Probabilistic Combination of Features for Music Classification

Arthur Flexer[1], **Fabien Gouyon**[2], **Simon Dixon**[2], **Gerhard Widmer**[2,3]
[1]Institute of Medical Cybernetics and Artificial Intelligence
Center for Brain Research, Medical University of Vienna, Austria
[2]Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria
[3] Department of Computational Perception
Johannes Kepler University, Linz, Austria
arthur.flexer@meduniwien.ac.at, fabien.gouyon@ofai.at,
simon.dixon@ofai.at, gerhard.widmer@jku.at

Abstract

We describe an approach to the combination of music similarity feature spaces in the context of music classification. The approach is based on taking the product of posterior probabilities obtained from separate classifiers for the different feature spaces. This allows for a different influence of the classifiers per song and an overall classification accuracy improving those resulting from individual feature spaces alone. This is demonstrated by combining spectral and rhythmic similarity for classification of ballroom dance music.

Keywords: music classification, combination

1. Introduction

Since the perceived similarity between pieces of music is defined by a whole range of different aspects (timbre, rhythm, harmony, melody, socio-cultural, etc) it is only logical that any attempt at music classification should be based on a combination of these different dimensions of similarity. In the scientific field of statistical pattern recognition, there exist clear results as to how to achieve such a combination (see e.g. [7]). Bayesian theory tells us that all available information (i.e. features derived from different aspects of music similarity) should be considered simultaneously by using one overall classifier. However, this is very often not practical or even possible (due to exponential growth of the number of the parameters, different time scales or general incomparability of feature spaces, etc). The alternative then is to combine information from different sources which leads to the question of how to weigh information from these sources to reach an overall decision. In a probabilistic setting, the preferred approach is to train separate classifiers for the different feature spaces and then to combine posterior probabilities to obtain a joint decision. This allows for a different influence of the classifiers per song based on their posterior probabilities. Thereby different aspects of music similarity achieve different weights in the joint classification decision for every song.

Although this combination approach is well known in statistical pattern recognition, something related has barely been used within the Music Information Retrieval community so far [16, 14]. We therefore think it is beneficial to further explore this probabilistic approach to the combination of features for music classification by (i) reviewing the necessary theory, (ii) presenting experimental results on the combination of spectral and rhythmic similarity for classification of ballroom dance music.

2. Data

The musical data set used for training and testing contains excerpts from S=698 pieces of music, around 30 seconds long, amounting to around 20940 seconds of data. This data was originally downloaded from a web site [1] providing diverse resources related to ballroom dancing (online lessons, videos, books, etc.). Some characteristic excerpts of many dance styles are provided in the low quality Real Audio sound format (with a compression factor of almost 22 with respect to the common 44.1 kHz 16 bits mono WAV format), labelled with a specific dance style (G=8 music genres: Cha Cha Cha (111 pieces), Jive (60), Quickstep (82), Rumba (98), Samba (86), Tango (86), Viennese Waltz (65), Slow Waltz (110)). The data was subsequently converted to WAV format for experiments.

This data was first used in [6, 4]. It was also used in the *Tempo induction* and *Rhythm classification* contests organised during ISMIR 2004. The audio data and style and tempo annotations are publicly available. [2]

3. Music similarity

One concept of central importance in music information retrieval is the notion of musical similarity. Similarity metrics define the inherent structure of a music collection, and the acceptance of a music retrieval system crucially depends on

[1] http://www.ballroomdancers.com/
[2] http://ismir2004.ismir.net/

whether the user can recognise some similarity between the query and the retrieved sound files. Since usually no ground truth with respect to music similarity exists, genre classification is widely used for evaluation of music similarity. Each song is labelled as belonging to a music genre using e.g. music expert advice. High genre classification results indicate good similarity measures. Genre classification is also a goal in its own right: it allows labelling of a user's collection of music based on a subset of training songs with user defined genre labels.

Some of the most successful approaches to genre classification are based on the use of spectral features (see e.g. [10, 1, 15] and many more), rhythm [15, 6, 13] or cultural features [12, 16, 8]. It seems clear that combination of different aspects of music similarity is able to improve the performance achieved so far. Our experiments are on the combination of spectral and rhythmic similarity.

3.1. Spectral similarity

The following approach to music similarity based on spectral similarity (see [10, 1] for early references) is now seen as one of the standard approaches in the field of music information retrieval. For a given music collection of S songs, each belonging to one of G music genres, it consists of the following steps: (i) for each song, compute Mel Frequency Cepstrum Coefficients (MFCCs) for short overlapping frames; (ii) train a Gaussian Mixture Model (GMM) for each of the songs; (iii) compute an $S \times S$ distance matrix between all songs using the likelihood of a song given a GMM.

