Automatic Genre Classification of Latin Music Using Ensemble of Classifiers

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

Graduate Program in Applied Computer Science (PPGIA)
Pontifical Catholic University of Paraná (PUCPR)
Rua Imaculada Conceição 1155, 80215-901
Curitiba - PR - Brasil

{silla, kaestner, alekoe}@ppgia.pucpr.br

Abstract. *This paper presents a novel approach to the task of automatic music genre classification which is based on ensemble learning. Feature vectors are extracted from three 30-second music segments from the beginning, middle and end of each music piece. Individual classifiers are trained to account for each music segment. During classification, the output provided by each classifier is combined with the aim of improving music genre classification accuracy. Experiments carried out on a dataset containing 600 music samples from two Latin genres (Tango and Salsa) have shown that for the task of automatic music genre classification, the features extracted from the middle and end music segments provide better results than using the beginning music segment. Furthermore, the proposed ensemble method provides better accuracy than using single classifiers and any individual segment.*

1. Introduction

With the continuous expansion of the Internet, a huge quantity of data from different sources has been become available on-line. An study from the UC Berkeley shows that in 2002 there were about 5 million terabytes of new information produced in films, printed media or magnetic/optic storage media [Lyman and Varian 2003]. In the Web alone, more than 170 terabytes of information is available. However it is very difficult to use in an efficient manner such a huge amount of information. Many important problems such as search for information sources, retrieval/extraction of information, automatic summarization of information, etc. have been the subject of intensive research in the last years. Automatic music genre classification is also the underlying technology to construct digital libraries that intend to preserve a relevant socio-cultural aspect.

In this context, a research area that has been growing in the past few years is the multimedia information retrieval which aims at building tools to effectively organize and manage the huge amount of available multimedia information [Fingerhut 1999] [Pampalk et al. 2002]. The current practice for indexing multimedia data is based on textual meta-data information, which is the case of the ID3 tags in MP3 music files. Although ID3 tags are very useful for indexing, searching, and retrieval, usually, such tags are manually generated and associated with the multimedia data.

One of the most important type of multimedia data distributed over the Web is the digital music in MP3 format. There are many studies and methods related to the analysis of the music audio signal [Aucouturier and Pachet 2003], [Guo and Li 2003], [Li et al. 2003], [Pampalk et al. 2002], [Zhang and Kuo 2001]. One important component for a content-based music information retrieval system is a module for the automatic

music genre classification [Li and Ogihara 2005]. Music genres are categoric labels created by humans in order to determine the style of music. Even for humans to classify a music according to its genre involves a subjective judgement, depending on their socio and cultural aspects. [Lippens et al. 2004] report an experiment where human judges agree only in 79% of their classifications. The genre labels are related to the instrumentalization, rhythmic structure and harmonic content of the music. Even if the music genre is a somewhat ambiguous descriptor, it has been used to categorize and organize large collections of digital music [Aucouturier and Pachet 2003], [Pampalk et al. 2002], [Tzanetakis and Cook 2002].

The content-based music genre classification is a recent field of research. In the work of [Tzanetakis and Cook 2002] a set of features based on music signals which are related to the timbral texture, rhythm and pitch was proposed. Such features have been used together with two classifiers, namely k nearest neighbors (k-NN) and Gaussian mixture models (GMM). [Li et al. 2003] proposed a novel method for feature extraction based on the Daubechies Wavelet Coefficients Histogram (DWCH) and compared it with the feature set proposed in [Tzanetakis and Cook 2002]. In this work the classifiers evaluated were Support Vector Machines (SVM), k-NN, GMM and Linear Discriminant Analysis. The best results were achieved using the SVM classifier. An unsupervised approach using Hidden Markov Models (HMMs) was proposed in the work of [Shao et al. 2004]. One common aspect of most works in the area is that they often use only one feature vector extracted from a segment (usually 30 seconds) of a music file. One of the few exceptions is the work of [Costa et al. 2004] which introduced the idea of segmenting the music audio file into three 30-second segments, training a classifier for each segment, and combining the classifiers decision in order to improve the final prediction about the music genre. In this work the segmentation method was evaluated employing a k-NN and a Multi-Layer Perceptron Neural Network (MLP) classifier.

