Music Artist Style Identification by Semi-supervised Learning from both Lyrics and Content

Tao Li
Computer Science Dept.
University of Rochester
Rochester, NY 14627-0226
taoli@cs.rochester.edu

Mitsunori Ogihara
Computer Science Dept.
University of Rochester
Rochester, NY 14627-0226
ogihara@cs.rochester.edu

ABSTRACT

Efficient and intelligent music information retrieval is a very important topic of the 21st century. With the ultimate goal of building personal music information retrieval systems, this paper studies the problem of identifying "similar" artists using both lyrics and acoustic data. The approach for using a small set of labeled samples for the seed labeling to build classifiers that improve themselves using unlabeled data is presented. This approach is tested on a data set consisting of 43 artists and 56 albums using artist similarity provided by All Music Guide. Experimental results show that using such an approach the accuracy of artist similarity classifiers can be significantly improved and that artist similarity can be efficiently identified.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing,Information Search and Retrieval; I.2 [**Artificial Intelligence**]: Learning; I.5 [**Pattern Recognition**]: Applications; J.5 [**Arts and Humanities**]: music

General Terms

Algorithms, Performance, Experimentation

Keywords

semi-supervised learning, artist style, lyrics

1. INTRODUCTION

The rapid growth of the Internet has made it possible for users to have access to large amounts of on-line music data, including music sound signals, lyrics, biographies and etc. While the advancements in technologies for dealing with music data are presented in public forums, Internet record stores and on-line music programs use collaborative filtering to make community-based recommendations on music purchases. However, at this moment, these services are still in their infancy. The experiences with them are much less than satisfactory, mainly because the musical tastes vary significantly even in a community-based group. This leads us to believe that music recommendation systems should be tailored to fit the tastes and needs of individual listeners at the very moment. In [6] Huron points out that since the preeminent functions of music are social and psychological, the most useful characterization would be based on four types of information: *genre*, *emotion*, *similarity*, and the *style*.

This paper addresses the issue of identifying the artist style. Ellis et al. [5] point out that similarity between artists reflects personal tastes and suggest that different measures have to be combined together so as to achieve reasonable results in similar artist discovery. Our focus is given to singer-songwriters, i.e., those who sing their own compositions. We take the standpoint that the artistic style of a singer-songwriter is reflected in consistent use of both the acoustic sounds and the lyrics, and hypothesize that by combining acoustic features and linguistic features of songs, the artistic styles of an artist can be better captured than by using only one type of features [1]. Although we believe that the degree at which a listener finds a piece of music similar to another is influenced by the listener's cultural and music backgrounds and by the listener's state of mind, to make our investigation more plausible we choose to use similarity information available at All Music Guide (`www.allmusic.com`). In our experiments two artists are thought of as similar if this guide asserts one to be an artist similar to the other on the "Similar Artist" lists. We take the standpoint that the artist similarity information in this guide summarizes the opinions of one or more listeners.

Identification of artistic styles based on sounds and lyrics boils down to the problem of learning from heterogeneous data. Here we take a semi-supervised learning approach, in which a classification algorithm is trained for each feature set but the target label is adjusted for input data so as to minimized disagreement between the classifiers.

2. HETEROGENEOUS FEATURE SETS

2.1 Content-Based Features

There has been a considerable amount of work in extracting descriptive features from music signals for music genre classification and artist identification [11, 8]. In our study,

MM'04, October 10-16, 2004, New York, New York, USA.
Copyright 2004 ACM 1-58113-893-8/04/0010 ...$5.00.

[1] Surely there are some songwriters and performers quite deliberately explore different lyrical and musical styles. Those types of music are beyond the scope of this paper.

we use timbral features along with Daubechies wavelet coefficient histograms(DWCH). The feature set consists of the following three parts and totals 35 features.

2.1.1 Mel-Frequency Cepstral Coefficients (MFCC)

MFCC is designed to capture short-term spectral-based features. After taking the logarithm of the amplitude spectrum based on short-term Fourier transform for each frame, the frequency bins are grouped and smoothed according to Mel-frequency scaling, which is design to agree with perception. MFCC features are generated by decorrelating the Mel-spectral vectors using discrete cosine transform.

