A MIXTURE OF SUPPORT VECTOR MACHINES FOR AUDIO CLASSIFICATION

Nicolas Scaringella

Signal Processing Institute
(ITS-LTS-3)
Swiss Federal Institute of Technology
(EPFL)
Lausanne, CH-1015 Switzerland
nicolas.scaringella@epfl.ch

Daniel Mlynek

Signal Processing Institute
(ITS-LTS-3)
Swiss Federal Institute of Technology
(EPFL)
Lausanne, CH-1015 Switzerland
daniel.mlynek@epfl.ch

ABSTRACT

This paper describes the algorithm submitted by the authors to the Audio Genre Classification Contest organised in the context of the 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005). The proposed algorithm parameterizes audio content by extracting 3 sets of features describing 3 different dimensions of music: timbre, energy and rhythm. Once features extracted, a mixture of Support Vector Machines (SVMs) is used for classification into musical genres. The underlying idea is to use separate models to approximate different parts of the problem and to combine the outputs from the experts with probabilistic methods. Using the proposed algorithm, classification of 73.11 % is achieved on the 2 databases used for the MIREX 2005 contest containing a total of 2929 songs.

Keywords: musical genres, features extraction, Support Vector Machines, mixture of experts.

1 INTRODUCTION

Musical genres are the main top-level descriptors used by music-dealers and librarians to organize their music collections. Though they may represent a simplification of one artist's musical discourse, they are of a great interest as summaries of shared characteristics in music pieces.

With Electronic Music Distribution (EMD), music catalogues tend to become huge; in that context, associating a genre to a musical piece is crucial to help users in finding what they are looking for. In fact, the amount of digital music data urges for new means of automatic annotation since manual labelling would be too time consuming.

Motivated by such concerns, a number of researchers have investigated methods for associating automatically a music genre to an audio excerpt. Consequently, state-of-the-art audio classification algorithms have been evaluated in the context of MIREX 2005.

This paper presents the algorithm submitted by the authors to the MIREX 2005 audio classification contest. It parameterizes audio content by extracting 3 sets of features describing 3 different dimensions of music: timbre, energy and rhythm. Once features extracted, a mixture of SVMs is used for classification into musical genres. The underlying idea is to use separate models to approximate different parts of a problem and to combine the outputs from the experts with probabilistic methods.

This paper is organized as follows. Section 2 describes the extraction of features from the audio signal. The architecture of the classifier is discussed in section 3 while section 4 presents the results obtained with this algorithm at the MIREX 2005 contest.

2 FEATURES EXTRACTION

2.1 Segmentation into analysis frames

The audio excerpts used are sampled at 44100 Hz and converted to mono signals. The first 30 seconds of the signals are discarded to avoid introductions that may not be representative of the rest of the excerpt. Only the next 30 seconds of the signal are kept for further analysis to limit further processing. The resulting signals are analysed through sliding windows of 20 ms overlapped by 50 %. In the case of genre classification, it is probable that these precision requirements could be relaxed. West and Cox [1] use audio signals sampled at 22050 Hz and no overlap between frames without significant loss in classification accuracy. Further experiments have to be run in our case to check if the system is robust to signals with reduced quality.

2.2 Texture windows

Frames of 20 ms are used for short time Fourier transform analysis since they allow representing the evolution of the spectrum with a good precision. Yet this time scale too many variations occur. Some integration process must be held to build more robust features. Not only does it reduce further computations but it is also more perceptually relevant. Consequently, *texture windows* are used to combine low-level features of adjacent analysis frames.

The impact of the size of the window over classification accuracy has been studied in [2]. The conclusion is that texture windows of 1 second are a good compromise since no significant gain in classification accuracy is obtained by taking larger windows while the accuracy decreases as the window is shortened.

We experimented in [3] with texture windows centred on time positions of musical beats. The sizes of the corresponding windows were selected in accordance with the local beat rate of the excerpt. Though this may allow a perceptually more relevant modelling of musical signals, no significant improvement of classification accuracy has been obtained with this technique, probably because of the weaknesses of state-of-the-art beat trackers. Consequently, the algorithm presented here use simple 1-second texture windows.

