# Time-Space Ensemble Strategies for Automatic Music Genre Classification

Carlos N. Silla Jr., Celso A. A. Kaestner, and Alessandro L. Koerich

Pontifical Catholic University of Paraná (PUCPR)
Postgraduate Programme in Computer Science (PPGIA)
Rua Imaculada Conceição 1155, 80215-901 Curitiba, Brazil
{silla,kaestner,alekoe}@ppgia.pucpr.br

**Abstract.** In this paper we propose a novel time–space ensemble–based approach for the task of automatic music genre classification. Ensemble strategies employ several classifiers to different views of the problem–space, and combination rules in order to produce the final classification decision. In our approach we employ audio signal segmentation in time intervals and also problem space decomposition. Initially the music signal is split in time segments; features are extracted from these music signal segments and the one against all (OAA) and round robin (RR) strategies, which implement a space decomposition by using several binary classifiers, are applied. Finally, the outputs of the set of classifiers are combined to produce the final result. We test our proposition in a music database of 1.200 music samples from four different music genres. Experimental results show that time segment decomposition is more important than the space decomposition produced by the OAA and RR strategies, although they produce better results relative to the use of single classifiers and feature vectors.

## 1 Introduction

The large amount of multimedia information on the web surface nowadays makes it necessary to build a new class of automatic tools, capable of dealing with information from very different media [1]. In this context one of the most important tasks is the automatic content-based music genre classification. The music genre is one of the most important aspects to describe music, and it is mainly used to organize large collections of digital music [2].

From a pattern recognition perspective, music genre classification poses an interesting research problem, since music is a complex time-variant sinal. Another interesting aspect is that genre classification is naturally a multi-class problem. In order to deal with multi-class problems there are two basic possibilities: the first one is to use learning and classification techniques that can naturally handle multi-class problems – producing complex decision surfaces – like decision trees, $k$ nearest neighbors (k-NN), neural networks, etc.; the second option is to use a problem space decomposition strategy to break a multi-class problem into a series of binary problems that can be tackled using a set of binary classifiers – which produce simple decision surfaces – like support vector machines (SVM).

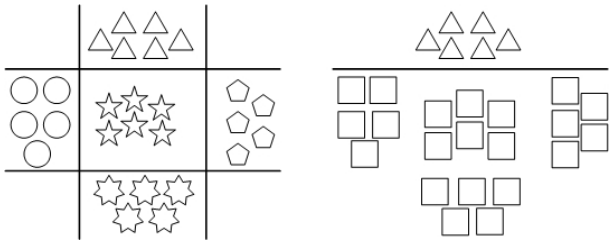

**Fig. 1.** An example illustrating the problem space decomposition strategy

In several pattern recognition areas [3] [4] the so called "ensemble approach" has been used with success. This approach consists of applying to the problem not a single classifier, but a collection of them, each one specialized in a specific view of the problem. In this way, each classifier is trained on different distributions, and the outputs of predictors are combined by a dynamic classifier combination model. This procedure may be viewed as either a version of mixture of experts [5] applied to classification, or a variant of the boosting algorithm [6]. A possible explanation for the success of the ensemble approach is that classifiers applied to partial views of the problem space produce simpler decision surfaces, and therefore, better classification results. Figure 1 shows an example of such an approach.

Fürnkranz [7] suggests that the problem space decomposition strategy can be used as an ensemble technique for any classifier (binary or not). A common problem space decomposition strategy is the one against all (OAA), where a classifier is created to recognize the set of patterns that belongs to one specific class. A second problem space decomposition strategy is the round robin (RR) [8] (a.k.a. pairwise comparison [7]) where a set of classifiers is created for every possible combination of two classes. A third possible approach is the random subspace method (RSM) [9], where classifiers are applied to a set of random selected projections of the problem space. In all cases the individual classification results are combined to produce the final classification.

