SEMI-SUPERVISED MUSIC GENRE CLASSIFICATION

Yangqiu Song, Changshui Zhang and Shiming Xiang

State Key Laboratory of Intelligent Technology and Systems
Tsinghua University, Department of Automation
Beijing, China, 100084

ABSTRACT

Music genre classification is a hot topic in pattern recognition and signal processing. Classical supervised methods need lost of labeled music data to train a classifier. In this paper, we propose a semi-supervised genre classification algorithm which is developed on several labeled music tracks and lots of unlabelled tracks. Three features are extracted from the each music track and manifold regularization method is used to design the classifier. Experiments on a large number of test music data show that semi-supervised method can improve the classification accuracy.

Index Terms— Semi-supervised Learning, Music Genre classification

1. INTRODUCTION

Large amount of music collections can be found in the Internet due to the increasing growth of bandwidth and storage. Automatically classifying music into different classes is of great practical importance in information retrieval, and question answer systems. Currently, most classification work is done by hand because it is difficult to give precise definition of different kinds of music. Therefore, automatic classification is significant for it can reduce much of the labeling work.

Music genre is one of the top-level descriptions for human to organize the music collections [1, 2]. Music genre classification, which is defined as the most restrict form (i.e., the computer classify each music audio signals to one class), can be divided into two stages: feature extraction and classifier design [1]. While there have been various feature extraction methods [3, 4, 5, 6, 7, 8, 9], the classification methods have only been compared in [1] to the best of our knowledge. Supervised methods such as support vector machines (SVM) may produce good results, however, they require the scrupulous labeling work by experts, which is time consuming and costly. Unsupervised methods could work automatically, but they are difficult to determine the cluster number and show worse results than the supervised methods [1]. Therefore, the idea of combining supervised and unsupervised methods is brought forward, which is the semi-supervised method [10].

Semi-supervised classification methods, have been proven that they can both reduce the labeling work and improve the accuracy rate [11, 12, 13]. For classification problem, it is intuitive that every point should be similar to the points in its local neighborhood. Especially when dealing with high dimensional data, the data are more likely to distribute on a low dimensional manifold in some subspace. Therefore, regularizing the learned functions being smooth on the manifold or graph constructed by the data is often helpful.

In this paper, we propose a semi-supervised method applied to the content based music genre classification problem. We use the manifold regularization method based on least-square framework [11]. The algorithm penalizes the function both in reproducing kernel Hilbert space (ambient space RKHS) and the intrinsic geometry. Therefore, the learned function is sufficiently smooth both in RKHS and on manifolds or graphs. By adding the unlabeled data, the accuracy rate can be improved. Fig. 1 shows that, a weighted graph can be constructed via k-nearest neighbor method to exploit the local geometrical (manifold) structure among music tracks based on a music similarity measure. The tracks which have the same genre, have a large weight; the tracks which have different genres may also have a weight due to the small similarity measure.

This paper is organized as follows. In section 2, we introduce the semi-supervised algorithm. In section 3, we show the features and similarity measurements we used in this paper. Experiments are given in section 4. Finally we conclude in section 5.

2. REGULARIZED LEAST-SQUARES FRAMEWORK

We first consider a binary class case. Suppose we have l labeled data and u unlabeled data. Given the training data set $\{\mathbf{x}_1, ..., \mathbf{x}_l, \mathbf{x}_{1+1}, ..., \mathbf{x}_{l+u}\}$ and $\{y_1, ..., y_l\}$, where $\mathbf{x}_i \in X$ is the feature vector and $y_i \in Y$ is the label, we not only want to estimate the soft labels $\{f(\mathbf{x}_1), ..., f(\mathbf{x}_{l+u})\}$ of the training data points, but also want to know what is the soft label $f(\mathbf{x})$ of a new test point $\mathbf{x}$. To incorporate the additional information of unlabeled data, we use the following regularization

This work is funded by the Basic Research Foundation of Tsinghua National Laboratory for Information Science and Technology (TNList).

