MIREX 2012: AUDIO TAG CLASSIFICATION USING NEW FEATURES AND GRID SEARCH FOR SVM

Simon Bourguigne **Pablo Daniel Agüero**

Facultad de Ingenieria,
Universidad Nacional de Mar del Plata
Mar del Plata, Argentina
simonbour@gmail.com

ABSTRACT

In our submission we use a straight forward method for the task of audio tag classification. This extended abstract briefly describes the features used and the classification method.

1. INTRODUCTION

In this paper we present our system for Audio Tag Classification for the MIREX 2012 competition. The proposed system takes a very simple approach. It uses the standard procedure of frame-level audio feature extraction and posterior aggregation. Once these features are obtained they are fed to a number of SVMs binary-classifiers that equate with the number of tags. One of the difficulties lies at optimizing each classifier for better performance. In order to do so, we iterate over different values for the SVM parameters doing a simple grid search and using n-fold inner cross validation.

2. FEATURE EXTRACTION

The feature extraction process is done on a frame-based fashion using an in-house feature extractor named Ursula. We use 50ms hamming windows and a hop-size of 25ms. After the feature extraction each frame has a feature array associated to it. This is followed by grouping contiguous frames into a 1s texture window and for each of the later we aggregate the features using mean and standard deviation values. Later on, we further aggregate these values computing the same statistics over them. This produces mean-mean, mean-std, std-mean, and std-std values that we use to aggregate the whole set of features for the specific sound clip. Once the aggregation is done, each numeric feature will have four times it's length, ie. mfcc will have 72 numeric values instead of 13. A detailed description of the features used is shown in Table 1.

Feature Description	Dim
lpcc	72
mfcc	52
lsp	72
spectral flatness	96
spectral crest factor	96
spectral flux	4
spectral decrease	4
loudness [3]	4
roll-off at 95%	4
zero crossing rate	4
formant band energy (250-2500Hz)	4
odd to even energy ratio	4
harmonic coefficient [1]	4
beat histogram (non-aggregated)	9

Table 1. The audio features used for classification.

2.1 Additional VQ feature

One of the present submissions (BA1) has an additional feature that is calculated using the distances between centroids. Each clip consists of N frames, and therefore N MFCC vectors of dimension 13 are available. By means of a vector quantization algorithm, the Linde, Buzo, and Gray (LBG) proposal [4], M centroids are obtained by clustering the N frames. As a consequence there exist M centroids for each clip in the feature extraction process.

Training data of each tag has N_P positive and N_N negative examples. The $N_P \cdot M$ centroids of the positive examples are clustered during training and classification to obtain P centroids. Some of these global P centroids are expected to represent the tag under consideration.

Then, for each clip with M centroids, the closest centroid to each global centroid is obtained, and the Euclidean distance for each pair is stored in a P dimensional vector. Each P dimensional vector of distances is considered as an additional feature for SVM classification algorithm.

2.2 MFCC sequence compression

The other submission (BA2) considers the time-series (and it's time structure) obtained when computing the MFCCs

for each individual frame. Linear Predictive Coding is used to compress the sequence down to 18 coefficients for each MFCC. This results in a 13×18 feature array that will be used to feed the SVM.

3. THE CLASSIFICATION METHOD

In this section we present our classification method. The tags will be assumed to be independent, that means that only one binary classifier will be used for each of them. The chosen classifiers are SVMs with linear kernels, which have only two parameters to be optimized.

In a linear SVM there are two parameters to be set, $C > 0$ is the penalty parameter of the error term of the classifier, and W is a penalty of the wrong classification for positive $(+1)$ and negative (-1) examples. To find the optimal set of these parameters we perform a simple grid search. We define two arrays for different values for C and W. Later on the various pairs of (C; W) values are tried and the one with the best inner cross-validation F-measure is picked [2]. The inner cross-validation in the submissions consists of three folds.

4. REFERENCES

[1] W. Chou and L. Gi : "Robust singing detection in speech/music discriminator design," *Proceedings of ICASSP*, Salt Lake, 2001.

[2] C.-W. Hsu, C.-C. Chang, C.-J. Lin : "A practical guide to support vector classification," *Technical report*, Department of Computer Science, National Taiwan University. July, 2003.

[3] S. Streich : "Music Complexity: a multi-faceted description of audio content," *Ph.D. Dissertation*, UPF, Barcelona, 2007.

[4] Y. Linde, A. Buzo, R. M. Gray, : "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Communications*, 1980.