MIREX-2011 SINGLE-LABEL AND MULTI-LABEL CLASSIFICATION TASKS: IRCAMCLASSIFICATION2011 SUBMISSION

Damien Tardieu, Christophe Charbuillet, Frédéric Cornu, Geoffroy Peeters
Ircam Sound Analysis/ Synthesis Team - CNRS STMS
1, pl. Igor Stravinsky - 75004 Paris - France
{dtardieu,charbuillet,cornu,peeters}@ircam.fr

1. INTRODUCTION

This extended abstract describes the system submitted by IRCAM to the MIREX Evaluation 2011 for the tagging and classification (Train/Test) tasks. The system, named ircamclassification11, is briefly described here. We first describe the feature extraction methods then we explain the classification method. Finally some implementation details are presented.

2. FEATURE EXTRACTION

2.1 Low Level Features

We extract two low level features, Mel Frequency Cepstral Coefficients (MFCC) and Spectral Flatness Measure (SFM). They are extracted using a 40 ms Blackmann window with a 20 ms overlap.

2.2 GMM Supervector

2.2.1 Universal Background Model

The Universal Background Model (UBM) aims at modeling the overall data distribution. It consists of a classical Gaussian Mixture Model. The UBM is usually composed of Gaussian models with diagonal covariance matrix. The loss of modeling ability due the diagonal covariance matrix can be compensated by increasing the number of Gaussian in the mixture The UBM is trained using a large and representative set of data by using the Expectation Maximization (EM) algorithm. The system is provided with a pre trained UBM.

2.2.2 UBM adaptation

The UBM adaptation is the process of modifying the UBM parameters in order to fit a particular data distribution. In our application, this subset is the data extracted from a track to model. This adaptation is made using the Maximum A Prosteriori (MAP) approach [3].

2.2.3 GMM supervector

To summarize, a music track model is directly derived from a generic GMM, estimated using a large set of representative data (the so called UBM). During the adaptation process, only the mean vectors of the Gaussians are modified to fit the particular music track distribution. Consequently, all the the music track models have both the same covariance matrix and weight. Knowing the parameter of the UBM, a particular music model can be summarized by the mean vectors of its Gaussian mixture components. The mean vectors are thus stacked in a one dimensional vector.

2.3 Autoregressive Vector Model

An autoregressive vector model (ARV) is also computed on the low level feature vector [1].

2.4 Autoregressive Vector Model of the residual

In the second configuration of the system, a gmm supervector is computed on the residual of the first ARV model.

3. CLASSIFICATION

As a classifier, we use support vector machines (SVM) with a Gaussian radial basis function kernel. We set $\gamma = 1/d$ where d is the dimension of the feature set and $C = 1$. The implementation is the one of LIBSVM [15]. To make a multi-class classifier from the 2-class SVM we use the one versus all method. We train a classifier for each class versus all the remaining classes. To make a decision we compare the posterior probabilities provided by LIBSVM. In the multi-label case, a tag is affected to the incoming song if the posterior probability of this tag is higher than .5. In the single-label case, the class with the highest probability is attributed to the song.

4. IMPLEMENTATION DETAILS

The low level feature extraction module is based on the executable *mirex_mfccsfm_extractor*, which outputs the computed features in a binary format. The computation of the supervectors and ARV models is done in matlab. The overall computation time for a 30s, 22kHz, 16 bit mono WAV audio is around .4 seconds as measured on an Intel Xeon 64 bit CPU at 2.4GHz and 16GB RAM. The memory usage is very small. The required disk space per audio file is around 125kB. These figures hold for both systems. SVM training and classification are performed by the libsvm library [2]. The total training runtimes depends heavily on the size of the database and number of classes. It can range from a few minutes for small databases with a small number of classes (5 to 10 classes) to around one hour for large databases with a large number of classes (around 100). As an exemple the training and classification procedure for a 14 tags problem on a 3600 songs database lasts around 25 minutes using 4 cores (by setting the environment variable OMP_NUM_THREADS to 4) on a Intel Xeon 64 bit CPU

at 2.4GHz and 16GB RAM. The memory usage is around 2 GB.

5. ACKNOWLEDGEMENTS

This work was partly supported by “Quaero” Programme, funded by OSEO, French State agency for innovation

6. REFERENCES

[1] F. Bimbot, L. Mathan, A. De Lima, and G. Chollet. Standard and target driven ar- vector models for speech analysis and speaker recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 5–8, 1992.

[2] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1—-27:27, 2011.

[3] DA Reynolds, TF Quatieri, and RB Dunn. Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, 2000.