The idea of developing an ensemble strategy based on problem space decomposition for the task of music genre recognition was introduced by Grimaldi et al. [10, 11]. They evaluate the performance of different ensemble methods (OAA, RR, RSM). The experiments were performed using features based on the discrete wavelet packet transform (DWPT), which where extracted from the whole music signal. A major limitation in their work is that the efficiency of the ensemble methods were not evaluated alone, since the presented results also include feature selection. Li et al. [8] present an analysis of different classifiers applied to the music genre recognition task. In this work they employ the OAA and the RR problem space decomposition techniques using a set of SVM classifiers. The experimental results show that the OAA strategy achieves slightly better results than the RR strategy. One limitation of this work is that the decomposition strategies were evaluated only with SVM classifiers.

Costa et al. [12] exploit a different way of using the ensemble approach for music classification. In their work each classifier of the ensemble was trained and applied to different time intervals – or segments – of the music, taking into account the temporal nature of the music signal. They separate three segments from the beginning, middle and end of the music signal, in order to produce a set of feature vectors and obtain the corresponding classifications. Music classification problems usually employ the beginning or the whole music signal in order to deal with classification [1], [11]. Costa et al. justify their proposition because a music signal can strongly vary in the time dimension – for example a rock music can start as a classical music, so it seems to be a good idea to treat different segments and combine the obtained classification results. In their work the k-NN classifier and multilayer perceptron (MLP) neural networks classifiers were used in each segment.

In this paper we apply the ensemble approach for automatic music genre classification in a new way. We consider the time dimension of the music signal and also space decomposition by using OAA and RR strategies for sets of binary classifiers, attempting to exploit the advantages offered by both strategies. The performance of the proposed approach is compared with other ensemble methods. We also employ a broader range of classifiers than the ones used in the previous works. This paper is organized as follows: Section 2 presents the foundations of the ensemble methods. Section 3 describes the target music genre classification task. Section 4 presents the experiments carried out and an analysis of the achieved results. Finally the last section presents our conclusions and concluding remarks.

## 2 Ensemble Methods

An "ensemble approach" to a pattern recognition problem consists of a decomposition of the original problem space by using a collection of classifiers, each of them dedicated to a specific view of this space. In a first step each classifier – usually a binary one – is trained and applied to its view, producing an individual classification; in a second step these classification results are dynamically combined to produce the final classification.

The main motivation to apply problem-space decomposition using methods like OAA and RR is that multi-class classification is intrinsically harder than binary classification, because the classification algorithm has to construct a high number of separation boundaries, whereas binary classifiers have to determine only one appropriate decision function [13].

### 2.1 One Against All (OAA) Strategy

Given a $n$-class pattern recognition problem, the OAA strategy consists of creating a set of $n$ binary classifiers, one for each class. Each classifier is trained through re-labeling of the same training dataset, in order to distinguishing between one single class and its complement in the problem space. For instance,

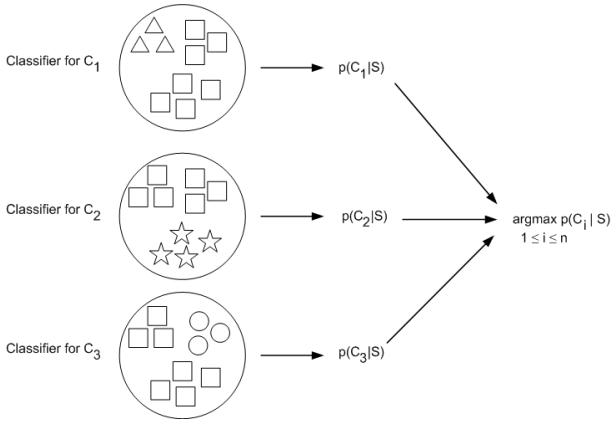


**Fig. 2.** Illustration of the OAA strategy for a three class problem

the classifier for class $C_i$ is trained using the elements of $C_i$ as positive examples and the remaining of the data-set as negative examples, producing a specialized classifier for class $C_i$. For an unseen example represented by a feature vector $s$, given the $n$ individual classifications, and considering that each individual classifier assigns to $s$ a probability $p$ (or a confidence score) that is directly related to the conformity of this example with its class, the final class $\hat{C}$ assigned to $s$ is given, as usual, by:

$$\hat{C} = C_k|k = \arg\max_{1 \leq i \leq n} p(C_i|s) \tag{1}$$

where $p(C_i|s)$ is the a posteriori probability of class $C_i$ given a feature vector $s$ and $\hat{C}$ is the winner class, that is, the one which provides the highest a posteriori probability. Figure 2 illustrates this approach.

