METRIC LEARNING FOR MUSIC SIMILARITY

Daniel {Rosenberg, Erenrich}

ABSTRACT

Automatic music classification is an important problem in music analysis. In this paper, we use data from the "Million Song Dataset" to construct and evaluate music similarity metric and metric learning techniques. While others have done metric learning on music before, we evaluate which standard techniques perform best in both accuracy and time on this much larger dataset.

1. INTRODUCTION

Machine music analysis is a growing field in both industry and academia. One of the major difficulties with this work is that it requires large labeled corpuses which are difficult to obtain. The relatively new "Million Song Dataset" provides the music analysis community an opportunity to scale research to very large datasets with little organizational difficulty [2]. Already many papers of capitalized on this data to to solve such problems as playlist generation [4], music classification and song cover-identification [1].

Music similarity algorithms are especially useful because they can be used as subcomponents of algorithms like cover-identification and playlist generation. Previous work on music analysis was often performed on datasets as small as just 5000 songs [5]. It is now important that we reevaluate these techniques to determine how these algorithms scale in terms of accuracy and speed. It is not immediately obvious that all of our music metric-learning techniques will be able to scale. Others have already shown algorithmic techniques to speed up basic metrics of music similarity [3].

We performed metric learning across the "Million Song Dataset" in order to determine song similarity. As a ground truth for similarity we used both artist and genre as labels.

2. DATA REPRESENTATION

2.1 Audio representation

The "Million Song Dataset" provides features and metadata for a million songs as generated by the "EchoNest's" music analysis software. This data includes such information as volume, length and beats per minute. In order to fit the entire dataset in memory and effectively work with it we projected down to a much smaller set of features. This set of features was strongly influenced by [5] though we chose to drop several features as they are not consistantly populated across the dataset. Less than one percent of the dataset appears to be invalid or improperly populated since some values are clearly incorrect. For example. songs which are listed as having zero beats were dropped from the analysis.

We also added several features derived from the pitches contained in the songs. We chose to include the covariance matrix of pitches to get an idea of how pitches are grouped together. For instance, this should capture if a particular chord appears often. We also included the co-occurence counts so we can have a way to detect combinations that appear strongly correlated just because they occur infrequently. By adding quartiles of the max segment loudnesses, we hope to capture the song's progression over time. For instance, if a certain genre of songs tends to have a louder section towards the end, and another ends with a quiet repetition of the piece's theme, this feature wil capture that.

- Song length
- Mean segment length - Echo Nest divides songs into segments of length less than a second
- Segment length variance
- Mean segment loudness
- Segment loudness variance
- Third quartile of max segment loudness - This gives us an idea of the distrubtion of loudness across segments
- First quartile of max segment loundess
- Mean segment begin loudness
- Segment begin loudness variance
- Beat interval variance
- Tatum confidence
- Mean tatum length

- Tatums per beat
- Time signature
- Time signature confidence
- Song mode
- Song mode confidence
- Pitch covariance matrix
- Pitch cooccurence counts

By eliminating all features that vary with time along the song the dimensionality of the data is reduced 308 numbers. This means that the 500GB dataset becomes just 1GB in size and so is much easier to analyze. Exactly how much information has been lost in the transformation is unclear.

2.2 Data labels

We used two different types of labels in our experiments as ground-truth for music similarity. Songs are alternately considered "similar" if they share a common artist, if they share a certain number of music tags as provided by EchoNest or if the artists share similar genres depending.

3. ALGORITHMS

3.1 Whitening

It has been shown throughout the literature that first whitening the dataset in order to ensure that the covariance matrix of the data is the identity matrix improves performance. We use KNN using euclidean distance and cosine similarity in conjunction with and without whitening. Evaluation of the distance metrics will done by determining how well a KNN classifier is able to find songs that we have prelabeled labeled as similar.

Here we show our results on a compressed version of the dataset which only includes the top 500 most frequent artists and 25000 songs. A holdout test set of 5000 songs was used. This was done to decrease artist sparsity.

KNN	Euclidean	Whitened	Cosine	Whitened
1	12.3%	19.3%	14.8%	20.3%
3	12.1%	19.0%	14.5%	20.1%
5	11.6%	18.2%	13.5%	19.4%

3.2 Mahalanobis distance methods

The next step is to create a more sophisticated notion of distance than either euclidean or cosine similarity. Whitening is an unsupervised process and so is not taking advantage of all the information available to us.

We will learn a positive-semidefinite matrix A such that the distance between songs is defined to be $||A(x-y)||_2$. To do this we will use neighborhood component analysis where we require that $A = M^T M$ such which ensures that A is positive semidefinite. We also examine MCML. These methods though are slow and so we will transition into online learning models such as POLA which is faster though often less accurate.

4. REFERENCES

[1] T. Bertin-Mahieux and D.P.W. Ellis. Large-scale cover song recognition using hashed chroma landmarks. In *Proceedings of IEEE WASPAA*, New Platz, NY, 2011. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).

[2] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[3] B. McFee and G. R. G. Lanckriet. Large-scale music similarity search with spatial trees. In *12th International Symposium for Music Information Retrieval (ISMIR2011))*, October 2011.

[4] B. McFee and G. R. G. Lanckriet. The natural language of playlists. In *12th International Symposium for Music Information Retrieval (ISMIR2011))*, October 2011.

[5] Malcolm Slaney, Kilian Weinberger, and William White. Learning a metric for music similarity.