2.3 Timbre features

Mel-frequency Cepstral Coefficients (MFCCs) are computed from the analysis frames. MFCCs are widely used descriptors for timbre modelling coming from the speech recognition literature [4]. Each analysis frame is parameterized with 6 MFCCs. The number of MFCCs used has been chosen to limit the further computations rather than by a careful analysis of its impact on the classification accuracy, though the number of MFCCs is a subject of debate in the literature [5]. Mean, standard deviation and skewness over the texture window are evaluated for each MFCC resulting in a vector of dimension 18.

2.4 Energy features

Log-compressed energies in 6 frequency bands are extracted from each analysis frame. Each band covers roughly one octave. Mean, standard deviation and skewness of each coefficient are evaluated over the texture window. The low-energy feature is also computed. It measures the percentage of frames within the texture window that have energy lower than the mean energy across the texture window (notice that we evaluate the low-energy feature with the energy across all bands rather than the energy over the 6 bands). This results eventually in a vector of dimension 19 for each texture window.

2.5 Rhythm features

A number of features describing local rhythm are evaluated from a *periodicity function* obtained in two steps. Firstly, banks of comb-filters are used to analyse the fluctuation of energy in 40 Mel frequency bands. The fluctuations in each band are then combined in a manner similar to [6]. The result is a measure of energy as a function of time periodicity. The periodicity function of each analysis frames are then averaged over the texture window. The tactus (the time between beats) and the bar measure length are extracted from the averaged periodicity function as described in [6]. The ratio of the measure length over the tactus is computed as well. Eventually, mean, standard deviation, skewness and kurtosis of the periodicity function are computed resulting in a vector of dimension 7.

3 CLASSIFICATION

3.1 Support Vector Machines

Support Vector Machines (SVMs) have shown excellent results for data classification and regression [7]. Their success in practice is based on two properties: margin maximization (which allows for a good generalization of the classifier) and non-linear transformation of the feature space with kernels (a data set is indeed more easily separable in a high dimensional feature space). SVMs are designed to discriminate between 2 classes. A number of approaches exist to extend SVMs for multi-class classification [8]. This algorithm uses a simple one-versus the rest strategy - for each class, a SVM is trained to separate the current class from all of the other classes.

3.2 Strategy to handle temporal patterns

SVMs like most machine learning algorithms are only able to manipulate static patterns while music has an inherent temporality. A simple solution to handle temporal sequences is to build a spatial representation out of it and use it as input of the classifier. Like the texture window at the analysis frame rate, sequences of adjacent feature vectors are presented to the classifier through a tapped delay line rather than isolated vectors. Notice however that this scheme suffers from a number of weaknesses:

1. Since feature vectors are concatenated into a larger one, the number of parameters of the classifier is increased and thus a larger number of examples are needed for a good training.
2. The classifier is not invariant to time shifting i.e. a very large number of example patterns is needed for every output class and every position in the delay line.
3. The classifier is sensitive to time variation i.e. it requires the delays to precisely match the input time intervals (this may be corrected by having feature vectors synchronised on beat positions).

Anyway, this simple scheme proved to be very efficient in earlier experiments reported in [3]. The proposed algorithm uses a delay line of 3 feature vectors or in other words, each pattern is represented as the concatenation of 3 adjacent feature vectors in such a way that contextual information is taken into account to classify a single vector.

3.3 Mixture of experts

A mixture of experts (or classifiers) solves a classification task by decomposing it into a series of subproblems. Not only does it reduce the complexity of each single task but it also improves the global accuracy by combining the results of the different classifiers. Of course, the number of needed classifiers is increased but each having a simpler problem to handle, the overall required computational power is reduced. It is particularly meaningful with SVM experts, which suffer from the complexity of their training (which is at least quadratic with respect to the number of examples).

When using a mixture of classifiers, each subtask may focus either on a subset of the attributes (feature selection); on a different sample of the data (subsampling, bagging, boosting…); or on a different relabelling of the data (decomposition of polychotomies into dichotomies). This allow to handle the scaling problem since each expert manipulate less data than in the original problem while at the same time it allows to improve the global accuracy beyond that of the best expert since the errors of the learners will not be too positively correlated because of the different data set given to each expert.