### 2.2 Round Robin Strategy

Fürnkranz [7] presents the RR problem space decomposition as an ensemble strategy, in order to allow binary classifiers to deal with multi-class problems. The RR method converts a $n$-class problem into a series of binary problems, by creating a set of $k = n(n-1)/2$ classifiers, one for each pair of classes. Unseen samples are classified by presenting them to the set of $k$ binary classifiers. In this case when an unseen example $e$ is presented to each one of the $k$ binary classifiers, a class is directly assigned to $e$. The $k$ assigned classes are finally combined into the final result through majority voting.

Contrary to the OAA ensemble strategy, in this case when a binary classifier is constructed, let's say for classes $C_i$ and $C_j$, only the examples of these two classes are used, and the rest of the dataset is ignored. According to Fürnkranz [7], this leads to an easier decision about the boundary between the two classes. Figure 3 illustrates this approach.

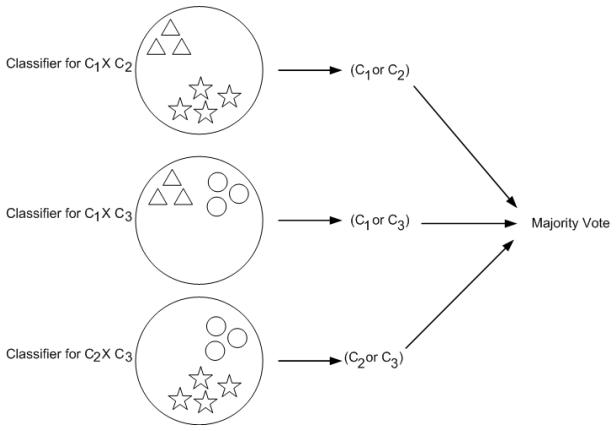


**Fig. 3.** Illustration of the RR strategy for a three class problem

### 2.3 Segment–Based Ensemble Strategy

A different ensemble strategy was proposed by Costa et al. [12] for the task of automatic music genre classification. The proposed approach can be easily extended to other time-variant signals or to time-dependent classification tasks. The music signal $m$ is split into time intervals or segments, and features are extracted separately from each segment. The same features are extracted from each segment and three classifiers are trained.

However, music is a time-variant signal, and the decomposition is made according to the time dimension, producing different views of the same object. In this case, as already noted by Kittler et al. [3], it is possible to use alternative ensemble methods. When a new unseen music $m'$ is presented, the corresponding temporal segment are extracted, producing three different views of $m'$. The specific classifier of each segment is then applied, and the final classification decision (in this case, the music genre) is carried out through the majority voting principle.

### 2.4 A Time–Space Ensemble Approach

It is possible to combine both ensemble techniques described previously in order to perform a new ensemble-based approach to music genre classification. In the first (time decomposition) step the music signal is segmented according to a set of time intervals. Features are then extracted from these segments and used in a second ensemble decomposition (space decomposition) using OAA and RR strategies and a set of binary classifiers. Finally, a compositional rule is employed, taking into account the individual classification decisions (for both time and space decompositions) to provide the final class label. Figure 4 illustrates the proposed approach.