We divide the raw audio data into overlapping frames of short duration and use Mel Frequency Cepstrum Coefficients (MFCC) to represent the spectrum of each frame (see e.g. [9]). The frame size for computation of MFCCs for our experiments was $23.3ms$ (1024 samples), with a hop-size of $11.6ms$ (512 samples) for the overlap of frames. We used the first 8 MFCCs for all our experiments.

A Gaussian Mixture Model (GMM) models the density of the input data by a mixture model of the form

$$p(x) = \sum_{m=1}^{M} P_m \mathcal{N}[x, \mu_m, U_m] \qquad (1)$$

where P_m is the mixture coefficient for the m-th component, $\mathcal{N}$ is the Normal density and μ_m and U_m are the mean vector and covariance matrix of the m-th mixture. For a data set X^i containing T data points given a GMM trained on song j, the negative log-likelihood function is given by

$$L(X^i|GMM_j) = -\frac{1}{T}\sum_{t=1}^{T} log(p_j(x_t^i)) \qquad (2)$$

For learning a GMM for a song i, $L(X^i|GMM_i)$ is minimised both with respect to the mixing coefficients P_m and with respect to the parameters of the Gaussian basis functions using Expectation-Maximisation (see e.g. [2]). For all our experiments we used $M = 10$ components and diagonal covariances. Computation of $L(X^i|GMM_j)$ for all possible combinations of songs i and GMMs j gives an $S \times S$ distance matrix D_S.

3.2. Rhythmic similarity

There are many ways to compute rhythmic similarity, e.g. [15, 6, 13]. In this paper, the definition of rhythmic similarity focuses on a single rhythmic dimension: the tempo.

Tempo is a musical attribute of prime importance and recent research showed, on the data used here, the high relevance of tempo for ballroom dance music classification [6, 4]. Common musical knowledge (e.g. instructional books, dance class websites) suggests that tempo is a fundamental feature in the definition of musical styles, and on the other hand, [11] shows on a large amount of data (more than 90000 instances) that different dance music styles ("trance, afro-american, house and fast") show clearly different tempo distributions, centred around different "typical" tempi.

The tempo induction algorithm used is one of the algorithms that entered the MIREX 2005 competition on *Perceptual tempo induction*. It is referred to as `Algorithm1` in [5] and consists of the following processing steps: (i) framing of the signal; (ii) computation of the magnitude-normalised derivative of the energy in 8 frequency bands; (iii) computation of the autocorrelation in each band; (iv) parsing of periodicity function peaks and global tempo computation as in [3]. This algorithm yields one single tempo feature t_i for each song i. The distance $d_R(i, j)$ between two songs i and j is computed as the Euclidean distance $(t_i - t_j)^2$. Computation of all possible combinations of songs i and j gives an $S \times S$ distance matrix D_R.

4. Combination

Following earlier work on combination of classifiers [7], let us consider a pattern recognition problem where a pattern Z is to be assigned to one of G classes $(\omega_1, ..., \omega_G)$. Let us further assume we have R classifiers each receiving distinct measurement vectors x_i (e.g. one classifier trained on spectral similarity, another one on rhythmic similarity). In measurement space x_i each class ω_k is modelled by the probability density function $p(x_i|\omega_k)$, with $P(\omega_k)$ being the corresponding prior probability. According to Bayesian theory a pattern Z with measurements $x_i, i = 1, ..., R$, should be assigned to class ω_j provided that the corresponding posterior probability is maximal:

$$P(\omega_j|x_1, ..., x_R) = \max_k P(\omega_k|x_1, ..., x_R) \qquad (3)$$

This means that to utilise all available information, all available measurements should be considered simultaneously. However, very often this is not practical or even not feasible. Simultaneous use of all measurements can lead to very

large feature spaces which are hard to model due to the curse of dimensionality (i.e. exponential growth of the number of model parameters with number of features, see e.g. [2]). Often subsets of the measurements are hard to compare due to their different origin and it is unclear how to weight or normalise them. Sometimes they exist on different time scales (e.g. one spectral feature vector every $11.6ms$ compared to one single rhythmic feature for a whole song) and cannot be concatenated at all. Therefore a promising approach is to use individual classifiers for subsets of the measurements and to combine the classifier outcomes instead.