2.1.2 Other Timbral Features

Spectral Centroid is the centroid of the magnitude spectrum of short-term Fourier transform and is a measure of spectral brightness. *Spectral Rolloff* is the frequency below which 85% of the magnitude distribution is concentrated. It measures the spectral shape. *Spectral Flux* is the squared difference between the normalized magnitudes of successive spectral distributions. It measures the amount of local spectral change. *Zero Crossings* is the number of time domain zero crossings of the signal. It measures noisiness of the signal. *Low Energy* is the percentage of frames that have energy less than the average energy over the whole signal. It measures amplitude distribution of the signal.

2.1.3 DWCH

To extract DWCH features, the Db8 filter with seven levels of decomposition is applied to three seconds of sound signals. After the decomposition, the histogram of the wavelet coefficients is computed at each subband. Then the first three moments of a histogram plus the subband energy are used [8] to approximate the probability distribution at each subband.

2.2 Lyrics-Based Style Features

Recently, there has appeared some work that exploits the use of non-sound information for music information retrieval. Whitman and Smaragdis [14] study the use of the descriptions (obtained from All Music Guide) and the sounds of artists together to improve style classification performance. Whitman, Roy, and Vercoe [13] show that the meanings the artists associate with words can be learned from the sound signals. Researchers also present probabilistic approaches to model music and text jointly [3, 4]. From these results, it can be hypothesized that by analyzing how words are used to generate lyrics artists can be distinguished from others and similar artists can be identified.

Previous study on stylometric analysis has shown that statistical analysis on text properties could be used for text genre identification and authorship attribution [10, 1] and over one thousand stylometric features (style makers) have been proposed in variety research disciplines. To choose features for analyzing lyrics, one should be aware of some characteristics of popular song lyrics. For example, song lyrics are usually brief and are often built from a very small vocabulary. In song lyrics, words are uttered with melody, so the sound they make plays an important in determination of words. The stemming technique, though useful in reducing the number of words to be examined, may have a negative effect. In song lyrics, word orders are often different from those in conversational sentences and song lyrics are often presented without punctuation. To account for the characteristics of the lyrics, our text-based feature extraction consists of four components: bag-of-words features, Part-of-Speech statistics, lexical features and orthographic features. The features are summarized in Table 1. Their descriptions are as follows: *Bag-of-words:* We compute the TF-IDF measure for each words and select top 200 words as our features. We did not apply stemming operations. *Part-of-Speech statistics:* We also use the output of Brill's part-of-speech(POS) tagger [2] as the basis for feature extraction. POS statistics usually reflect the characteristics of writing. There are 36 POS features extracted for each document, one for each POS tag expressed as a percentage of the total number of words for the document. *Lexical Features:* By lexical features, we mean features of individual word-tokens in the text. The most basic lexical features are lists of 303 generic function words taken from [9][2], which generally serve as proxies for choice in syntactic (e.g., preposition phrase modifiers vs. adjectives or adverbs), semantic (e.g., usage of passive voice indicated by axillary verbs), and pragmatic (e.g., first-person pronouns indicating personalization of a text)planes. Function words have been shown to be effective style markers. *Orthographic features:* We also use orthographic features of lexical items, such as capitalization, word placement, word length distribution as our features. Word orders and lengths are very useful since the writing of lyrics usually follows certain melody.

Type	Number
Function Words (FW)	303
Token Place	5
Capitalization	10
Start of ...	9
Word Length	6
Line Length	6
Average Word Length	1
Average Sentence Length	1
POS features	36
Bag-Of-Words	200

Table 1: Summary of Feature Sets for Lyric Styles.

3. SEMI-SUPERVISED LEARNING

With both content-based and lyrics-based features, identification of artistic styles based on sounds and lyrics boils down to the problem of learning from heterogeneous data. Here we take a semi-supervised learning approach, in which a classification algorithm is trained for each feature set but the target label is adjusted for input data so as to minimize the disagreement between the classifiers.