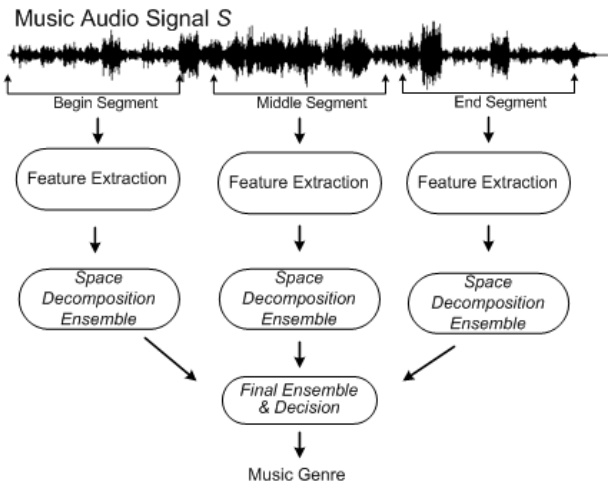


**Fig. 4.** The main components of the time–space ensemble–based approach

## 3 Music Genre Classification

The problem of music genre classification can be formally defined as the task of assigning a boolean value $T$ ($true$) or $F$ ($false$) to each pair $\langle m, g\rangle \in \{\mathcal{M} \times \mathcal{G}\}$, where $\mathcal{M}$ is a domain of music (signals) and $\mathcal{G}$ is a predefined set of music genres. A value $T$ assigned to $< m, g >$ indicates that $m$ belongs to the genre $g$, while $F$ indicates that $m$ does not belong to $g$. In a classification process we try to approximate an unknown *target function* $\Phi^{real} : \mathcal{M} \times \mathcal{G} \rightarrow \{T, F\}$ by means of a function $\Phi : \mathcal{M} \times \mathcal{G} \rightarrow \{T, F\}$ called the classifier (a.k.a. rule, or model) such as $\Phi^{real}$ and $\Phi$ "coincide as much as possible" [14].

In our case, the final classifier $\Phi$ is not obtained directly from the whole music signal, but from the music segments and a set of classifiers, according to the OOA or the RR strategy. In each segment and binary space decomposition we employ conventional algorithms.

We split each music signal into three 30 seconds segments – which is often used in the literature [1], [8] – from the beginning, middle and end of the music signal. If a music signal is composed by $f$ frames, we consider the beginning of the music the frames from 0 to 1.153 (corresponding in a MP3 file to 30 seconds), the middle segment as the frames from $(\frac{f}{3} + 500)$ to $(\frac{f}{3} + 1.453)$ and the final segment as the frames $(f - 1.453)$ to $(f - 300)$. During the segmentation stage each segment from the audio signal is extracted after the conversion of the MP3 File to 16 bits mono Wav.

The feature extraction from the music segments is carried out through the Marsyas framework [15], which implements the original feature set proposed by Tzanetakis & Cook [1]. The features used in this work can be divided into three groups: timbral texture, beat related and pitch related. The features based on the timbral texture account for the mean and variance of the spectral centroid,

rolloff, flux, the time zero domain crossings, the first 5 MFCC's and low energy. Features that are beat related include the relative amplitudes and the beats per minute. Pitch related features include the maximum periods of the pitch peak obtained from the pitch histograms. All this features are concatenated to form a 30-dimensional feature vector. More details about these features can be found in [1].

We use the following machine learning algorithms as component classifiers for the ensemble methods: Decision Trees (DT), k-NN, Naïve Bayes (NB), Support Vector Machines (SVM) with pairwise classification and an MLP neural network classifier trained with the backpropagation momentum algorithm [16]. The employed classification framework is based on the Weka Datamining Tool [17].

The final classification result is obtained from these partial classifications, by means of a combination rule. In our case the partial results are combined by using the majority vote rule. It is important to notice that we have not mixed different component classifiers into the ensembles. The ensembles are made up by homogeneous classifiers.

## 4 Experimental Results

The main objective of our experiments is to evaluate the performance of the different ensemble methods for the task of automatic music genre classification, and if it is advantageous or not to combine time segmentation and problem space decomposition into a unique ensemble approach to deal with this classification problem.

A database containing music samples from four different Latin genres was available for the experiments. We have selected 300 samples from each genre (Tango, Salsa, Forró, and Axé) and split them into training (50%), validation (20%) and test (30%) sets according to a classical holdout procedure. Due to time constraints cross-validation was not employed. The music samples have been selected randomly from a large database without reposition to avoid bias in the experiments.

In order to have a baseline (BL) we have also used the classifiers in a conventional manner – so, with no space decomposition – to each music segment. Also it is important to note that since the baseline SVM classifier uses the RR decomposing strategy, its results in both columns are the same.

Table 1 presents the accuracy results achieved by the different classifiers on individual segments. It seems that for the beginning segment the best classification accuracy is often achieved by the RR ensemble strategy. This is not true only for the DT classifier which presents the best result without using any ensemble technique and the NB classifier which presents the best result using the OAA ensemble. For the 1-NN the result is the same regardless of the strategy employed.