Assuming that the measurements $x_i, i = 1, ..., R$ are conditionally statistically independent, we can rewrite the decision rule given by Eqn. 3 to:

$$P(\omega_j|x_1, ..., x_R) = \max_k \prod_{i=1}^{R} P(\omega_k|x_i) \qquad (4)$$

This means we can express the posterior probability given the joint measurements as the product of the posteriors computed in the individual measurement spaces. This is known as the product rule [7] and for many applications the above independence assumption provides an adequate and practicable approximation to a reality that might be more complex. It is important to note that the product rule provides a different influence of the classifiers per song since the weight given to each of the classifiers changes with the posteriors obtained for each song. If a classifier is very insecure about its decision all posterior probabilities will be at the same level and not influence the other classifier posteriors at all. If the decision of a classifier is very clear, the corresponding posterior probability will be very high and dominate the product of the posteriors.

A simple way to directly estimate posterior probabilities is to use K-nearest neighbour classification [2]. The numbers of songs belonging to genres $(\omega_1, ..., \omega_G)$ in the set of the K nearest neighbours are an approximation of the true posterior probabilities. To avoid zero probabilities we added a pseudo-count to all numbers of songs belonging to genres.

A related approach to combination of music similarity features has been reported in [16]. Contrary to the above described approach no proper posterior probabilities were used and therefore, following a suggestion by [7], the average instead of the product was computed for combinations of estimated posteriors. The data set in this study was rather small (25 artists) and the goal was classification of artists, not songs, into five genres. Combination of audio and community meta-data improved the results achieved on the individual feature spaces. Not directly related is an approach at the symbolic level on combining predictive models of sequential pitch structure in melodic music [14].

Our own previous efforts on combination [13] relied on a linear combination of distance matrices which were computed for the individual feature spaces. This new combined distance matrix was then used as input for one overall classifier. Contrary to our new approach this requires one overall weighting for all songs in a training set. It is also not clear how to obtain the specific weighting which is optimal for a specific data set. This would require a meta-search strategy evaluating all different possible weights for the linear combination and a final evaluation of the winner using previously unseen test data to allow for fair performance evaluation. Nevertheless, linear combination of distance matrices is also applicable to problems outside of music classification since no class labels are needed to compute the combination weights.

5. Results

We computed distance matrices D_S and D_R for all $S = 698$ pieces of music in our data base as described in Secs 3.1 and 3.2. We also computed a combined distance matrix D_C using a linear combination: $D_C = (D_S + D_R)/2$. Since we have no information as to which weighting to prefer we decided to use this simple average. All distances in D_S and D_R were normalised to zero mean and unit standard deviation before this combination to guarantee comparability as suggested in [13].

In a 10-fold cross validation we did the following experiments:

1. do K=10-nearest neighbour classification for all songs in the test fold using distance matrices D_S, D_R and D_C to compute respective posterior probabilities; assign each song to the class with maximal posterior probability

2. use the product rule to combine posterior probabilities obtained from K=10-nearest neighbour classification using D_S and D_R separately; assign each song to the class with maximal posterior probability

Average classification accuracies plus standard deviations for the four different methods are given in Table 1. As can be clearly seen, the combination based on the product rule is able to enhance the performance considerably when compared with classification based on spectral or rhythmic similarity alone. Both results are statistically significant when using paired t-tests: spectral vs. product $t = |-16.10|$, rhythm vs. product $t = |-4.61| > t_{(99,df=9)} = 3.25$. The product rule also outperforms the linear combination ($t = |-5.55| > t_{(99,df=9)} = 3.25$), which even falls behind the classification based on rhythmic similarity alone.

spectral	rhythm	linear	product
33.39 ± 6.16	58.59 ± 5.34	48.67 ± 7.77	66.89 ± 5.26

Table 1. Mean accuracies and standard deviations for the four methods.

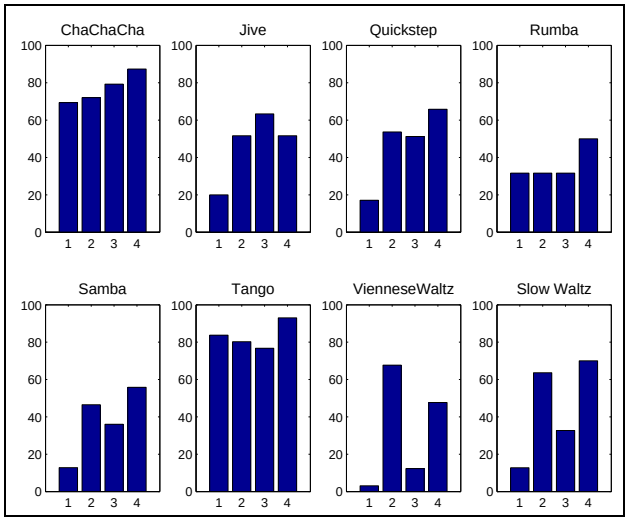

Figure 1. Mean percentages of correctly classified songs per Genre given for the four methods: (1) spectral, (2) rhythm, (3) linear combination, (4) posterior product combination.

