Multimedia Systems (2006) 12(1): 3–13
DOI 10.1007/s00530-006-0032-2

REGULAR PAPER

Michael I. Mandel · Graham E. Poliner · Daniel P. W. Ellis

Support vector machine active learning for music retrieval

Published online: 7 April 2006

Abstract Searching and organizing growing digital music collections requires a computational model of music similarity. This paper describes a system for performing flexible music similarity queries using SVM active learning. We evaluated the success of our system by classifying 1210 pop songs according to *mood* and *style* (from an online music guide) and by the performing artist. In comparing a number of representations for songs, we found the statistics of mel-frequency cepstral coefficients to perform best in precision-at-20 comparisons. We also show that by choosing training examples intelligently, active learning requires half as many labeled examples to achieve the same accuracy as a standard scheme.

Keywords Support vector machines · Active learning · Music classification

1 Introduction

As the size of digital music libraries grows, identifying music to listen to from a personal collection or to buy from an online retailer becomes increasingly difficult. Since finding songs that are similar to each other is time consuming and each user has unique opinions, we would like to create a flexible, open-ended approach to music retrieval.

Our solution is to use relevance feedback, specifically support vector machine (SVM) active learning, to learn a classifier for each query. A search is then a mapping from low level audio features to higher level concepts, customized to each user. To begin a search, the user presents the system with one or more examples of songs of interest or "seed" songs. The system then iterates between training a new classifier on labeled songs and soliciting new labels from the user for informative examples. Search proceeds quickly, and at every stage the system supplies its best estimate of appropriate songs. Since it takes a significant amount of time to listen to each song returned by a search, our system attempts to minimize the number of songs that a user must label for a query.

Active learning has two main advantages over conventional SVM classification. First, by presenting the user with and training on the most informative songs, the algorithm can achieve the same classification performance with fewer labeled examples. Second, by allowing the user to dynamically label the set of instances, a single system may perform any number of classification and retrieval tasks using the same precomputed set of features and classifier framework. For example, the system may be used to search for female artists, happy songs, or psychedelic music.

This flexibility depends on the information the acoustic feature vectors make explicit, leading to our comparison of various features for these tasks. On a collection of 1210 pop songs, the features that produced the most accurate artist classifier were the statistics of MFCCs calculated over entire songs. These same features achieved the best precision on the top 20 ranked results for music categorizations culled from allmusic.com.

We have also developed an automatic tester for our SVM active learning system, showing that an SVM active learner trained with half as many examples can perform as well as a normal SVM or, alternately, can increase its precision by ten percentage points with the same number of labeled examples.

1.1 Music similarity

The idea of judging the similarity of music by a direct comparison of audio content was proposed by Foote [13]. For computational simplicity, his system used discrete distributions over vector-quantizer symbols and was evaluated on a database of a few hundred 7-s excerpts. Logan and Salomon [21] was able to compare continuous distributions over thousands of complete songs, using the Earth Mover's Distance

M. I. Mandel (✉) · G. E. Poliner · D. P. W. Ellis
Department of Electrical Engineering, 1312 S.W. Mudd,
500 West 120th Street, New York, NY 10027
E-mail: {mim, graham, dpwe}@ee.columbia.edu

to calculate dissimilarity between mixtures of Gaussians. There have followed a number of papers refining the features, distance measures, and evaluation techniques, including our own work [3–5, 12]; Aucouturier and Pachet [2] provides an excellent review, where they characterize these approaches as "timbre similarity" to emphasize that they are based on distributions of short-term features and ignore most temporal structure.

Particular tasks for music similarity are largely defined by the availability of ground truth. Tzanetakis and Cook [30] popularized the use of genre classification, whereas Whitman et al. [31] proposed artist identification as a more interesting task, with the attraction of having largely unambiguous ground-truth. Here, we consider versions of both these tasks.