For the middle segment, the RR ensemble often presents the best classification accuracy. Similar to the beginning segment, the DT holds better results

**Table 1.** Accuracy (%) using different strategies on individual music segments

| | Begin | | | Middle | | | End | | |
|---|---|---|---|---|---|---|---|---|---|
| Classifier | BL | OAA | RR | BL | OAA | RR | BL | OAA | RR |
| DT | **67.22** | 63.05 | 64.72 | **68.61** | 65.27 | 66.94 | 50.00 | 58.88 | **67.22** |
| 1-NN | 68.33 | 68.33 | 68.33 | 73.61 | 73.61 | 73.61 | 68.05 | 68.05 | 68.05 |
| 3-NN | 67.22 | 67.22 | **68.05** | 75.83 | 75.83 | **76.66** | 71.66 | 71.66 | **73.05** |
| 5-NN | 68.33 | 68.33 | **70.27** | 76.38 | 76.38 | **76.66** | **70.27** | **70.27** | 69.72 |
| 7-NN | 70.83 | 70.83 | **71.94** | 74.44 | 74.44 | **75.83** | **73.05** | **73.05** | 70.27 |
| MLP | 76.94 | 80.55 | **83.33** | 80.83 | **85.27** | 77.77 | 66.11 | **68.33** | 59.72 |
| NB | 69.16 | **70.27** | 69.16 | 76.94 | 76.94 | 76.94 | 61.94 | **62.50** | 61.66 |
| SVM | **81.11** | 66.38 | **81.11** | **86.66** | 74.44 | **86.66** | **70.00** | 60.27 | **70.00** |

using only a BL classifier. For the 1-NN and the NB classifier, regardless of the strategy adopted, the results are similar. Also for the MLP classifier, the best accuracy was achieved using the OAA ensemble.

For the segment that represents the end part of the music, the OAA ensemble often presents the best results. This is true for the MLP and the NB classifier and also for the 1-NN, 5-NN and 7-NN classifier. However, the results achieved using the OAA ensemble of k-NN classifiers are similar to the results achieved by the BL approach. The RR ensemble outperformed the OAA only for the DT and 3-NN classifier.

In summary, for the initial segment the best result was achieved using the MLP with RR (83.33%); for the middle segment the best result was achieved using the SVM with RR (86.66%); for the end segment the best results were achieved using the 3-NN with RR, and 7-NN with OAA (73.05%).

It is also possible to combine the decisions of the ensemble of classifiers specialized in each segment of the music and build a time-space ensemble approach based on majority voting which takes into account the final class label provided by the ensemble performed on each music segment. The accuracy results achieved by this approach are presented in Table 2.

Table 2 shows that the ensemble of the DT individual results is the same as for the for the RR ensemble. For the k-NN classifiers using only the individual segments often achieves better results than using the OAA or RR ensembles. This is not true only for the 5-NN, where the RR time-space ensemble performs slightly better. For the MLP and the NB classifier the time-space OAA ensemble strategy provides better results than using the BL classifiers or the RR ensemble. For the SVM classifier the results achieved are the same for the BL and the RR ensemble, because as mentioned before, the decomposition strategy used for handling multi-class problems in the SVM was the RR.

When evaluating the results achieved by the ensembles and comparing them with the results achieved by the classifiers on the individual segments, we see that the majority vote ensemble provides better accuracy in any case only for the DT and k-NN. For the MLP classifier, the ensemble of BL classifiers pro-

**Table 2.** Accuracy (%) using the majority vote rule for segment ensemble

| | 3 Segment Majority Vote | | |
|---|---|---|---|
| Classifier | BL | OAA | RR |
| DT | **75.27** | 73.61 | **75.27** |
| 1-NN | **78.33** | **78.33** | 78.05 |
| 3-NN | **81.38** | 80.55 | 81.11 |
| 5-NN | 81.38 | 81.11 | **81.94** |
| 7-NN | **82.50** | 81.94 | 80.83 |
| MLP | 81.38 | **83.33** | 79.72 |
| NB | 72.50 | **76.94** | 72.22 |
| SVM | **86.11** | 77.50 | **86.11** |

vides better results than any of the individual segment/classifier. This does not happen for the time-space ensembles which present better results using only one of the individual segments. The majority vote ensemble does not improve the performance of the NB classifier as the best accuracy is achieved with the middle segments. For the SVM classifier the accuracy using the middle segment outperforms the performance of the time-space ensembles.

