MUSIC TYPE GROUPERS (MTG): GENERIC MUSIC CLASSIFICATION ALGORITHMS

N. Wack, E. Guaus, C. Laurier, O. Meyers, R. Marxer, D. Bogdanov, J. Serrà, P. Herrera
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
{nicolas.wack,enric.guaus,cyril.laurier,owen.meyers,ricard.marxer,
dmitry.bogdanov,joan.serraj,perfecto.herrera}@upf.edu

ABSTRACT

This paper outlines our submissions to different music classification tasks for the Music Information Retrieval Evaluation eXchange (MIREX) 2009. We detail here three different algorithms tested in mood and genre classification tasks, and in classical composer identification. These algorithms are based on Support Vector Machines, Disjoint Principal Components Models, and RCA-kNN. The last one utilizes Euclidean distances in a reduced space using Relevant Component Analysis and Kullback Leibler divergence on Mel Frequency Cepstrum Coefficients (MFCC).

1. FEATURE EXTRACTION

The submissions are coded in C++ and python. For the feature extraction part, we use an internal library of the Music Technology Group called *Essentia* [1]. This library contains the features outlined below. We divide our features in two main categories. The "base" features which are state-of-the-art MIR features and the "high-level" features. We aggregate frame-based descriptions using mean and derivatives until second order, variance and derivatives until second order, minimum, and maximum.

1.1 Base features

In Table 1 we summarize the set of base features that performed the best in our preliminary experiments made with in-house genre, artist, and mood ground truths.

Type	Features
Low level:	barkbands spread, skewness, kurtosis, dissonance, hfc pitch and confidence, pitch salience, spectral complexity spectral crest, spectral decrease, energy, spectral flux spec spread/skewness/kurtosis, spec rolloff, strong peak ZCR, barkbands, mfcc
Rhythm:	bpm, beats loudness, onset rate
Sound FX:	inharmonicity, odd2even, pitch centroid, tristimulus
Tonal:	chords strength (frame), key strength(global), tuning freq

Table 1. Feature set used in all submitted approaches.

1.2 High-level features

One key part of our approach is the integration of high level descriptors. Our assumption is that low level features are both powerful and limited. Powerful because they can model many problems, but limited because using only those we miss to model semantic aspects of our decision mechanisms. Based on this idea, we added high level features of different categories: genre, mood, etc. These models are pre-trained using an SVM algorithm on various groundtruth data sets, and the resulting feature space contains the probability values of each class for each SVM [8]. Those are then added to our bag of features.

2. CLASSIFICATION

The three classification algorithms are coded in C++ and python. They are implemented using Gaia, a library for manipulating datasets and computing similarity distances. These algorithms are based on Support Vector Machines, Disjoint Principal Components Models, and RCA-kNN. Each algorithm has the option to look for its best parameters with a grid-search cross-validation approach on the training data. Therefore, the total number of submitted algorithms was 6: 3 with best parameter grid-search and 3 without.

2.1 Relevant Component Analysis and Nearest Neighbours

Relevant Component Analysis (RCA) is a supervised transformation which aims at maximizing the global variance of a dataset while reducing the intra-class variance (representing unwanted variability). The algorithm is split in two parts. The first part is the dimensionality reduction that consists in applying a modified version of the Fisher Linear Discriminant (FLD) where we only use part of the classified vectors for training. This transformation amounts to resolving the following estimator:

[1] Essentia & Gaia: audio analysis and music matching C++ libraries developed by the MTG (Resp.: N. Wack), http://mtg.upf.edu/technologies/essentia

$$\max_{A \in M_{P \times Q}} \frac{A^t S_t A}{A^t S_w A}, \quad (1)$$

and transforming from a space with P dimensions to a space with Q dimensions where A is the searched transformation matrix, $M_{P \times Q}$ is the space of all transformations, S_t is the total covariance matrix and S_w is the inner-class covariance matrix.

The second part consists in applying the actual RCA transformation, which scales down those dimensions that have great variability within our classes by whitening the resulting feature space. We first calculate the covariance for all the centered data-points in the classes:

$$\hat{C} = \frac{1}{p} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ji} - \overline{x}_j)(x_{ji} - \overline{x}_j)^t, \quad (2)$$

where p is the total number of points in the classes and $\overline{x}_j$ is the mean of the data-points of the class j. Finally we obtain the whitening matrix

$$W = \hat{C}^{-\frac{1}{2}}, \quad (3)$$

so the new feature space is given by

$$x_{new} = Wx. \quad (4)$$

Our classification algorithm is made of a K-NN classifier using a weighted distance based on two distances. One is from the reduced space mentioned previously where we use the euclidean distance. The other is the Kullback-Leibler distance applied to MFCCs.

$$Dist = \alpha(KL_{MFCC}) + (1 - \alpha)(Euclidean_{RCA}) \quad (5)$$

We optimize the weight α between both distances with a cross-validation technique on the training set.