The proposed algorithm uses three SVM experts, each one using one of the three features set presented earlier (timbre, energy or rhythm). The architecture of the classifier is depicted on figure 1. The decomposition of the training data set into small subsets given to more experts is planned for future works to allow for faster training and possibly better generalization as suggested in [9].

A range of solutions has been proposed in the literature for the combination of different models into a global system. The simplest solution is to use a majority vote of the different experts and has been applied to music genre classification in [10]. The so-called mixture of experts' architecture is a probabilistically motivated method for combining models: the outputs of the experts are combined using a gate whose outputs are probabilities of selecting the experts given the inputs. More formally, the posterior probability of class c given input vector x is given by:

$$p(c|x) = h\left(\sum_{m=1}^{M} w_m(x) s_m(x)\right) \quad (1)$$

where M is the number of experts in the mixture, $s_m(x)$ is the posterior probability of class c given input vector x and expert m, $w_m(x)$ is the probability of selecting expert m given input vector x and is given by a gater module. h is a transfer function, typically an hyperbolic tangent for classification tasks. The gater module is implemented with a single layer neural network while the experts are SVMs.

A number of experiments have been made with this model. Yet the algorithm submitted for MIREX 2005 implements a simpler scheme since the one presented here was not working properly at the time of the contest. The results presented in the next section were obtained by simply averaging influences of each expert

$$p(c|x) = \frac{1}{M}\sum_{m=1}^{M} s_m(x) \quad (2)$$

A complete musical excerpt is classified by averaging posterior probabilities over all input feature vectors and by selecting the class with the highest averaged posterior probability.

4 EXPERIMENTAL RESULTS

4.1 Datasets

The algorithm presented in this paper was evaluated in the context of MIREX 2005 on two different sets of polyphonic musical audio files in PCM format (mono files samples at 44100 Hz):

1. The MAGNATUNE dataset consists of 1005 training files and 510 testing files distributed over a hierarchical genre taxonomy of 10 genres.
2. The USPOP dataset consists of 940 training files and 474 testing files distributed over a flat genre taxonomy of 6 genres.

4.2 Results

Classification results obtained on the 2 databases are reported in the confusion matrices of table 1 and 2 illustrating information about actual and predicted classifications. Report to the MIREX website[1] for details on the evaluation procedure.

The classification accuracy obtained on the USPOP dataset is of 75.74 %. The classification accuracy on the same dataset with normalization of the results is of 77.67 %.

The classification accuracy obtained on the MAGNATUNE dataset if of 66.14 % while it is of 70.74 % when considering the hierarchical structure of the taxonomy, Normalized classification accuracies are of 67.12 % and 72.30 %, respectively for the raw and hierarchical measures.

The mean unnormalized classification accuracy over the 2 databases is of 73.11 %.

5 CONCLUSION

The obtained classification accuracy of 73.11 % is encouraging and comparable to other state-of-the-art algorithms evaluated in the MIREX 2005 contest (12 algorithms with accuracies between 53.45 % and 81.77 %). By analysing carefully confusion matrices, one can notice that classification errors make sense: for example, on the MAGNATUNE dataset, 20.59 % of Punk excerpts are classified as Rock. There is indeed a clear overlap between these two genres and the misclassified examples may have been probably better described as belonging to both classes.

As a matter of fact, the classification paradigm used in the audio genre classification contest was thought for strict classification: one excerpt must belong to one class. Yet it may be hard to fit unambiguously one song into one box. Taking into account ambiguity or in other words, allowing multiple labels classification is probably closer to the human experience in general, for sure to the artist's point of view. Artists usually produce music without concerning themselves in which genre they are working. Furthermore in most Internet based classifications, artists, albums or titles are typically associated to a number of genres.

In other fields of information retrieval and machine learning in which research is probably more advanced, solutions have already been proposed to deal with ambiguity of a real-world classification problem. The next step for MIR researcher is to follow these advances and to consider multiple labels classification schemes.