## 5 Concluding Remarks

In this work we propose a new approach to the task of automatic music genre classification based on time-space ensemble decomposition. Time decomposition is achieved by breaking up the music signal in several temporal segments. The proposed approach uses three feature vectors extracted from the beginning, middle and end parts of the music. This procedure tries to assure that the most important temporal patterns of the music signal to be considered in the classification. Space decomposition is obtained by applying sets of binary classifiers to a naturally multi-class problem. This procedure tends to produce simpler separation surfaces in specific views of the problem space. We employ two different space-decomposition ensemble strategies, namely One-Against-All (OAA) and Round Robin (RR). Final results are obtained by simple composition rules. We use the classical DT, Naïve Bayes, k-NN and SVM as classification algorithms.

The experiments were performed using the classifiers and a large dataset composed by 1.200 music samples of different Latin music genres, namely Tango, Salsa, Forró and Axé. The achieved results show that the accuracy in music genre classification achieved by the RR decomposition provides better results than the OAA ensemble and baseline classifiers when considering the individual segments. Contrary to our expectative, the complete ensemble approach – using time and space decompositions – does not present superior results in comparison with the OAA and RR ensemble strategies on individual segments.

One solution that might improve the performance of the time-space ensemble approach in this scenario would be to base the final decision on the vote of all

space component classifiers, instead of using only the final label provided in each segment. For example, in the case of the RR approach it would be possible to make the final decision based on the majority vote of all the trained classifiers instead of only the output produced by each segment. Also it would be possible to use more robust rules instead of only the majority voting. Both aspects are subject of our current research.

## References


1. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing **10** (2002) 293–302
2. Aucouturier, J.J., Pachet, F.: Representing musical genre: A state of the art. Journal of New Music Research **32** (2003) 83–93
3. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Transaction on Pattern Analysis and Machine Intelligence **20** (1998) 226–239
4. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. Journal of Machine Learning Research **5** (2004) 101–141
5. Jacobs, R., Jordan, M., Nowlan, S., Hinton, G.: Adaptative mixtures of local experts. Neural Computation **3** (1991) 79–87
6. Schapire, R.: The strength of weak learnability. Machine Learning **5** (1990) 197–227
7. Fürnkranz, J.: Pairwise classification as an ensemble technique. In: Proceedings of the 13th European Conference on Machine Learning, Helsinki, Finland, Springer-Verlag (2002) 97–110
8. Li, T., Ogihara, M., Li, Q.: A comparative study on content-based music genre classification. In: Proceedings of the 26th annual international ACM SIGIR Conference on Research and Development in Informaion Retrieval, Toronto, Canada (2003) 282–289
9. Ho, T.K.: Nearest neighbors in random subspaces. In: Proc. of the 2nd Int'l Workshop on Statistical Techniques in Pattern Recognition. (1998) 640–648
10. Grimaldi, M., Cunningham, P., Kokaram, A.: An evaluation of alternative feature selection strategies and ensemble techniques for classifying music. In: Workshop on Multimedia Discovery and Mining at ECML/PKDD-2003. (2003)
11. Grimaldi, M., Cunningham, P., Kokaram, A.: A wavelet packet representation of audio signals for music genre classification using different ensemble and feature selection techniques. In: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval, ACM Press (2003) 102–108
12. Costa, C.H.L., Valle Jr, J.D., Koerich, A.L.: Automatic classification of audio data. In: IEEE International Conference on Systems, Man, and Cybernetics. (2004) 562–567
13. Dietterich, T.G.: Ensemble methods in machine learning. In: Proc. 1st International Workshop on Multiple Classifier System. Number 1857 in Lecture Notes in Computer Science, Springer (2000) 1–15
14. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys **34** (2002) 1–47
15. Tzanetakis, G., Cook, P.: Marsyas: A framework for audio analysis. Organized Sound **4** (1999) 169–175
16. Mitchell, T.M.: Machine Learning. McGraw-Hill (1997)
17. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. 2nd edn. Morgan Kaufmann, San Francisco (2005)