The main motivation of this work is to analyze Latin music audio signals, which present a significant variation in time. To account for such a variation it would be better to extract feature vectors from the whole music, however this is a computationally expensive process. In order to overcome this problem, the strategy that is often adopted is extracting features from only one segment of the music. However, this approach is not reliable since the classification of different music segments can lead to different classification outputs and high error rates. For this reason, in this work we have extended the method proposed by [Costa et al. 2004] with other learning algorithms (Decision Trees, SVM and Naïve Bayes), a different feature set and ensemble of classifiers. The experiments were performed on a novel data set which is composed of Latin music. The reason for considering Latin Music is because we believe that the development of tools for different music styles is as important as the development of tools for other languages than English. For music, the main reason is that different music genres have different influences and instrumentalization.

This paper is organized as follows: Section 2 presents an overview of the system. Section 3 reports the experimental results and their analysis. Section 4 presents the conclusions and discusses future work.

2. System Overview

The Latin music genre classification system proposed in this paper is composed of three main phases (Fig. 1): feature extraction, classification and decision based on an ensemble. Initially features are extracted from three 30-second segments of the audio signal. These segments are chosen from the beginning, middle and end portion of the music since for

many music the audio signal has a significant variability in time. In this way each segment of the music is represented by a feature vector.

Since this is a system that employs supervised classification algorithms, it operates in two modes: training and classification. In the training mode the feature vectors are used with their respective labels by the learning algorithms. The labels consist in the textual information that represents the musical genre assigned to the music by human experts. In the classification mode, a music file, which genre is unknown, is provided to the system. Similarly to the training mode, three 30-second segments of the music are selected and feature vectors are generated. Each feature vector feeds a single classifier which will assign a genre to the music. The output of the classifiers are then combined through the majority vote rule and based on such a combination, a genre is assigned to the music. In the next section, the feature set, the classifiers, and the ensemble method are presented in details.

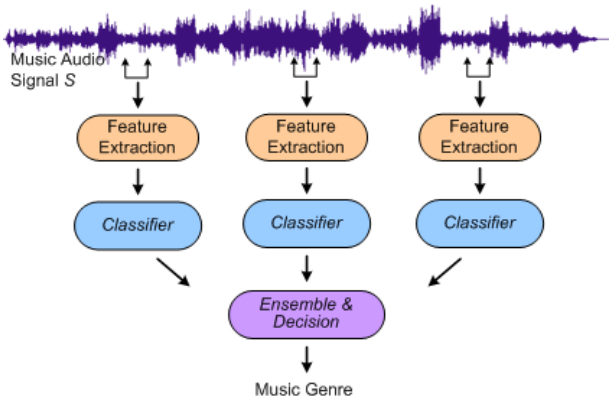

Figure 1: System Overview

2.1. Feature Extraction

In this work the problem of automatic music genre recognition is viewed as a pattern recognition problem where a music sample is represented in terms of feature vectors. The aim of feature extraction is to represent a music into a compact and descriptive way and that is suitable to deal with learning algorithms.

Since good quality digital music have about 1MB per minute, the extraction of features from the whole music can be prohibitive due to the required processing time [Costa et al. 2004]. For that reason features are extract from three 30-second music segments. The music segments have the same duration (which are equivalent to 1.153 frames in MP3 files). It is important to notice that regardless of the bitrate of the file, when dealing with MP3 files, the number of frames (which denotes the duration of the music) is always the same [Hacker 2000]. For this reason we use the following strategy to extract features from three segments of each music:

- the 1st segment is extracted from the beginning of the music, from frame 0 to frame 1.153;
- Let L denotes the total number of frames of a music, the 2nd segment is extracted from the middle of the music, from frame $(L/3) + 500$ to frame $(L/3) + 1.653$;
- 3rd is extracted from the end part of the music but a particular strategy is adopted to avoid getting noisy or silenced endings that are common in some MP3 files. The 3rd segment is extracted from frame $L - 1.453$ to frame $L - 300$).

For the extraction of features from the music segments, the Marsyas[1] framework [Tzanetakis and Cook 2000] was employed. The Marsyas framework implements the

[1]Available at: http://marsyas.sourceforge.net

original feature set proposed by [Tzanetakis and Cook 2002]. The features used can be divided into three groups: Timbral Texture, Beat Related and Pitch Related. The features based on the Timbral Texture are extracted based on the means and variance of the spectral centroid, rolloff, flux, the time zero domain crossings, the first 5 MFCCs (Mel-frequency cepstral coefficients) [Tzanetakis and Cook 2002] and low energy. Features that are beat-related include the relative amplitudes and the beat per minute. Pitch related features include de maximum periods of the pitch peak in the pitch histograms. The final feature vectors are 30-dimensional (Timbral Texture: 9 STFT (Short Time Fourier Transform) + 10 MFCC; Beat: 6; Pitch: 5) [Tzanetakis and Cook 2002].