3.1 Minimizing the Disagreement

Suppose we have an instance space $X = (X_1, X_2)$ where X_1 and X_2 are from different observations. Let D be the distribution over X. If f is the target function over D, then for any example $x = (x_1, x_2)$ we would have $f(x_1, x_2) = f_1(x_1) = f_2(x_2)$ where f_1 and f_2 are the target functions over X_1 and X_2 respectively [3]. It has been shown that

[2] Available on line at
http://www.cse.unsw.edu.au/∼min/ILLDATA/Function.word.htm

[3] For music artist style identification, $X1$ represents the lyrics-Based features and $X2$ represents the content-based features. The target functions correspond to the classifiers for style identification built on the features.

```
ALGORITHM Co-updating
Input: A collection of labeled and unlabeled data
       $\alpha$ — default 0.15, $T$ — default 30
Output: Two classifiers that predict class labels for new.
       instances based on different information sources.
1: Build $f_1^0$ using the first component of labeled samples.
2: Build classifier $f_2^0$ using the second component.
3: Loop for $T$ times:
3.1:    Step I: Using $f_1^{i-1}$ get the labels of all the unlabeled samples based on their first component; using
        $f_2^{i-1}$ on their second component.
3.2:    Step II: With probability $\alpha$, select the most confident unlabeled samples on which the two classifiers
        have the same predictions. Rebuild the classifiers $f_1^i$ and $f_2^i$ using the labeled samples and selected
        unlabeled samples.
4: Output $f_1^T$, $f_2^T$.
```

Figure 1: The algorithm description of the semi-supervised approach

minimizing the disagreement between two individual models could lead to the improvement of the classification accuracy of individual models. The detailed proof of Theorem 1 can be found in [7].

THEOREM 1. *Under certain assumptions, the disagreement upper bounds the misclassification error for the nontrivial classifier.*

3.2 Co-updating Approach

Based on Theorem 1, we have developed a co-updating approach to learn from both labeled and unlabeled data which aims to minimizing the disagreement on unlabeled data. The approach is an iterative Expectation-Maximization (EM)-type procedure. Its basic idea is as follows: The labeled samples are first used to get weak classifiers f_1^0 on X_1 and f_2^0 on X_2. Then for each iteration, the expectation step uses current classifiers to predict the labels of unlabeled data, the maximization step re-builds the classifiers using the labeled samples and a random collection of unlabeled samples on which the classifiers agree (i.e., they have the same predictions). This process is then repeated until some termination criterion is met. The detailed description of the algorithm is given in Figure 1.

The intuition behind the approach is that we stochastically select the unlabeled samples on which the two component classifiers agree and confident, and then use them along with the labeled samples to train/update the classifiers. The approach iteratively updates classifier models by using current models to infer (a probability distribution on) labels for unlabeled data and then adjusting the models to fit the (distribution on) filled-in labels.

4. EXPERIMENTS

4.1 Data Description

56 albums of a total of 43 artists are selected. The sound recordings and the lyrics from them are obtained. Similarity between artists is identified by examining All Music Guide artist pages. If the name of an artist appears on the "Similar Artist" of the web page of another artist, then X and Y are thought of as similar. Based on this relation artists having a large number of neighbors are selected. There are three of them, Fleetwood Mac, Yes, and Utopia. These three artists form a triangle, so the neighbors of these three are chosen as a cluster. Of the remaining nodes two clusters are identified in a similar manner. The clusters are listed in Table 2. The similarity graph of these nodes are shown in Figure 2. The goal for artist style identification is to distinguish each cluster from the rest. Our subjective evaluation does not completely agree with the artist clusters or the similarity itself. Nonetheless we use it as the ground truth.

