MIREX 2011: AUDIO TAG CLASSIFICATION USING WEIGHTED-VOTE NEAREST NEIGHBOR CLASSIFICATION

Mohamed Sordo
Music Technology Group
Universitat Pompeu Fabra
mohamed.sordo@upf.edu

Òscar Celma
Gracenote
ocelma@gmail.com

Dmitry Bogdanov
Music Technology Group
Universitat Pompeu Fabra
dmitry.bogdanov@upf.edu

ABSTRACT

In this long abstract, we present an algorithm for automatically annotating music with tags that is fast, scalable and relatively easy to implement. It uses acoustic similarity for propagating tags among audio items. The algorithm makes use of a variety of acoustical features, ranging from spectral features, to rhythm, tonal and highlevel features (such as mood, genre, gender). These features are then transformed into a reduced d–dimensional space, and finally combined with tempo and semantic features. A k–Nearest Neighbor classifier — with a modified weighting function and two different distance measures — is performed in order to propose tags to new music items.

1. INTRODUCTION

We present an algorithm for automatically tagging music that is fast, scalable and relatively easy to implement. It is, in fact, an enhanced and modified version of the method presented in [4]. Figure 1 illustrates the general structure of our algorithm. The Ground Truth training dataset is transformed by extracting acoustic features and performing a feature selection. Classification is then achieved by applying the same set of transformations for each test song, and then using similarity distances to infer tags from neighbors in the training dataset.

2. FEATURE EXTRACTION AND SELECTION

2.1 Feature Extraction

Table 1 summarizes the list of features that are used by our autotagging algorithm, which were extracted with the `Essentia` library [5]. The audio features are captured on a short-time frame-by-frame basis, using sliding windows of 46ms, and a hop size of 23ms. For tonal features, we set these values to 92ms and 46 ms, respectively. For a detailed description, please refer to [1–3]. The features are then averaged over the whole audio excerpt. We take the means, variances and their corresponding deltas. These values are then used to represent each audio excerpt as an N–dimensional vector.

Low level	average loudness, barkbands, HFC, MFCC, dissonance, zero crossing rate, pitch, silence rate, spectral features (centroid, rollof, kurtosis,...)
Rhythm/Tempo	beats (position, loudness), bpm, onset
Tonal	chords, key, hpcp, tuning
High level	genres, moods, gender, speech/music,...

Table 1. Summary list of the audio features used.

2.2 Feature Selection

A set of additional steps are performed to further reduce the dimensionality representation of each audio exceprt. We remove features with constant and invalid values. We then normalize the feature vectors, in order to apply Principal Component Analysis (PCA). PCA projects the original feature vectors into a reduced d–dimensional space, while still keeping the variance of the original data. The reduced d–dimensional vectors will be then used alone, or combined with other set of features, for the task of audio automatic tagging (or simply autotagging).

3. AUTOTAGGING

For each new audio excerpt (see Figure 1), we apply the same feature extraction and feature selection as in the case of the Ground Truth dataset. The resulting audio vector is then queried into the dataset. With the help of a k–NN method, our algorithm proposes tags from the k most similar songs, by using a weighted vote. We tried two different distance measures for retrieving similar audios. The first

