

A comparison of statistical approaches to symbolic genre recognition

Carlos Pérez-Sancho, Pedro J. Ponce de León and José M. Iñesta
Departamento de Lenguajes y Sistemas Informáticos
{cperez,pierre,inesta}@dlsi.ua.es
Universidad de Alicante
P.O. box 99, E-03080 Alicante, Spain

Abstract

Previous work in genre recognition and characterization from symbolic sources (melodies extracted from MIDI files) carried out by our group pointed our research to study how the different utilized approaches perform and how their different abilities can be used together in order to improve both the accuracy and robustness of their decisions. Results for a corpus of Jazz and Classical music pieces are presented and discussed.

1 Introduction

Some recent works explore the capabilities of machine learning or pattern recognition methods to recognize music genre, either using audio (Zhu, Xue, and Lu 2004; Whitman, Flake, and Lawrence 2001; Soltau, Schultz, Westphal, and Waibel 1998), or symbolic (Cruz, Vidal, and Pérez-Cortes 2003; McKay and Fujinaga 2004) sources. After a period of time researching on the use of statistical models and classification paradigms for music genre (or style) characterization (Ponce de León and Iñesta 2003; Pérez-Sancho, Iñesta, and Calera-Rubio 2004) from symbolic data, we show here a comparison of the performance of both paradigms pointing to what they share and how they can complement their work.

Current research on the use of different statistical description methods for classification of symbolic music files into genres is presented in this paper. For it, MIDI files have been used as the primary source of musical data. The paper presents the data, the description methods, and the classification techniques in a comparative fashion.

2 Music data

The corpus used is a set of MIDI files from *Jazz* and *Classical* music collected from different sources, without any processing before entering the system, except for manually

checking the presence and correctness of key, tempo, and meter meta-events, as well as the labeling of the melody track (monophonic) since we are interested in the part of music genre that may be conveyed by melody.

The corpus is made up of 110 files, 45 of them being classical music and 65 of jazz, with a total length around 10,000 bars (more than six hours of music). The music pieces have been selected from well-known authors from both styles, ranging a broad range of styles (see (Ponce de León and Iñesta 2003)). Also, a different set of 42 files, 21 in each style, was used for validating the performance of the system using an ensemble of classifiers.

Two different ways of describing the content of the melody track have been used. The first one is based on melodic, harmonic, and rhythmic statistical descriptors and the second one describes melodic content in terms of strings of symbols corresponding to melody subsequences. Both description methods are briefly described in the following sections.

3 Statistical description models

For both approaches explained below, a sliding window of width ω bars traverses the melody, shifting its position 4 beats each time, and provides statistical information about the music content for each window position. This way, a new dataset is constructed for each ω . Integer values $\omega \in [1, 100]$ have been used, providing datasets of different size and granularity in order to analyse the models' behaviour.

3.1 Shallow statistical descriptors

The first group of description models that have been used are based on descriptive statistics that summarise the content of a melody in terms of pitches, intervals, durations, silences, harmonicity, rhythm, etc. This kind of statistical description of musical content is sometimes referred to as *shallow structure description* (Pickens 2001).

In these models, each window is described by a vector of statistical descriptors, labeled with the genre of the original melody. A set of 28 descriptors has been defined, based on several categories of features that assess melodic, harmonic, and rhythmic properties of a melody. These descriptors are summarized in table 1. The first column indicates the musical property analysed and the other columns indicate the kind of statistics describing the property. A blank entry in the table means that a particular statistic has not been computed.

Table 1: Shallow structure descriptors and the feature sets in which they were included.

Category	Count	Range	Avg.-rel.	Dev.	Norm.
Notes	•				
Significant silences	•				
Non significant silences	•				
Pitches		•	•	•	•
Note durations		•	•	•	•
Silence durations		•	•	•	•
Inter-onset intervals		•	•	•	•
Pitch intervals		•	•	•	•
Non-diatonic notes	•		•	•	•
Syncopations	•				

Durations are measured in ticks. For pitch and interval categories, the range values are the difference between maxima and minima, and average-relative descriptors are computed as the average value minus the minimum value. For durations (note durations, silence durations, and inter-onset intervals), the ranges are computed as the ratio between maximum and minimum values, and the average-relative descriptors are computed as the ratio between the average and the minimum value. Finally, normality descriptors are computed using the D’Agostino statistic (D’Agostino and Stephens 1986) for assessing the normality of the distribution of each category values within a window.