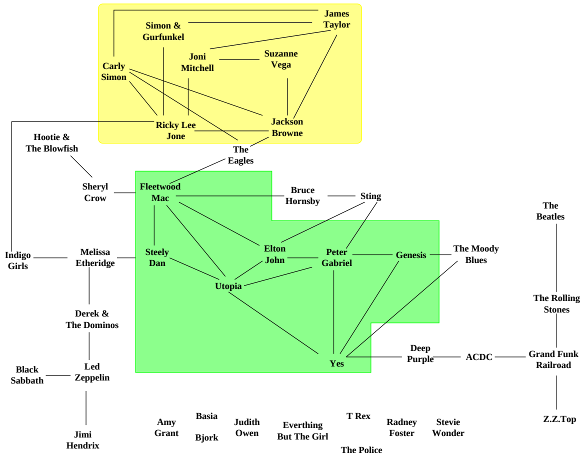

Figure 2: The artist similarity graph.

Clusters	Members
No. 1	{ *Fleetwood Mac, Yes, Utopia, Elton John, Genesis, Steely Dan, Peter Garbriel* }
No. 2	{ *Carly Simon, Joni Mitchell, James Taylor, Suzanne Vega, Ricky Lee Jones, Simon & Garfunkel* }
No. 3	{ *AC/DC, Black Sabbath, ZZ Top, Led Zeppelin, Grand Funk Railroad, Derek & The Dominos* }

Table 2: Cluster Membership.

4.2 Experiments

Generally building models when one class is rare can be quite difficult because there are often many unstated assumptions [15]. It is conventional wisdom that classifiers built using all the data tend to perform worse on the minority class than on the majority class since the class *priors* in the natural distribution are biased strongly in favor of the majority class and the minority class has much fewer training and test samples [12]. Although the balanced distribution will not always yield optimal distribution, it will generally lead to results which are no worse than, and often superior to, those which use the natural class distribution [12]. We sample roughly balanced datasets from the original dataset and the distributions of samples are shown in Table 3. We train a classifier that distinguishes each cluster from the rest of the artists. To build the classifiers we use support vector machines with linear kernels. The performance of the classifiers is measured using *accuracy*, *precision*, and *recall*.

The unlabeled data were used for testing and the results of the three experiments are shown in Table 4, Table 5, and Table 6, respectively. Without co-updating (labeled data only) we have three choices for the data source: only lyrics, only acoustics, and both. Co-updating approaches use both types of data. Accuracy measures can be applied to the

Experiments	TS	PS	NS	PSL	NSL	PSU	NSU
Cluster 1	217	103	114	26	29	77	85
Cluster 2	196	75	121	19	31	56	90
Cluster 3	200	80	120	20	30	60	90

Table 3: The distribution of the samples used in the experiments. The TS column is the number of total samples used, the PS column is the total number of positive samples, the NS column is the total number of negative samples, the PSL column is the number of positive labeled samples, the PSU column is the number of unlabeled positive samples, the NSL column is the number of negative labeled samples, and the NSU column is the number of unlabeled negative samples.

lyrics-based classifier in its final form (after co-updating), acoustic-based classifier in its final form, and the combination of the two [4]. So, the tables below have six row each.

We observe that the accuracy measures of a classifier built using labeled lyrics data are almost equal to those of a classifier built using labeled acoustic data and that combining the two sources improve the accuracy of classifier in the case of labeled data. The use of co-updating significantly improves accuracy for each of the three cases of data sources, but there is a slight gap between the two classifiers at the end.

We can conclude from these experiments that artist similarity can be efficiently learned using a small number of labeled samples by combining multiple data sources. We looked at the performance of the classifiers for Cluster 1 in more detail. The core of the cluster consists of Fleetwood Mac, Yes, and Utopia. We examined for which music tracks the combined classifier made an error after co-updating. Of the 71 errors it made, 38 were from albums of Peter Gabriel, Genesis, Elton John, and Steely Dan, none of which are not in the core of the cluster. Using analytical similarity measures to obtain the ground truth about artist similarity, thereby improving upon the data provided by web information resources, will be our future goal.