1-4244-0728-1/07/$20.00 ©2007 IEEE

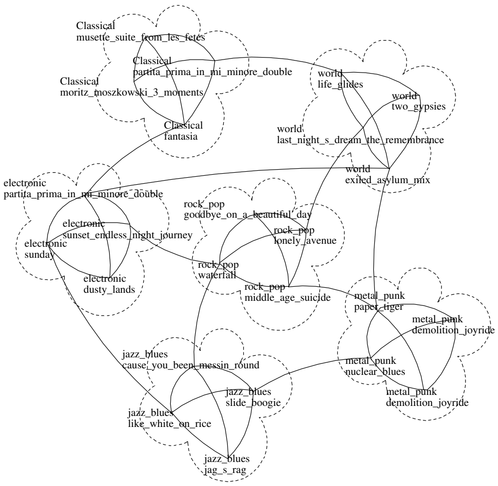


Fig. 1. Music Graph.

framework:

$$f^*(\mathbf{x}) = \arg\min_{f\in\mathcal{H}_K} \int_{X\times Y} V(y,f(\mathbf{x}))dP(\mathbf{x},y) + \lambda_1||f||_K^2 + \lambda_2||f||_I^2 \tag{1}$$

where $\int_{X\times Y} V(y,f(\mathbf{x}))dP(\mathbf{x},y)$ is *expected risk*, and the loss function can be arbitrary form $V(y,f(\mathbf{x}))$. The solution $f^*(\mathbf{x})$ is a function $X \rightarrow \mathcal{R}$ which lies in a bounded convex subset of RKHS $\mathcal{H}_K$ defined by a positive definite kernel function $K : X \times X \rightarrow \mathcal{R}$. The kernel function satisfies the Mercer condition. $||f||_K^2$ is the traditional Tikhonov regularization term in RHKS [14], and $||f||_I^2$ is the new regularization term based on graph or manifold [11, 12]. λ_1 and λ_2 are the parameters which control the tradeoffs of these two terms.

To simplify the representation of (1), we first use the *empirical risk* to replace the *expected risk*. Here, we use the least-squares loss function. Thus, the *empirical risk* is given by:

$$\frac{1}{l}\sum_{i=1}^{l}(y_i - f(\mathbf{x}_i))^2 \tag{2}$$

Second, according to Representer Theorems [15], under very general conditions on the loss function V, the optimal solution of f can be given as:

$$f(\mathbf{x}) = \sum_{i=1}^{l+u}\alpha_i K(\mathbf{x}_i, \mathbf{x}) \tag{3}$$

Substituting (2) and (3) into (1), the objective can be rewritten as a function of the $l+u$ dimensional vector $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}\in\mathcal{R}^{l+u}} \frac{1}{l}(\mathbf{y} - \mathbf{JK}\boldsymbol{\alpha})^T(\mathbf{y} - \mathbf{JK}\boldsymbol{\alpha}) + \lambda_1\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} + \lambda_2\boldsymbol{\alpha}^T\mathbf{KLK}\boldsymbol{\alpha} \tag{4}$$

where $\mathbf{y} = [y_1, ..., y_l, 0_{l+1}, ..., 0_{l+u}]^T$ is a $\mathcal{R}^{l+u}$ vector. $\mathbf{J}$ is a diagonal matrix, which is $\mathbf{J} = diag(1_1, ..., 1_l, 0_{l+1}, ..., 0_{l+u})$. $(\mathbf{K})_{(l+u)\times(l+u)}$ is the Gram matrix which satisfies that $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$. By using the Gaussian kernel, we have:

$$K(\mathbf{x}_i, \mathbf{x}_j) = exp\{-\frac{D(\mathbf{x}_i, \mathbf{x}_j)^2}{2\sigma^2}\} \tag{5}$$

where $D(\mathbf{x}_i, \mathbf{x}_j)$ is the distance measure of music similarity, which will be presented in section 3.