3.2 n -word based descriptors

The n -word based models make use of text categorization methods. The technique encodes note sequences as character strings, therefore converting a melody in a text to be categorized. Such a sequence of n consecutive notes is called a n -word. All possible n -words are extracted from a melody, except those containing a silence lasting four or more beats. The encoding for n -words used in this work has been derived from the method proposed in (Doraisamy and Ruger 2003). This method generates n -words by encoding pitch interval and duration information. For each n -note window, all possible intervals and duration ratios are obtained, respectively, by the equations:

$$I_i = \text{Pitch}_{i+1} - \text{Pitch}_i \quad (i = 1, \dots, n-1)$$

$$R_i = \frac{\text{Onset}_{i+2} - \text{Onset}_{i+1}}{\text{Onset}_{i+1} - \text{Onset}_i} \quad (i = 1, \dots, n-2)$$

and each n -word is defined as a string of symbols:

$$[I_1 R_1 \dots I_{n-2} R_{n-2} I_{n-1} R_{n-1}] \quad (1)$$

where the intervals and duration ratios have been mapped into alphanumeric characters (see (Perez-Sancho, Inista, and Calera-Rubio 2004) for details).

This method represents a musical piece as a vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{i|\mathcal{V}|})$, where each component represents the presence of the word w_i in the melody, and $|\mathcal{V}|$ is the size of the vocabulary.

A common practice in text classification is to reduce the dimensionality of those vectors (usually very high) by selecting the words that contribute most to discriminate the class of a document (a melody here). The *average mutual information* (AMI) (Cover and Thomas 1991) has been used in this work to rank the words. This method gives a high value to those words that appear often in the melodies of one genre and are seldom found in the melodies of the other genres. The n -words are sorted using this value, and only information about the most informative words are provided to the classifier.

4 Classification techniques

4.1 Classifiers for shallow statistical features

Four conceptually different classification paradigms have been used with the four description models presented in section 3.1: Nearest-neighbour classifier, bayesian classifier, multilayer perceptron and support vector machines (Duda, Hart, and Stork 2000). These are standard machine learning techniques used nowadays for classification.

Thus, given a window of length ω (i.e. a dataset), four different classifiers have been trained. In order to estimate the accuracy of such classifiers, a 10-fold cross-validation scheme was used on each dataset.

4.2 Naive Bayes Classifier for n -words

For n -word based classification, the naive Bayes classifier (McCallum and Nigam 1998), has been used. Here, the classifier is based on the Bayes rule, but applying the *naive Bayes assumption*: all the n -words extracted from a melody sample are independent of each other, and also independent of the order they were generated. This assumption is clearly false, but naive Bayes can obtain near optimal classification errors in spite of that (Domingos and Pazzani 1997).

The class-conditional probability of a melody $P(x_i|c_j)$ is given by the probability distribution of note sequences (n -words) for genre c_j , which can be learned from a labeled training set, $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$.

Two different distribution models have been used for the class-conditional probability: a Multivariate Bernoulli (MB) model, where the vector components are $x_{it} \in \{0, 1\}$ and a Multinomial (MN) model, where components are $x_{it} \in \{0, 1, \dots, |\mathbf{x}_i|\}$, being $|\mathbf{x}_i|$ the number of n -words extracted from the melody.

In the MB model, each class follows a multivariate Bernoulli distribution where the parameters to be learned from the training set are the class-conditional probabilities for the words in the vocabulary.

The MN model, the probability that a melody has been generated from a genre c_j is a multivariate multinomial distribution, where the melody length is assumed to be class-independent.

4.3 Classifier ensembles

After analysing the performance of the different classifiers studied, we have found a diversity of errors among the decisions taken by the different classifiers. This diversity has been suggested by some authors (Kuncheva and Whitaker 2003) as an argument for using classifier ensembles with good results. These ensembles could be regarded as committees of ‘experts’ in which the decisions of individual classifiers are considered as opinions supported by a measure of confidence usually related to the accuracy of that particular classifier. The final classification is taken either by majority vote or by a weighing system.