2.2 Disjoint Principal Components Models

The disjoint principal components modeling architecture was proposed by Wold [3]. In our submissions, we use its Soft Independent Modeling of Class Analogies (SIMCA) [4] implementation for classification. This implementation is specially useful for high-dimensional classification problems because it uses Principal Components Analysis (PCA) applied to each category individually for dimensionality reduction. By using this structure, SIMCA provides information on different groups such as the relevance of different dimensions and measures of separation. It is the opposite when applying PCA to the full set of observations because the same reduction rules are applied through all the original categories. The goal of SIMCA is to obtain a classification rule for a set of m known groups.

Let X^j be the m groups where j indicates the class membership ($j = 1 \ldots m$). The observations of group X^j are represented by x_i^j, where $i = 1 \ldots n_j$ and n_j is the number of elements in the group j. Now, let p be the number of variables for each element providing $x_i^j = (x_{i1}^j, x_{i2}^j, \ldots, x_{ip}^j)'$. Finally, let Y^j be the validation set, with $j = 1 \ldots m$. The goal of SIMCA is not only the classification itself but also to enhance the individual properties of each group. Then, PCA is performed on each group X^j independently. This produces a matrix of scores T^j and loadings P^j for each group. Let $k^j << p$ be the retained number of principal components for group j. Let y be a new observation to be classified, and let $\tilde{y}^{(l)}$ represent the projection of this observation on the PCA model of group l:

$$\tilde{y}^{(l)} = \bar{x}^l + P^l (P^l)' (y - \bar{x}^l) \quad (6)$$

where $\bar{x}^l$ is the mean of the training observations in group l. The classification is carried out in terms of a linear combination of the *orthogonal distance* (OD) and the *score distance* (SD), which are the euclidian measure of an observation to the spaces panned by the first k PCs and a Mahalanobis-like measure of distance of an observation within the PC space, respectively. For further details of this modeling architecture and applications to MIR tasks we refer to [5].

2.3 Support Vector Machines

Support Vector Machines (SVM) [1], is a widely used supervised learning classification algorithm. It is known to be efficient, robust and to give relatively good performance and it protects against overfitting because of the structural risk minimization that is at the core of the algorithm. Indeed, this classifier is widely used in Music Information Retrieval (MIR) research. In a previous MIREX classification task (Audio Mood Classification), we submitted an algorithm based on SVMs that performed relatively well [6]. Indeed, most of the best performing algorithms for classification use a SVMs. The submitted algorithm is based on a famous implementation of SVMs called libsvm [2]. In preliminary analysis, we tried different kernel methods: linear, polynomial, radial basis function (RBF) and sigmoid. We found that the better and more robust kernel is the RBF. Even if an RBF kernel is not always recommended for large feature sets compared to the size of the dataset [1], we had a good accuracy using this kernel for all tasks. It may not be the best solution always, but it offers a good compromise in average. To find the best parameters, we implemented a grid search on the training data (with a cross-validation approach).

3. EVALUATION

We submitted our algorithms to all the train/test tasks: audio classical composer identification (CI), audio genre classification (GC), and audio music mood classification (MC). The CI data set consisted of 2772 30 second audio clips and the composers represented were: Bach, Beethoven, Brahms, Chopin, Dvorak, Handel, Haydn, Mendelssohnn, Mozart, Schubert, and Vivaldi. The goal was to correctly identify the composer who wrote each of the represented pieces. The GC task was evaluated on two different collections. The first collection was the so-called "mixed set" collection. It was composed of 7000 30-second audio clips

Task	Num. of Participants	Num. of Algorithms	Best MTG Ranking	Best MTG Algorithm	Worst MTG Ranking
CI	16	30	1	RCA2	22
GC - Latin	17	34	6	SVM2	30
GC - Mixed	17	31	5	SVM1	23
MC	17	33	4	SVM1	24

Table 2. Summary of our best and worst submitted algorithms.

	RCA1		RCA2		SIM1		SIM2		SVM1		SVM2		Best
	Acc.	#	Acc.	#	Acc.	#	Acc.	#	Acc.	#	Acc.	#	subm.
CI	54.7	9	**62.1**	**1**	48.12	22	48.2	21	49.8	18	50.4	17	62.1
GC - Latin	61.7	10	62.4	9	45.8	30	47.8	28	61.1	11	**63.1**	**6**	74.7
GC - Mixed	64.8	18	–	–	64.1	22	64.0	23	**70.4**	**5**	–	–	73.3
MC	57.7	23	57.5	24	59.8	12	59.3	16	**62.8**	**4**	59.5	14	65.7

Table 3. Obtained results for our submitted algorithms.