The mean percentages of correctly classified songs per genre are given in Fig. 1. Whereas combination based on the product rule improves performance for every genre except 'Viennese Waltz' and 'Jive', linear combination increases performance for only two genres ('ChaChaCha', 'Jive') but decreases it for five of the remaining six.

We also tried different weights for the linear combination of D_S and D_R. Changing the weights in steps of 0.1 gave best performance of 59.97 ± 9.63 for the combination $D_C = (0.1 * D_S + 0.9 * D_R)$. Although this post-hoc choice of an optimal classifier is not statistically sound and should lead to over-optimistic performance estimates, the result is still just at the level of classification based on rhythmic similarity alone. Combination of posterior probabilities using the average instead of the product as done by [16] yielded an average accuracy of 63.17 ± 5.77. This is better than the results based on spectral or rhythmic similarity alone as well as based on the linear combination. But it does not reach the product rule's performance (product vs. average $t = |5.01| > t_{(99,df=9)} = 3.25$).

6. Conclusion

We presented a general framework for combination of music similarity feature spaces in the context of music classification. Combination of separate classifiers trained on the individual feature spaces allows for a different weighting of the separate classifiers per song which results in a considerable increase in genre classification accuracy. This probabilistic approach to combination of classifiers is well known in statistical pattern recognition and our results obtained for a combination of spectral and rhythmic similarity applied to ballroom dance music confirm its applicability in the field of Music Information Retrieval.

7. Acknowledgments

Parts of the MA Toolbox [3] and the Netlab Toolbox [4] have been used for this work. This research was supported by the EU project S2S2, [5] and the WWTF project CI010 *Interfaces to Music*. OFAI acknowledges support from the ministries BMBWK and BMVIT.

References

[1] Aucouturier J.-J., Pachet F.: Music Similarity Measures: What's the Use?, in Proc. of the 3rd Int. Conf. on Music Information Retrieval, pp. 157-163, 2002.

[2] Bishop C.M.: Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 1995.

[3] Dixon S., Pampalk E., Widmer G.: Classification of dance music by periodicity patterns, in Proc. 4th Int. Conf. on Music Information Retrieval, 2003.

[4] Gouyon F., Dixon S.: Dance music classification: A tempo-based approach, in Proc. of the 5th Int. Conf. on Music Information Retrieval, 2004.

[5] Gouyon F., Dixon, S.: Influence of input features in perceptual tempo induction, in 2nd Annual Music Information Retrieval eXchange (MIREX), 2005.

[6] Gouyon F., Dixon S., Pampalk E., Widmer G.: Evaluating rhythmic descriptors for musical genre classification, in Proc. 25th International AES Conference, pp.196-204, 2004.

[7] Kittler J., Hatef M., Duin R.P.W., Matas J.: On combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3), 1998.

[8] Knees P., Pampalk E., Widmer G.: Artist Classification with Web-based Data, Proc. of the 5th Int. Conf. on Music Information Retrieval, 2004.

[9] Logan B.: Mel Frequency Cepstral Coefficients for Music Modeling, Proc. of the International Symposium on Music Information Retrieval, 2000.

[10] Logan B., Salomon A.: A music similarity function based on signal analysis, IEEE International Conf. on Multimedia and Expo, Tokyo, Japan, 2001.

[11] Moelants D.: Dance music, movement and tempo preferences, in Proc. 5th Triennal ESCOM Conference, 2003.

[12] Pachet F., Westermann G., Laigre D.: Musical Data Mining for Electronic Music Distribution, Proc. of the first WedelMusic conference, 2001.

[13] Pampalk E., Flexer A., Widmer G.: Improvements of Audio-Based Music Similarity and Genre Classification, Proc. of the 6th Int. Conf. on Music Information Retrieval, 2005.

[14] Pearce M.T., Conklin D., Wiggins G.A.: Methods for combining statistical models of music, in Wiil U.K. (Ed.), Computer music modelling and retrieval, (pp. 295-312), Heidelberg, Germany, Springer Verlag, 2005.

[15] Tzanetakis G., Cook P.: Musical genre classification of audio signals, IEEE Transactions on Speech and Audio Processing, Vol. 10, Issue 5, 293-302, 2002.

[16] Whitman B., Smaragdis P.: Combining Musical and Cultural Features for Intelligent Style Detection, Proc. of the 3rd Int. Conf. on Music Information Retrieval, 2002.

[3] `http://www.ofai.at/~elias.pampalk/ma`
[4] `http://www.ncrg.aston.ac.uk/netlab`
[5] `http://www.s2s2.org`