4.3.1 Voting schemes

Designing a suitable method of decision combinations is a key point for the ensemble’s performance. In this paper, two different possibilities that are presented below have been proposed and compared. In the discussion that follows, N stands for the number of samples, contained in the training set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$, M is the number of classes in a set $\mathcal{C} = \{c_j\}_{j=1}^M$, and K classifiers, C_k , are utilized.

1. Best-worst weighted majority. In this ensemble, the best and the worst classifiers in the ensemble are identified using their estimated accuracy. A maximum authority, $a_k = 1$, is assigned to the former and a null one, $a_k = 0$, to the latter, being equivalent to remove this classifier from the ensemble. The rest of classifiers are rated linearly between these extremes. The values for a_k are calculated as follows:

$$a_k = 1 - \frac{e_k - e_B}{e_W - e_B} ,$$

where

$$e_B = \min_k \{e_k\} \quad \text{and} \quad e_W = \max_k \{e_k\}$$

and e_k is the number of errors made by C_k .

2. Quadratic best-worst weighted majority. In order to give more authority to the opinions given by the most accurate classifiers, the values obtained by the former approach are squared. This way,

$$a_k = \left(\frac{e_W - e_k}{e_W - e_B} \right)^2 .$$

4.3.2 Classification

Once the weights for each classifier decision have been computed, the class receiving the highest score in the votation is the final class prediction. If $\hat{c}_k(\mathbf{x}_i)$ is the prediction of C_k for the sample \mathbf{x}_i , then the prediction of the ensemble can be computed as

$$\hat{c}(\mathbf{x}) = \arg \max_{c_j \in \mathcal{C}} \sum_k w_k \delta(\hat{c}_k(\mathbf{x}_i), c_j) , \quad (2)$$

being $\delta(a, b) = 1$ if $a = b$ and 0 otherwise.

Since the weights represent the normalized authority of each classifier, it follows that $\sum_{k=1}^M w_k = 1$. This makes possible to interpret the sum in Eq. 2 as $P(\mathbf{x}_i | c_j)$, the probability that \mathbf{x}_i is classified into c_j .

5 Results

The classifiers described in section 4 have been applied to our dataset with different parameter values both for the feature extraction and classifier tuning. A study of their performance as a function of the window length is presented in Fig. 1. In both cases the classification percentage for the different classification approaches are represented together.

It is remarkable how the different approaches have led to so similar results. This points to the compatibility of both techniques.

Two ensembles have been constructed using the votation methods described above (represented as V1 and V2). The decisions of the ensembles are displayed in Table 2 (# errors column). Note that the ensemble’s performance using the quadratic best-worst strategy improves the behaviour of the best of the individual classifiers: just two errors against the three errors made by 7-nearest neighbour classifier based on the whole set of shallow descriptors.

XXX Note how, although the results are not the same as earlier, the ensembles maintain a high standard of precision, with just 4 errors. This clearly improves the performance of the now best classifier (7 errors), so the ensemble seems quite robust and performs well, specially with the best-worst strategies introduced here.