ACKNOWLEDGEMENTS

The authors would like to thanks the EC IST FP6 for the partial funding of the AXMEDIS project (www.axmedis,org) and to express gratitude to all AXMEDIS project partners internal and external to the project including the Expert User Group and all affiliated members for their interests, support and collaboration.

REFERENCES

[1] K. West, S. Cox, "Features and classifier for the automatic classification of musical audio signals", in Proc. Of the 5th Int. Conf. on Music Information Retrieval (ISMIR), Barcelona, Spain, 2004.

[2] G. Tzanetakis, P. Cook, "Musical genre classification of audio signals", in IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 5, July 2002.

[1] [1] http://www.music-ir.org/mirexwiki/index.php/MIREX_2005

[3] N. Scaringella, G. Zoia, "On the modelling of time-information for automatic genre recognition systems in audio signals", in Proc. Of the 6th Int. Conf, on Music Information Retrieval, London, UK, 2005.

[4] L. Rabiner, B.H. Juang, "Fundamentals of speech recognition", Englewood Cliffs, NJ, Prentice-Hall, 1993.

[5] J.J. Aucouturier, F. Pachet, "Improving timbre similarity: how high's the sky?", in Journal of Negative Results in Speech and Audio Sciences, 2004.

[6] A. Klapuri, A. Eronen, J. Astola, "Analysis of the meter of acoustic musical signals", in IEEE Trans. Speech and Audio Proc., 2004.

[7] C.J.C. Burges, "A tutorial on Support Vector Machines for pattern recognition", in Data Mining and Knowledge Discovery, 2:121-167, 1998.

[8] E. Allwein, R. Schapire, Y. Singer, "Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers", in Journal of Machine Learning Research 1, pp. 113-141, 2000.

[9] R. Collobert, S. Bengio, Y. Bengio, "A Parallel Mixture of SVMs for Very Large Scale Problems", in Neural Computation, Vol. 14, No.5, pp. 1105-1114(10), May 2002.

[10] C. Costa, J. Valle Jr., A. Koerich, "Automatic classification of audio", in IEEE Int. Conf. on Systems, Man, Cybernetics, The Hague, Netherlands 2004.

Truth / Prediction	C	ED	NA	RH	RE	RO
C	**94.05**	4.48	0.00	2.56	0.00	18.56
ED	1.19	**52.24**	0.00	5.98	11.11	4.19
NA	0.00	10.45	**95.24**	0.00	0.00	2.40
RH	0.00	10.45	0.00	**84.62**	5.56	4.79
RE	1.19	4.48	0.00	2.56	**72.22**	2.40
RO	3.57	17.91	4.76	4.27	11.11	**67.66**

Table 1. Confusion matrix for the USPOP dataset (C: Country – ED: Electronica & Dance – NA: New Age – RH: Rap & Hip-Hop – RE: Reggae – RO: Rock)

Truth / Prediction	A	B	C	EL	ET	F	J	NA	P	R
A	**26.47**	0.00	0.00	2.44	1.20	0.00	0.00	0.00	0.00	0.00
B	2.94	**100,00**	0.00	12.20	3.61	0.00	4.55	5.88	0.00	9.52
C	5.88	0.00	**100.00**	2.44	7.23	0.00	0.00	5.88	0.00	1.19
EL	14.71	0.00	0.00	**42.68**	7.23	8.33	4.55	23.53	2.94	9.52
ET	17.65	0.00	0.00	6.10	**59.04**	8.33	9.09	11.76	0.00	1.19
F	0.00	0.00	0.00	0.00	0.00	**70.83**	0.00	0.00	0.00	1.19
J	2.94	0.00	0.00	2.44	4.82	0.00	**72.73**	0.00	0.00	1.19
NA	11.76	0.00	0.00	3.66	6.02	4.17	0.00	**44.12**	0.00	2.38
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**76.47**	3.57
R	17.65	0.00	0.00	28.05	10.84	8.33	9.09	8.82	20.59	**70.24**

Table 2. Confusion matrix for the MAGNATUNE dataset (A: Ambient – B: Blues – C: Classical – EL: Electronic – ET: Ethnic – F: Folk – J: Jazz – NA: New-Age – P: Punk – R: Rock)