in 22.05kHz mono WAV format drawn from 10 genres (700 clips from each genre). The genres were: blues, jazz, country/western, baroque, classical, romantic, electronica, hip-hop, rock, and hard rock/metal. The second collection was the so-called "latin set" collection. It was composed of Latin popular and dance music, sourced from Brazil and hand-labeled by music experts. This collection was likely to contain a greater number of styles of music that will be differentiated by rhythmic characteristics than the "mixed set" collection. The "latin set" collection contained 3,227 audio files from 10 Latin music genres: axé, bachata, bolero, forró, gaúcha, merengue, pagode, sertaneja, and tango. The MC task dataset and evaluation procedure was the same as in previous editions [7]. More details about the evaluation procedure can be found in the MIREX wiki [2]. Our algorithms follow this naming convention: (1) RCA1: Relevant Component Analysis (MTG1 in the wiki), (2) RCA2: Relevant Component Analysis with with best parameter grid-search (MTG2 in the wiki), (3) SIM1: Soft Independent Modeling of Class Analogies (MTG3 in the wiki), (4) SIM2: Soft Independent Modeling of Class Analogies with with best parameter grid-search (MTG4 in the wiki), (5) SVM1: Support Vector Machines (MTG5 in the wiki), and (6) SVM2: Support Vector Machines with with best parameter grid-search (MTG6 in the wiki).

4. RESULTS

We now briefly highlight the results obtained by our submissions across the different tasks. For more details we refer to the Results MIREX wiki [3]. In Table 2 we show a summary of the best and worst rankings achieved by our algorithms. Our best performing algorithms for the different tasks were RCA for the CI task and SVM for the other tasks. In Table 3 we show a summary for all our proposed algorithms.

[2] http://www.music-ir.org/mirex/2009

[3] http://www.music-ir.org/mirex/2009/index.php/MIREX2009_Results

5. DISCUSSION

In general the submitted algorithms performed quite well in all tasks. The ranks ranged from 1 to 23 and they didn't reach the very last positions. The fact that the same classifier performs differently for different tasks highlights that maybe there is no general architecture that is best-performing across all tasks. However, one should note that the algorithms submitted by Cao & Li outperformed the rest in the majority of the tasks. When information about these algorithms is available we should study what are their strengths and our weaknesses. As the diversity of submitted algorithms is high it could be possible that the error patterns would be equally diverse. In this case, it would be interesting to study their combination as a mixture of experts, if only using a simple voting scheme.

Focusing in our results, we observe how, although RCA reaches the 1st position for the CI task, SVM is in general the best classification technique. Rankings for RCA decrease below those obtained by SVM for GC and MC tasks. Focusing on the results obtained by the 2 variations of the SIMCA algorithm (referred as MTG3 and MTG4 in the wiki), we realize that results are not so good as expected. They rank at 21st. and 22nd. position for CI, 28th. and 30th. position for GC-Latin, 22th. and 23th. position for GC-Mixed, and 12th. and 16th. position for Mood. We will analyze these results focusing on the implementation and the classifier itself. On the other hand, no clear conclusions are extracted when comparing results of the algorithms that use grid search for best parameter research (RCA2, SIM2, SVM2) with respect those than do not use it (RCA1, SIM1, SVM1).

Finally, the MC task allows a comparison of this year submissions with last year's ones. In our case, the accuracy of our algorithms ranges from 57.5 to 62.83. Whereas the highest value represents a small improvement over the 61% of the 2008 submission (and it corresponds in both cases to a SVM classifier), the remaining algorithms have shown a decrease in accuracy that needs to be carefully studied.

6. ACKNOWLEDGMENTS

We want to thank the people from the Music Technology Group (Universitat Pompeu Fabra, Barcelona) and especially those who contributed to Essentia, particularly Eduard Aylon and Roberto Toscano. This research has been partially funded by the EU Project PHAROS IST-2006-045035.

7. REFERENCES

[1] B. E. Boser, I. M. Guyon and V. N. Vapnik. A training algorithm for optimal margin classifiers. *In COLT '92: Proceedings of the fifth annual workshop on Computational learning theory, (pp. 144-152)*. New York, NY, USA: ACM, 1992.

[2] C. C. Chang, C. J. Lin. LIBSVM: a library for support vector machines, 2001. *Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm*

[3] S. Wold. Pattern recognition by means of disjoint principal components models. *Pattern Recognition, Vol. 8, pp. 127-139*. 1976.

[4] K. Vanden and M. Hubert. Robust classification in high dimensions based on the SIMCA method. *Chemometrics and Intelligent Laboratory Systems, Vol. 79, pp. 10.21*. 2005.

[5] E. Guaus. Audio content processing for automatic music genre classification: descriptors, databases, and classifiers. PhD Thesis. Universitat Pompeu Fabra, 2009.

[6] C. Laurier and P. Herrera. Audio music mood classification using support vector machine. *Music Information Retrieval Evaluation eXchange (MIREX) extended abstract*, 2007.

[7] X. Hu, J. S. Downie, C. Laurier, M. Bay and A. Ehmann. The 2007 MIREX Audio Mood Classification Task: Lessons Learned. *9th International Symposium on Music Information Retrieval (ISMIR 2008)*, Philadelphia, Sept. 2008.

[8] D. Bogdanov, J. Serrà, N. Wack and P. Herrera. From Low-level to High-level: Comparative Study of Music Similarity Measures. *International Workshop on Advances in Music Information Research*, in Press.