2.2. Classification and Ensemble

The problem of music genre classification can be summarized as follows: given a music segment represented by its feature vector, one must assign a label (music genre or class) which better represents these input.

Machine Learning (ML) techniques are usually employed for classification. Tipically a supervised ML algorithm works in two modes [Mitchell 1997]: (i) in *training mode* a set of previously classified objects is used to produce a decision structure or model; (ii) in *classification mode* the generated model is applied to an unseen object and assigns a class to it.

In this work we have used the following supervised machine learning algorithms as component classifiers for the ensemble methods: Decision Trees (J48), k-NN, Naïve Bayes, Support Vector Machines with the pairwise classification decomposition strategy and an MLP Neural Network trained with the backpropagation with momentum learning algorithm which has 16 neurons in the hidden layer and 2 neurons in the output layer. The classification framework is based on the Weka Datamining Tool [Witten and Frank 2005].

In order to combine the decisions of the individual classifier trained on each segment of the same music, their output are combined to improve the final classification performance. The combination of the results is achieved through the majority voting rule taking into account only the predicted class labels from each single classifier. In case there is a draw between the different classifiers (i.e., each one outputting a different class) the following strategy is adopted: between the classifiers that are drawn, the genre will be labeled according to the class with the highest confidence score. In case there is another drawn, this time with the confidence scores, the decision will be made based on the drawn segments. If one of the drawn segments is the 2nd segment then the final label will be given by its classifier otherwise the final label will be given by the 1st segment classifier.

3. Experiments

We have selected 300 music samples from each genre and split them into training set with 150 (50%) samples from each class, that is, for each music genre; validation set with 60 (20%) samples from each genre; and test set with 90 (30%) music samples from each genre. It is important to notice that to avoid any biasing in the experiments, all the available music has been random selected without reposition from the database. Another important aspect of this dataset is that each music sample was labeled by an human expert after manual inspection. Regardless of Pachet's suggestion [Aucouturier and Pachet 2003] of using CDs from collections of CDs or theme, in the case of Latin music this approach is inefficient for labeling. For example in the case of the four cd collection (*Los 100 Mayores Exitos De La Musica Salsa*) only half (50 of 100) of the songs can be classified as Salsa, the remaining of the collection belongs to other music genres like Merengue, Lambada and even Samba. Another approach that could have been used for automatically

Table 1: Accuracy using Individual Segments

Classifier	1st Segment	2nd Segment	3rd Segment
J48	93.33%	93.88%	90.55%
k-NN (k=1)	93.88%	99.44%	98.33%
k-NN (k=3)	96.66%	99.44%	97.77%
k-NN (k=5)	96.11%	99.44%	98.88%
k-NN (k=7)	96.11%	99.44%	98.33%
MLP	98.33%	100%	98.88%
Naïve Bayes	95.00%	99.44%	95.00%
SVM	98.33%	100%	98.88%

labelling the music would be the classification of albums based on the artist profile (which is common practice in the area). However this approach does not seem to be adequate especially when working with Latin Music genres. For example, if we wanted to add songs by Carlos Gardel (who is a famous Tango Composer) all his songs would be labeled as Tango and although he has over 500 compositions only around 400 of them are Tangos, introducing unnecessary noise from a machine learning perspective. But even with other artists from a specific genre, like Salsa, hardly ever all the tracks from an album are all from the same genre. One interesting fact that was perceived during the dataset creation is when working with Latin Music Genres, usually at least one to three songs are not from the main style of the Artist profile.

Table 1 shows that in the case of Latin rhythms using only the beginning music segment is not a good strategy. In all cases, the best results were achieved on the middle segment, and with the exception of the J48 classifier, in all other cases the 2nd best classification accuracy was achieved using the 3rd segment which represents the end of the music.

The results achieved using the method of combination and decision based on the majority vote rule are presented in Table 2. With the exception of the Naïve Bayes classifier, all the results are at least as good as the performance achieved using only a single music segment. In the case of the J48 classifier, the accuracy was improved in 5%.