To use the graph or manifold regularization term, we define $G = (V, E)$ as an weighted neighborhood graph. V is the vertex set of graph, which can be defined on the training set, including both labeled and unlabeled data. E is the edge set which contains the pairs of neighboring vertices $(\mathbf{x}_i, \mathbf{x}_j)$. The neighboring vertices can be defined as such that either $D(\mathbf{x}_i, \mathbf{x}_j) < r$ or $\mathbf{x}_j$ $(\mathbf{x}_i)$ is among k nearest neighbors of $\mathbf{x}_i$ $(\mathbf{x}_j)$. Then the adjacency matrix $\mathbf{W}$ of graph is defined as:

$$\mathbf{W}_{ij} = \begin{cases} exp\{-\frac{D(\mathbf{x}_i, \mathbf{x}_j)^2}{2\sigma^2}\} & if\ (\mathbf{x}_i, \mathbf{x}_j) \in E \\ 0 & otherwise \end{cases} \tag{6}$$

and the normalized graph Laplacian [16] is:

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}} \tag{7}$$

where the diagonal matrix $\mathbf{D}$ satisfies $\mathbf{D}_{ii} = d_i$, and $d_i = \sum_{j=1}^{l+u}\mathbf{W}_{ij}$ is the degree of vertex $\mathbf{x}_i$. Here the adjacency matrix and the normalized graph Laplacian are both symmetric.

Note that (4) is convex, by differentiating (4) and let the derivative to be zero, we obtain the optimal solution [11]:

$$\boldsymbol{\alpha}^* = (\mathbf{JK} + \lambda_1 l\mathbf{I} + \lambda_2 l\mathbf{LK})^{-1}\mathbf{y} \tag{8}$$

where $\mathbf{I}$ is the identity matrix. Therefore, we can submit (8) to (3) to obtain the soft labels.

If $\lambda_2 = 0$, we have the following objective function:

$$f^*(\mathbf{x}) = \frac{1}{l}\sum_{i=1}^{l}(y_i - f(\mathbf{x}_i))^2 + \lambda_1||f||_K^2 \tag{9}$$

This is the supervised case which is the traditional regularized least-squares. The solution is then:

$$\boldsymbol{\alpha}^* = (\mathbf{K} + \lambda_1 l\mathbf{I})^{-1}\mathbf{y} \tag{10}$$

where $\mathbf{K}$ is an $l \times l$ matrix and $\mathbf{y}$ is $[y_1, ..., y_l]^T$. It has been shown that this algorithm is competitive to the popular method support vector machine (SVM) [14].

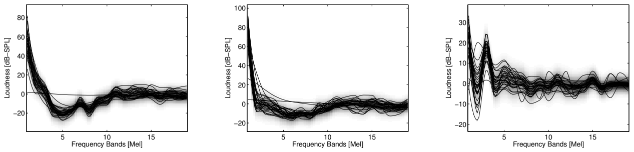
For multi-class classification, it is easy to use the one-against-one or the one-against-the-rest method to construct a classifier based on a set of binary classification algorithms. We also solve this as a one-against-the-rest problem. This is reasonable because it only requires to change the form of the labels of the data points, and does not need to modify the algorithm framework. Therefore, for each labeled point, we set label as $y_i = [-1, ..., y_i^{(j)} = 1, -1, ..., -1]$ if a point $\mathbf{x}_i$ is in the jth class.

3. MUSIC SIMILARITY MEASUREMENT

In this section, we present three music similarity measures which calculated by the music analysis toolbox [6].

