allows for multiple competing opinions about the genre of an album. As an example of this approach we use the FreeDB system. In contrast to AllMusic, FreeDB is user-maintained and uncurated. Consequently, the genre labels in FreeDB have high variance and are drawn from a large set. FreeDB users are free to add their own genre tags if they wish. To compare, there are 32 popular and classical genres in AllMusic versus more than 500 in FreeDB⁴.

It is possible in FreeDB to have many competing genre tags for a single album. If we include all the albums for a given artist, we can build a histogram of genre labels that describes a *probability distribution* over a wide range of genre for that artist. We observe that this distribution could be useful for both predicting the canonical genre from sources like AllMusic and also, more generally, as a descriptor for music similarity. In this work we investigate the former point by performing a set of machine learning experiments with the FreeDB database. We test the ability of a simple model to predict the more canonical AllMusic genre information from the FreeDB histogram. The latter point, regarding music similarity, is left for future work.

2. Correlating AllMusic and FreeDB

To quantify the extent to which AllMusic genre can be predicted from FreeDB, we collected genre histograms for a small but representative sample of artists in FreeDB, and trained a supervised learning model to predict the AllMusic genre from the FreeDB genre.

2.1. Data Collection

The entire FreeDB database is available from the FreeDB website. Although the database is large (roughly 8GB in the unix distribution) we distilled it to a manageable size by grouping entries that correspond to the same artist (according to approximate string-matching of artist fields). After trimming artists with fewer than 10 disc entries, 20470 artists remained. After trimming artists with fewer than 50 disc entries, 2388 artists remained. We chose our dataset artists randomly from among this set of the most popular 2388. We selected 500 artists at random from the top 2388 as the subjects of our experiment. There are 639 genres in FreeDB that occur in at least 10 album labels. The most popular of these are Rock, Classical, and Pop, while the less-frequently used genre labels explore a wide space of possibilities (eg. weihnachtslieder, hungarian folk, viking metal). Among albums of the 500 artists that we chose, 408 genres were mentioned.

⁴ Because FreeDB is dynamic, uncurated and prone to typographical errors, it is difficult to offer a definitive count.

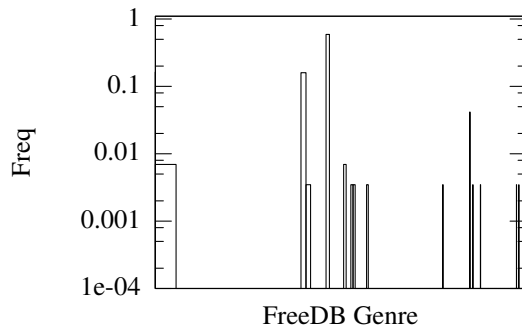


Figure 1. FreeDB genre histogram for 2Pac (logscaled y-axis). The histogram is dominated by Rap and HipHop, but many genres are present.

We collected data from AllMusic by hand, from their web interface at (<http://allmusic.com>). We recorded the genre associated with each artist, and when multiple genres were associated with an artist, we recorded all of them. Out of the 500 artists we chose from FreeDB, 30 of them were either repeated names, or else were not found in AllMusic, so our effective dataset had 463 examples. AllMusic defines 32 popular and classical genres. The most popular genre by far is Rock, which includes about half of the popular artists from FreeDB. In contrast to FreeDB, Classical is not a genre in AllMusic, but instead, classical music is divided among 13 genres such as Ballet, Choral Music, Keyboard Music, Opera, and Symphony.

2.2. Experiment

We used a machine learning algorithm to build a predictive model of AllMusic’s genre, given FreeDB’s genre information. We defined the input vector for a given artist to be the histogram of the genres assigned to their albums in FreeDB. Similarly, we defined the target vector for an artist to be the histogram of their genres in the AllMusic database. In many cases, especially among pop artists, the target histogram has only one genre. In contrast, FreeDB typically associates each artist with a large number of genres. For example, Figures 1 and 2 show the histograms associated with rapper 2Pac and classical composer Wolfgang Amadeus Mozart.

As a learning problem, we cast this as logistic regression. We assumed that the map between the different genre histograms would be relatively simple, so we chose a very simple neural network architecture to learn the map. Our neural network had no ‘hidden layer’, but was just a linear transform from the 563-dimensional input to the 32-dimensional output, followed by a softmax transform to ensure the output was a valid histogram. Thus we had $563 \times 32 = 17248$ parameters. We set them by gradient-descent, using the Kullback-Liebler divergence of the prediction with the target value as a cost function.