Most work has sought to define a low-dimensional feature space in which similarity is simply Euclidean distance, or measured by the overlap of feature distributions. Here, we use a more complex classifier (the SVM) on top of an implicit feature space of very high dimension; the related regularized least squares classifier was used for music similarity by Whitman and Rifkin [32]. The Fisher Kernel technique we use was introduced for audio classification by Moreno and Rifkin [22].

1.2 Relevance feedback

While the idea of relevance feedback had been around for a number of years, Tong and Koller [28] first described using support vector machines for active learning. Tong and Koller [29] discussed the *version space* of all possible hyperplanes consistent with labeled data along with methods for reducing it as quickly as possible to facilitate active learning. Refs. [27, 28] applied SVM active learning to text and image retrieval.

Recently, improvements in SVM active learning have been made in the areas of sample selection, scalability, and multimodal search. Chang et al. [7] described a number of methods for selecting the most informative database items to label, with the *angle diversity* selection criterion producing the best active retrieval results. The same paper describes a multimodal search in which the user can limit the pool of images to search by supplying relevant keywords. In order to scale to large databases, Lai et al. [19] describes methods for disambiguating search concepts and using a hierarchical data structure to more efficiently find data points.

Hoashi et al. [15, 16] used relevance feedback for music retrieval, but their approach suffers from some limitations. Their system was based on the TreeQ vector quantization from Ref. [13], with which they must re-quantize the entire music database for each query. Relevance feedback was incorporated into the model by modifying the quantization weights of desired vectors. Our approach calculates the features of a song offline and uses SVM active learning, which has a strong theoretical justification, to incorporate user feedback.

1. Seed the search with representative song(s).
2. Acquire initial negative examples by e.g. presenting randomly selected songs for labeling
3. Train an SVM on all labeled examples
4. Present the user with the most relevant songs (those with the greatest positive distance to the decision boundary)
5. If the user wishes to refine the search further, present the most informative songs (those closest to the decision boundary) for labeling and repeat 3-5.

Fig. 1 Summary of SVM active learning algorithm

2 SVM active retrieval

SVM active learning combines the maximum margin classification of SVMs with ideas from relevance feedback. See Fig. 1 for a summary of the active learning algorithm, which lends itself to both direct user interaction and automated testing.

2.1 Support vector machines

The support vector machine (SVM) is a supervised classification system that minimizes an upper bound on its expected error. It attempts to find the hyperplane separating two classes of data that will generalize best to future data. Such a hyperplane is the so called maximum margin hyperplane, which maximizes the distance to the closest points from each class.

More concretely, given data points $\{\mathbf{X}_0, \ldots, \mathbf{X}_N\}$ and class labels $\{y_0, \ldots, y_N\}$, $y_i \in \{-1, 1\}$, any hyperplane separating the two data classes has the form

$$y_i(\mathbf{w}^T \mathbf{X}_i + b) > 0 \; \forall i. \tag{1}$$

Let $\{\mathbf{w}_k\}$ be the set of all such hyperplanes. The maximum margin hyperplane is defined by

$$\mathbf{w} = \sum_{i=0}^{N} \alpha_i y_i \mathbf{X}_i, \tag{2}$$

and b is set by the Karush Kuhn Tucker conditions [6] where the $\{\alpha_0, \alpha_1, \ldots, \alpha_N\}$ maximize

$$L_D = \sum_{i=0}^{N} \alpha_i - \frac{1}{2} \sum_{i=0}^{N} \sum_{j=0}^{N} \alpha_i \alpha_j y_i y_j \mathbf{X}_i^T \mathbf{X}_j, \tag{3}$$

subject to

$$\sum_{i=0}^{N} \alpha_i y_i = 0 \quad \alpha_i \geq 0 \; \forall i. \tag{4}$$

For linearly separable data, only a subset of the α_is will be non-zero. These points are called the *support vectors* and all classification performed by the SVM depends on only these points and no others. Thus, an identical SVM would