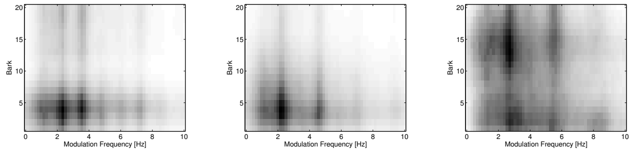
3.1. MFCC-EMD

MFCCs (Mel frequency cepstrm coefficients) are useful features to characterize music timbre. However, it is not easy to measure the similarity of MFCCs between two music tracks. One smart method is first calculating the histogram of MFCC of each track, and using k-means [7] or GMM [9] to approximate it, then calculating the Kullback Leibler (KL) distance of two histograms. The Earth Mover's Distance (EMD) which incorporates the KL distance can be used [7]. We compare three typical songs, whose names are "Sonata I Minuet" (classical), "Sonata II Minuet" (classical) and "Heavenly Rain" (rock). We use the first 20 MFCC coefficients (expect the 0th coefficient) and 30 k-means centers to gain the features. The features are shown in Fig. 2. We can see that the former two songs are more similar to each other.



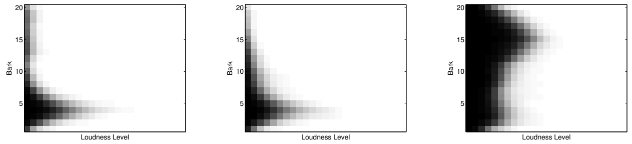
(a) Sonata I Minuet (b) Sonata II Minuet (c) Heavenly Rain

Fig. 2. Illustration of MFCC-EMD Features. (The solid curve represents the center of each cluster.)



(a) Sonata I Minuet (b) Sonata II Minuet (c) Heavenly Rain

Fig. 3. Illustration of Fluctuation Pattern Features.



(a) Sonata I Minuet (b) Sonata II Minuet (c) Heavenly Rain

Fig. 4. Illustration of Spectrum Histogram Features.

3.2. Fluctuation Pattern

Fluctuation pattern (FP) is a feature to describe periodicities of music tracks [4]. It uses two stages to extract the features. In the first step, a track of music is segmented into several 6 second piece sequences. The piece of music contains loudness information of a time in a specific critical-band. In the second step, it calculates the rhythm pattern which depict the strength and rate of beats within the frequency bands. Then each FP feature can be represented by a 20×60 matrix and the similarity is calculated by Euclidean metric. The FP features of the three songs are plotted in Fig. 3.

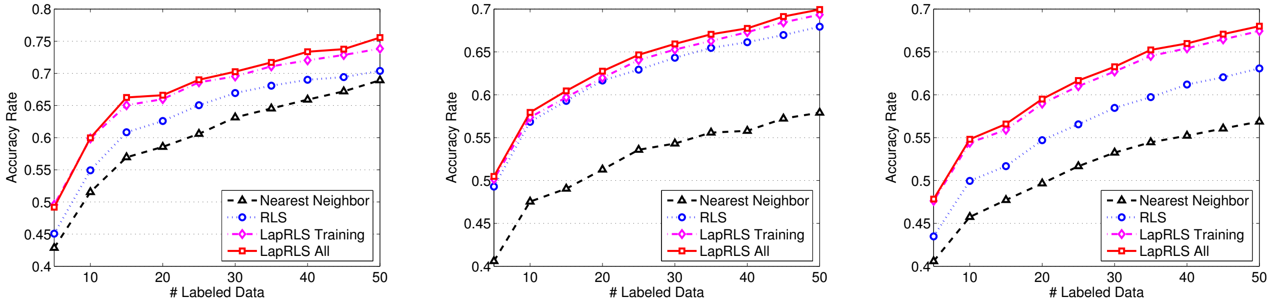
3.3. Spectrum Histogram

Spectrum histogram (SH) is a two dimensional feature which also characterizes the timbre of a music track. The advantage of this feature is the distance can be calculated by Euclidean metric, while the calculation of MFCC distance is more complex. It counts the times of a loudness level in a specific critical band being reached or exceeded. Finally we get a 20×50 matrix since there are 20 rows for critical-bands and 50 columns for loudness resolution [5]. The SH features of the three songs are plotted in Fig. 4.