We evaluated the performance by 10-fold cross-validation, giving us 423 training examples for each fold. Even our

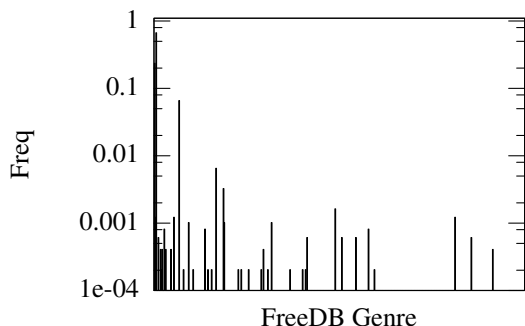


Figure 2. FreeDB genre histogram for Mozart, W.A. (logscaled y-axis). The distribution takes a peak at Classical, but many genres are present.

Table 1. Predictive power of FreeDB genre histograms.

Algorithm	Measure	Mean	$\pm 95\%$
random guess	KL	3.665	0.0219
rock guess	KL	2.269	0.0374
linsoft	KL	0.922	0.0539
random guess	0-1	3.00%	0.000742
rock guess	0-1	47.09%	0.002921
linsoft	0-1	74.10%	0.002003

simple network had far more capacity than was required for this task, so we regularized learning by starting the gradient descent with small weights, and using an early-stopping heuristic (learning is stopped once 50 iterations fails to improve on a held-out validation set).

2.3. Results and Discussion

The results of our experiment are posted in Table 1 and in Figure 3. In the table, three algorithms appear, and two scoring measures. The first algorithm *random guess* outputs a random histogram without making reference to the data in any way. The second algorithm, *rock guess*, guesses each output label according to the frequency in the training data, and represents the best classifier possible that does not use the input data. The third algorithm *linsoft* is the classifier described above, trained by gradient descent. The *KL* measure is the mean *KL*-divergence between test predictions and test targets. This is one metric for measuring the distance between two probability distributions. The 0 – 1 measure (classification accuracy) is the fraction of test examples such that the index of the maximum predicted value was equal to the index of an arbitrarily chosen maximum value in the target.

It was our impression that the artist names in AllMusic are correct, in keeping with our assumption that it is a well-managed database. However, a number of artists that appeared to be more popular outside of North America were missing from AllMusic (eg. Eri Esittajia). Furthermore, of the more exotic artists that did appear in AllMusic, many

entries were relatively incomplete; there was no background data on the artist, often nothing but a list of their albums and a genre (eg. Die Artze). In contrast, most popular artists (eg. Radiohead) have biographical sketches, a list of styles that are more precise than the genre, and a list of moods to which their music corresponds.

2.3.1. Classification Rates

While this task is not strictly classification, 92% (435/463) of the artists in our dataset have a single genre, so the 0 – 1 measure is relevant. The expected performance of the random guess algorithm is $\frac{1}{32} = 3\%$, in agreement with observation. The performance of the rock guess is more interesting, it correctly labeled 47% of the data. This is a consequence of the enormous class imbalance in AllMusic’s genre labels. The *linsoft* classifier correctly labeled 74% of test examples. This demonstrates a large degree of correlation between FreeDB and AllMusic. Still, there are a large number of mistakes, in view of the fact that the input and output are both genre histograms.

One shortcoming of the 0 – 1 measure is that it does not take into account how close the model is to getting the answer right. Another way to compare model performance is by looking at the relative ranking of correct labels. As is seen in Figure 3 the *linsoft* classifier converges towards 100% classification more quickly than the other two, indicating that its highly-ranked choices are often the right ones. For example, 89% of correct answers are found in the top 5 of 32 choices.

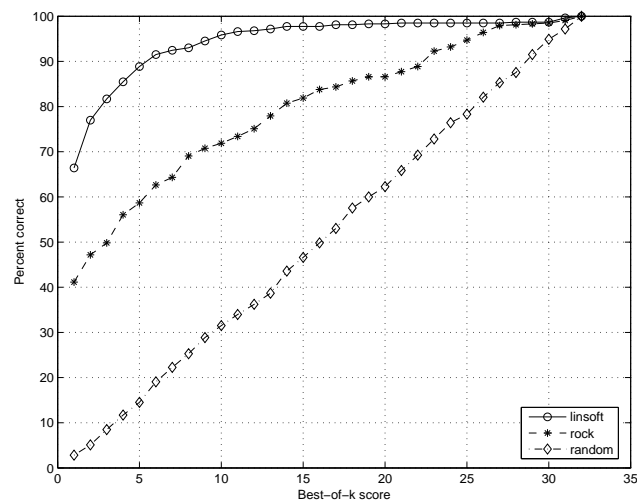


Figure 3. Cumulative sums showing the relative contributions in rank order. Faster convergence to 100% represents better performance in multi-task classification because it indicates that correct answers, even when not the first choice, are highly ranked by the classifier.