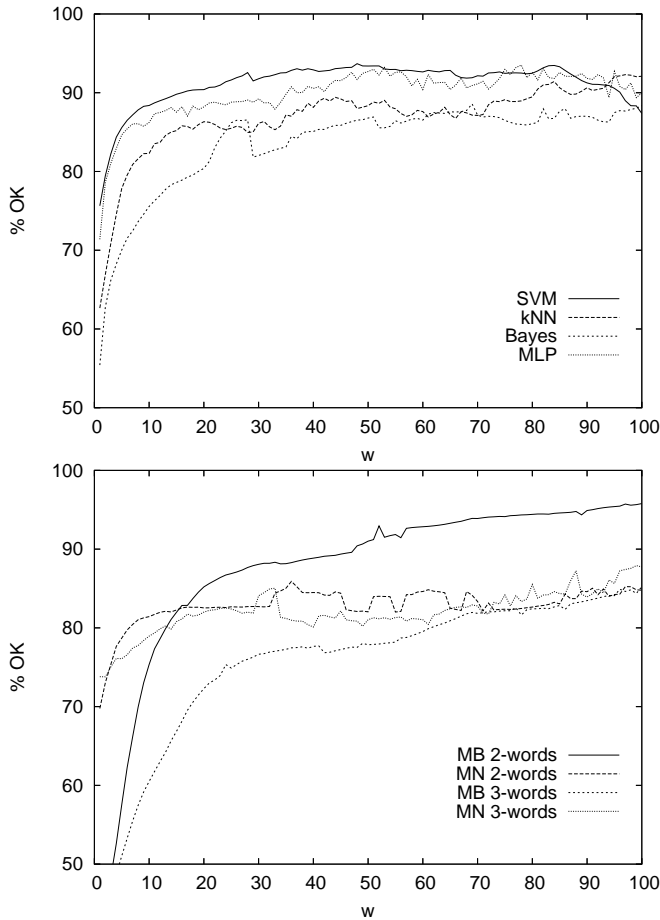


Figure 1: Performance of the classifiers for the shallow (top) and n -word (bottom) approaches.

6 Conclusions

We have shown the performance of classifier ensembles for classifying a symbolically represented melody into a given music genre. In previous works we have shown the feasibility of using these kind of data and representations to approach this problem, but by constructing an ensemble using different classifiers, their votes are “averaged” and this reduces the risk of choosing the wrong classifier.

Further work is needed to test the robustness of the capabilities of these approaches to classify other music genres.

7 Acknowledgments

This work was supported by the projects Spanish CICYT TIC2003-08496-C04, partially supported by EU ERDF.

Table 2: Ensemble’s performance.

Voting method	# err all	%
V1	2	95.2
V2	1	97.6
Best single classifier	# err all	%

References

- Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory*. John Wiley.
- Cruz, P. P., E. Vidal, and J. C. Pérez-Cortes (2003). Musical style identification using grammatical inference: The encoding problem. *LNCS 2905*, 375–382.
- D’Agostino, R. B. and M. A. Stephens (1986). *Goodness-of-Fit Techniques*. New York: Marcel Dekker, Inc.
- Domingos, P. and M. Pazzani (1997). Beyond independence: conditions for the optimality of simple bayesian classifier. *Machine Learning 29*, 103–130.
- Doraisamy, S. and S. Rürger (2003). Robust polyphonic music retrieval with n -grams. *Journal of Intelligent Information Systems 21*(1), 53–70.
- Duda, R. O., P. E. Hart, and D. G. Stork (2000). *Pattern Classification*. John Wiley and Sons.
- Kuncheva, L. I. and C. J. Whitaker (2003). Measures of diversity in classifier ensembles. *Machine Learning 51*, 181–207.
- McCallum, A. and K. Nigam (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, pp. 41–48.
- McKay, C. and I. Fujinaga (2004). Automatic genre classification using large high-level musical feature sets. In *Int. Conf. on Music Information Retrieval, ISMIR 2004*, pp. 525–530.
- Pérez-Sancho, C., J. M. Iñesta, and J. Calera-Rubio (2004). Style recognition through statistical event models. In *Proceedings of the Sound and Music Computing Conference, SMC ’04*.
- Pickens, J. (2001). A survey of feature selection techniques for music information retrieval. Technical report, Center for Intelligent Information Retrieval, Dept. Computer Science, Univ. Massachusetts.
- Ponce de León, P. J. and J. M. Iñesta (2003). Feature-driven recognition of music styles. *LNCS 2652*, 773–781.
- Soltau, H., T. Schultz, M. Westphal, and A. Waibel (1998). Recognition of music types. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP-1998)*, pp. 1137–1140.
- Whitman, B., G. Flake, and S. Lawrence (2001). Artist detection in music with minnowmatch. In *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pp. 559–568.
- Zhu, J., X. Xue, and H. Lu (2004). Musical genre classification by instrumental features. In *Int. Computer Music Conference, ICMC 2004*, pp. 580–583.