4. EXPERIMENTS

In the experiments, we use the *ISMIR2004 Audio Description Contest* [16] data set for testing our proposed algorithm. There are 729 training tracks and 729 test tracks, which are classified to six genres. The distribution of different genres in the training set is *classical* (320 tracks), *electronic* (115 track), *jazz_blues* (26 tracks), *metal_punk* (45 tracks), *rock_pop* (101 tracks) and *world* (122 tracks). The original tracks are used instead of the segmented small pieces. The stereo audio signals are reduced to mono and down-sampled from 44kHz to 11kHz. We use the feature extraction and similarity measurement methods presented in section 3 to evaluate the distances of music tracks. The parameters for feature extraction algorithms are the same as the default setups in the music analysis toolbox [6].

For comparison, we implemented the nearest neighbor (NN) algorithm; the supervised regularized least-squares (RLS) and the semi-supervised graph based regularized least-squares (LapRLS). We randomly select labeled examples in each class in the training set. If there are no enough tracks for the labeled number in one class, we only use all tracks of this class. For supervised methods NN and RLS, we use the labeled data to train the classifier. For the semi-supervised method LapRLS, we run it two times. First, we use the training set, including both labeled and unlabeled data, to train the classifier (named as LapRLS_Train). Second, we use the whole data set, including both training and test data, to train the classifier (named as LapRLS_All). For all the classifiers, we evaluate the accuracy rates on the test set. The final results are show in Fig. 5. Each



(a) Accuracy Rates (b) Accuracy Rates (c) Accuracy Rates

Fig. 5. SH Classification Results.

test accuracy plotted in the figures is an average of 50 random trials. We can see that the results prove our discussion. For the MFCC-EMD feature, the accuracy rate is more than 75% when we labeled only 50 tracks of each class.

5. CONCLUSION

This paper proposes a semi-supervised way to deal with the content based music genre classification problem. The approach uses both labeled and unlabeled data to train the classifier. This mechanism can both reduce the labeling work and improve the accuracy rate for three typical timbre and rhythm features.

6. REFERENCES

[1] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in *SIGIR*, 2003, pp. 282–289.

[2] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: A survey," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133 – 141, 2006.

[3] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, 2002.

[4] E. Pampalk, A. Rauber, and D. Merkl, "Content-based organization and visualization of music archives," in *ACM Multimedia*, 2002, pp. 570–579.

[5] E. Pampalk, S. Dixon, and G. Widmer, "Exploring music collections by browsing different views," in *ISMIR*, 2003, pp. 201–208.

[6] E. Pampalk, "A matlab toolbox to compute similarity from audio," in *ISMIR*, 2004.

[7] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in *ICME*, 2001, pp. 22–25.

[8] J.-J. Aucouturier and F. Pachet, "Music similarity measures: What's the use?," in *ISMIR*, 2002, pp. 13–17.

[9] J.-J. Aucouturier and F. Pachet, "Improving timbre similarity: How high is the sky?," *Journal of Negative Results in Speech and Audio Sciences*, vol. 1, no. 1, 2004.

[10] X. Zhu, "Semi-supervised learning literature survey," Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison.

[11] M. Belkin, P. Niyogi, and V. Sindhwani, "On manifold regularization," in *Proc. AISTATS*, 2005, pp. 17–24.

[12] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. NIPS 16*, 2003, pp. 321–328.

[13] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. ICML*, 2003, pp. 912–919.

[14] R. Rifkin, G. Yeo, and T. Poggio, "Regularized least squares classification," in *Advances in Learning Theory: Methods, Model and Applications, Chapter 7*, pp. 131–154. 2003.

[15] B. Schölkopf, R. Herbrich, and Alex J. Smola, "A generalized representer theorem," in *Proc. COLT*, 2001, pp. 416–426.

[16] "ISMIR audio description contest," http://ismir2004.ismir.net/genre_contest/index.htm, 2004.