2.3.2. KL Divergence

The KL divergence between the target and predicted histograms provides another perspective on the performance of the algorithms. When there is just one correct genre for an artist, the KL divergence is the negative logarithm of the probability mass that the prediction put on the right genre. When there are multiple correct genres, the KL divergence is the average negative logarithm of the masses on the correct genres.

$$KL(P||Q) = \sum_i P_i \log \left(\frac{P_i}{Q_i} \right) \quad (1)$$

A KL score of z means that the model got an average fraction of e^{-z} of the density that it was supposed to. In this way we can see that the KL score of random guess corresponds to the fraction 0.025, the score of rock guess corresponds to 0.10, while the score of linsoft corresponds to 0.40. Recalling that there are 32 genres in the target histogram, this represents a significant degree of learning.

If the scores seem low, remember that in terms of a learning problem, this would normally be a very difficult problem. Our model has over 17000 parameters, and our dataset has only 463 examples. Only 422 of those are used during training. Of those 422, 338 are used for fitting the parameters, and 84 are held out to guard against overfitting. At the same time consider that each example is described by a histogram over 540 genres, and we would like to assign a histogram over 32 genres. Also consider that half of the target labels are rock, severely limiting the value of the little training data that we have.

3. Conclusions and Future Work

Our results suggest that a linear model explains much of the variation in AllMusic given FreeDB input. Given the success of our simple model in what would usually be a difficult machine learning problem, and given that AllMusic is a gold-standard for stylistic music classification, we conclude that FreeDB is a useful resource for labeling music for the purpose of music classification.

This conclusion has particular relevance for the MIR community. We would like to see FreeDB used as a reference for labeling music for the purpose of comparing algorithms. Gathering music that is representative of current tastes, and labeling that music in order to test automatic labeling methods are both difficult time-consuming tasks. FreeDB can help in both. First, FreeDB can be used as a reference of what is popular music (there are charts posted on the website, and each album entry is dated). Second, FreeDB can be used in the manner described here to associate a vector with an artist, such that the prediction of that vector is tantamount to classifying the genre of the artist.

In this work we have not addressed the question of what kind of music similarity is induced by our notions of histogram similarity, such as the [exponential of the negative]

KL divergence. Future work will consider whether nearness relations correspond to acoustic similarity or stylistic similarity, or whether they are more determined by factors outside the content of the music. At the same time, we have not demonstrated that histograms related to FreeDB genres are any easier to predict from symbolic or recorded audio than are the canonical genres offered by AllMusic. Indeed, since there are more FreeDB genres, they should presumably be harder to predict. The reason that FreeDB genres are attractive compared to canonical genres for machine learning methods, is that that machine learning methods work best when there is an enormous amount of training data. It is difficult to obtain large amounts of training data for canonical genres, while FreeDB can be used to automatically label any large collection of music. Future work will explore machine learning methods for predicting FreeDB genres from symbolic and recorded audio using training databases of hundreds of thousands, or millions of songs.

4. Acknowledgments

The authors would like to thank NSERC and FQRNT for their generous support.

References

- [1] F. Pachet and D. Cazaly, "A taxonomy of musical genres," in *Proc. Content-Based Multimedia Information Access (RIAO)*, 2000.
- [2] C. McKay and I. Fujinaga, "Automatic music classification and the importance of instrument identification," in *Proceedings of the Conference on Interdisciplinary Musicology (CIM05)*, Montreal, Canada, 2005.
- [3] J. S. Downie, K. West, A. Ehmann, and E. Vincent, "The 2005 music information retrieval evaluation exchange (mirex 2005): Preliminary overview," in *Proceedings of the Sixth International Conference on Music Information Retrieval: ISMIR 2005* (J. D. Reiss and G. A. Wiggins, eds.), pp. 320–323, Sept 2005.
- [4] D. Perrott and R. Gjerdingen, "Scanning the dial," in *Society for Music Perception and Cognition Conference (SMPC)*, Evanston, IL, 2005.
- [5] J. Aucouturier and F. Pachet, "Representing musical genre: A state of the art," *Journal of New Music Research*, vol. 32, no. 1, pp. 1–12, 2003.