Classifier	Accuracy	Precision	Recall
Lyrics-based	0.506542	0.500000	0.384416
Content-Based	0.512150	0.509804	0.337662
Combined	0.530864	0.557143	0.467532
Co-updating/Lyrics	0.635802	0.572581	**0.622078**
Co-updating/Content	0.685285	0.654762	**0.714286**
Co-updating/Combined	**0.697531**	**0.694444**	0.649351

Table 4: The results on Cluster 1.

Classifier	Accuracy	Precision	Recall
Lyrics-based	0.506849	0.382353	0.464286
Content-Based	0.602740	0.426230	0.467742
Combined	0.630137	0.516667	0.553571
Co-updating/Lyrics	0.664384	0.563636	0.553571
Co-updating/Content	0.685285	0.583333	**0.625000**
Co-updating/Combined	**0.698630**	**0.61111**	0.589286

Table 5: The results on Cluster 2.

5. CONCLUSION

In this paper, we study the problem of music artist style identification from both lyrics and acoustic signals via a semi-supervised learning approach. Based on the theoretical foundations that the disagreement upper bounds the misclassification error, the semi-supervised learning approach then tries to minimize the disagreement on unlabeled data via an iterative Expectation-Maximization (EM)-type procedure. Experimental results on the data set consisting of 43 artists and 56 show the effectiveness and efficacy of our approach.

Classifier	Accuracy	Precision	Recall
Lyrics-based	0.643836	0.541081	0.616667
Content-Based	0.664384	0.537931	0.596667
Combined	0.686667	0.603175	0.633333
Co-updating/Lyrics	0.760000	0.700000	0.700000
Co-updating/Content	0.760000	0.662162	**0.816667**
Co-updating/Combined	**0.786667**	**0.741379**	0.786667

Table 6: The results on Cluster 3.

Acknowledgment

We would like to thank Sterling S. Stein for providing us the tools for extracting lexical and orthographic features.

6. REFERENCES

[1] Shlomo Argamon, Marin Saric, and Sterling S. Stein. Style mining of electronic messages for multiple authorship discrimination: first results. In *SIGKDD*, pages 475–480, 2003.

[2] Eric Bill. Some advances in transformation-based parts of speech tagging. In *AAAI*, pages 722–727, 1994.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[4] Eric Brochu and Nando de Freitas. Name that song!: A probabilistic approach to querying on music and text. In *NIPS*, 2002.

[5] D.P.W.Ellis, B. Whitman, A. Berenzweig, and S. Lawrence. The quest for ground truth in musical artist similarity. In *ISMIR*, pages 170–177, 2002.

[6] D. Huron. Perceptual and cognitive applications in music information retrieval. In *ISMIR*, 2000.

[7] Tao Li and Mitsunori Ogihara. Semi-supervised learning from different information sources. *Knowledge and Information Systems Journal*, 2004. In Press.

[8] Tao Li, Mitsunori Ogihara, and Qi Li. A comparative study on content-based music genre classification. In *SIGIR*, pages 282–289, 2003.

[9] R. Mitton. Spelling checkers, spelling correctors and the misspellings of poor spellers. *Information Processing and Management*, 23(5):103–209, 1987.

[10] Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–496, 2000.

[11] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), July 2002.

[12] G. Weiss and F. Provost. The effect of class distribution on classifier learning: An empirical study. Technical Report ML-TR 44, Rutgers University, 2001.

[13] B. Whitman, D. Roy, and B. Vercoe. Learning word meanings and descriptive parameter spaces from music. In *HLT-NAACL03 workshop*, 2003.

[14] B. Whitman and P. Smaragdis. Combining musical and cultural features for intelligent style detection. In *ISMIR*, pages 47–52, 2002.

[15] Bianca Zadrozny and Charles Elkan. Learning and making decisions when costs and probabilities are both unknown. In *SIGKDD*, pages 204–213, 2001.

[4] The combined classifier after co-updating is constructed by multiplying the probability outputs of the lyrics-based and content-based classifiers with the conditional Independence assumption.