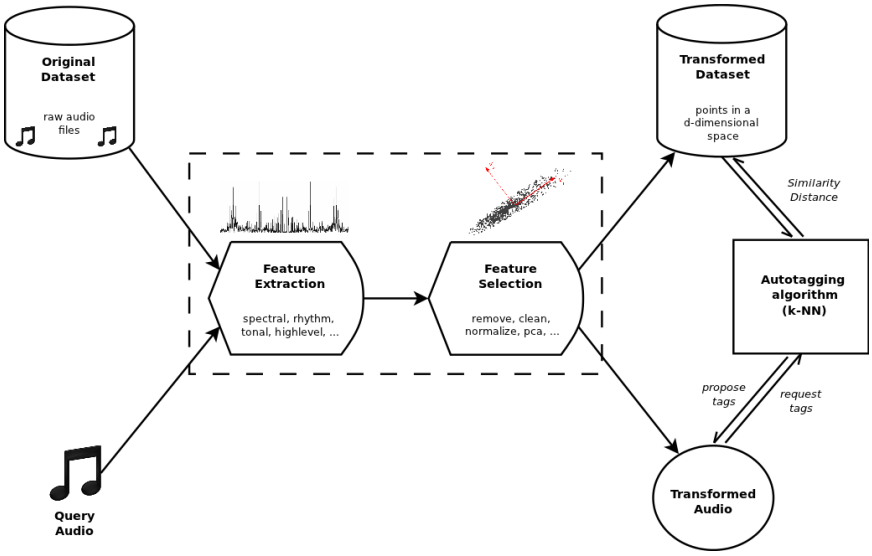

Figure 1. Framework of our autotagging approach

distance measure is an euclidean distance computed over the PCA reduced training data representation. The second measure is a hybrid distance that combines the output of the first distance with a Kullback-Leibler divergence based on single Gaussian MFCC modeling, a tempo-based distance, and a semantic classifier-based distance. The latter distance component employs probability estimations of different classes of genre, mood, and instrumentation made by Support Vector Machines. For more details on the hybrid measure, please refer to [6].

3.1 Audio tag classification

In the audio classification task (or binary relevance task), each one of the k nearest neighbors has equal vote. Additionally, a `voting threshold` (a value from 0 to 1) is defined to affect the number of proposed tags. For example, if we define a threshold of 0.4 for a 10–NN algorithm, only tags that appear at least $0.4 \times 10 = 4$ times in the top 10 neighbors are proposed. Empirical results have shown us that a threshold of 0.2 has a good trade off between precision and recall. For the MIREX task, we set $k = 18$, that is, for each song we take the 18 nearest neighbors [1] .

3.2 Affinity ranking

In the case of affinity ranking, the concept of voting threshold is removed, since we are ranking all the tags in the Ground Truth vocabulary. Instead, we take the R–nearest neighbors, where R is the size of the GT training dataset, and rank tags based on the following weighting function:

$$w_{ij} = \begin{cases} 1, & \text{if } j \leq k \\ \frac{1}{j^2}, & \text{otherwise} \end{cases} \quad (1)$$

where w_{ij} is the weight or score of tag i in rank j (j–nearest neighbor). The value of k is taken from the audio classification task. That is, the first k nearest neighbors will affect the classification equally, whilst the furthest neigbors $(R - k)$ are defined by a reciprocal quadratic function. This function is set to give a marginal weight for the furthest neighbors, so the nearest neighbors will have more influence on the highly ranked tags, while still allowing the ranking of all the tags in the Ground Truth vocabulary.

[1] This parameter was chosen from a previous experiment, using a different dataset.

4. REFERENCES

[1] P. Cano, M. Koppenberger, N. Wack, et al. "An Industrial-Strength Content-based Music Recommendation System" *Proceedings of 28th Annual International ACM SIGIR Conference*, 2005.

[2] E. Gómez "Tonal Description of Music Audio Signals" PhD Thesis. 2006.

[3] F. Gouyon "A computational approach to rhythm description — Audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing" PhD Thesis. 2005.

[4] M. Sordo, C. Laurier, and O. Celma "Annotating Music Collections: How content-based similarity helps to propagate labels" *Proceedings of the International Conference on Music Information Retrieval*, pp. 531–534, 2007.

[5] N. Wack "Essentia & Gaia: audio analysis and music matching C++ libraries developed by the Music Technology Group.," http://mtg.upf.edu/technologies/essentia.

[6] D. Bogdanov, J. Serrà, N. Wack, P. Herrera and X. Serra "Unifying Low-level and High-level Music Similarity Measures," *IEEE Transactions on Multimedia*, vol. 13, pp. 587–701, 2011.