As mentioned earlier, this method of ensemble and decision based on three segments extracted from the music piece was originally proposed in [Costa et al. 2004] where the method was evaluated using music from the musical genres Rock and Classic. In the previous experiments, the results were not improved significantly using this method. However in this work the music samples used are from different musical genres (Tango and Salsa) which seems to benefit from the ensemble strategy adopted. This might be

Table 2: Accuracy using Ensemble of Classifiers

Classifier	Accuracy
J48	98.33%
k-NN (k=1)	100%
k-NN (k=3)	100%
k-NN (k=5)	99.44%
k-NN (k=7)	99.44%
MLP	100%
Naïve Bayes	98.33%
SVM	100%

due to the nature of the genres, since Rock and Classic are usually more constant than Latin rhythms. In the case of Salsa, most of music samples starts slow (sometimes as slow as a Bolero) in the introduction and after a while they "explode" (at the time when all instruments come into play). The results are in accordance with the positioning of [Li and Ogihara 2005] who states that different strategies are needed for the classification of different music genres when some sort of hierarchical classification is taken into account. This indicates that the strategy of segmenting the music piece into three segments and the combination from the ensemble of the classifiers trained in these segments might be more appropriate to use with specific genres or sub-genres. Unfortunately a direct comparison with the experiments performed earlier in [Tzanetakis and Cook 2002], [Li et al. 2003] is not possible due to the fact that although the data set used is available it contains only the initial 30 seconds of each music sample.

4. Concluding Remarks

In this paper we presented an evaluation of different classifiers with an ensemble technique applied to three different segments of the same song for the task of automatic music genre classification. The genres used were two Latin Genres, namely Tango and Salsa. The results achieved by the use of the ensemble, with the exception of the Naïve Bayes classifier, were positive and in the worst case achieved at least the same accuracy as the best classifier of one of the individual segments.

An analysis of the achieved results shows that without the ensemble method, the segment of the middle of the song provides higher classification accuracy than the other two (beginning and end). This is an interesting finding since most work of the literature [Tzanetakis and Cook 2002] considers using only the beginning (initial thirty seconds) of each song.

As future work, we plan to use more sophisticated rules of combination instead of only the majority vote like the output confidence score provided by each classifier and also expand the number of Latin genres available in the data set adding rhythms like Samba, Merengue, Lambada, etc.

References

Aucouturier, J. J. and Pachet, F. (2003). Representing musical genre: A state of the art. *Journal of New Music Research*, 32(1):83–93.

Costa, C. H. L., Valle Jr, J. D., and Koerich, A. L. (2004). Automatic classification of audio data. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 562–567, Haia, Holanda.

Fingerhut, M. (1999). The ircam multimedia library: A digital music library. In *IEEE Forum on Research and Technology Advances in Digital Libraries*, pages 19–21.

Guo, G. and Li, S. Z. (2003). Content–based audio classification and retrieval by support vector machines. *IEEE Transactions on Neural Networks*, 14(1):209–215.

Hacker, S. (2000). *MP3: The Definitive Guide*. O'Reilly, 1st edition.

Li, T. and Ogihara, M. (2005). Music genre classification with taxonomy. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 197–200.

Li, T., Ogihara, M., and Li, Q. (2003). A comparative study on content-based music genre classification. In *Proceedings of the 26th annual international ACM SIGIR Confer-*

ence on Research and Development in Informaion Retrieval, pages 282–289, Toronto, Canada.

Lippens, S., Martens, J., Leman, M., Baets, B., Meyer, H., and Tzanetakis, G. (2004). A comparison of human and automatic musical genre classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 233–236.

Lyman, P. and Varian, H. R. (2003). How much information. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 on [06/25/2005].

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Pampalk, E., Rauber, A., and Merkl, D. (2002). Content–based organization and visualization of music archives. In *ACM Multimedia 2002*, pages 570–579, Juan-les-Pins, France.

Shao, X., Xu, C., and Kankanhalli, M. S. (2004). Unsupervised classification of music genre using hidden markov model. In *IEEE International Conference on Multimedia and Expo*, volume 3, pages 2023–2026.

Tzanetakis, G. and Cook, P. (2000). Marsyas: A framework for audio analysis. *Organized Sound*, 4(3).

Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition.

Zhang, T. and Kuo, C. C. J. (2001). Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing*, 9(4):441–457.