INFERRING EFFICIENT HIERARCHICAL TAXONOMIES FOR MIR TASKS: APPLICATION TO MUSICAL INSTRUMENTS

Slim ESSID
GET-Télécom Paris, CNRS LTCI,
37, rue Dareau, 75014 Paris
slim.essid@enst.fr

Gaël RICHARD
GET-Télécom Paris, CNRS LTCI,
37, rue Dareau, 75014 Paris
gael.richard@enst.fr

Bertrand DAVID
GET-Télécom Paris, CNRS LTCI,
37, rue Dareau, 75014 Paris
bertrand.david@enst.fr

ABSTRACT

A number of approaches for automatic audio classification are based on hierarchical taxonomies since it is acknowledged that improved performance can be thereby obtained. In this paper, we propose a new strategy to automatically acquire hierarchical taxonomies, using machine learning methods, which are expected to maximize the performance of subsequent classification. It is shown that the optimal hierarchical taxonomy of musical instruments (in the sense of inter-class distances) does not follow the traditional and more intuitive instrument classification into instrument families.

Keywords: Hierarchical taxonomy, musical instrument, clustering, probabilistic distance.

1 INTRODUCTION

Recently, hierarchical taxonomies have been profitably used for audio classification tasks, especially musical instrument classification [1, 2, 3, 4] and genre classification [5, 6, 7]. In the first place, by recurring to hierarchical classification, it is desired to achieve better classification performance than the so-called "flat" systems, wherein all classes are put at the same level without any arrangement. Furthermore, classification scalability is thereby obtained in the sense that "coarse" classification yielding top-level (more vague) labelling of some sound properties is made possible.

In most studies, straightforward taxonomies were considered which were borrowed from other areas of activity. Typically, taxonomies used in instrument classification [1, 2, 3] are highly inspired by instrument family divisions derived from instrument physics and/or musicology based categories, whereas taxonomies exploited in musical genre classification essentially originate from the music industry.

Such taxonomies present the advantage of being habitual and intuitive, hence enabling a certain ease of application for any potential end-user. On the other hand, they suffer from two major drawbacks. First, on the basis of intuition, a high number of alternative taxonomies can be potentially used leading to heterogeneous systems and contradictory classifications. Second, such taxonomies are not necessarily meant to maximize the accuracy of the classification tasks.

Attempts to address both issues were made in previous work. Pachet & Cazaly proposed "guiding principles" to be used in building a music genre taxonomy [5]. The application of Multi-Dimensional Scaling (MDS) analysis to observe dissimilarities among musical instruments [8, 3] can also be considered as an important step towards finding "natural" organizations among sound classes. Finally, very recently, a taxonomy of musical genres was induced by grouping genres which were the most frequently confused by a given classifier [7].

We propose an algorithm to acquire automatic taxonomies using unsupervised machine learning techniques in order to obtain solutions which are expected to yield the best classification performance. Our approach makes use of hierarchical clustering to produce a tree data structure wherein nodes represent optimal groupings of classes with respect to a robust probabilistic distance criterion.

We start by describing our algorithm and the related machine learning concepts. Subsequently, we present applications of our method to the case of musical instruments. Finally, we suggest some conclusions.

2 ALGORITHM DESCRIPTION

2.1 Overview

We aim to obtain a hierarchical taxonomy of some musical descriptions which are associated with target classes (for example instruments, orchestrations or genres, etc.). These are materialized by the leaf nodes of the taxonomy tree representation. To this end, we organize target classes using a hierarchical clustering algorithm. This is known to be an optimal and natural way of arranging the data since the most similar classes with respect to the chosen closeness criterion are then put in the same clusters.

Thus, the choice of the closeness criterion is critical. We need robust distances enabling us to reduce the effect of noisy features on the clustering performance. Also, the

distances are required to be matched with the behavior of the classifiers to be used. A convenient and robust means for measuring the closeness or separability of data classes is to use probabilistic distance measures between them, *i.e* distances between their probability distributions [9]. This is an interesting alternative to classic Euclidean distance between feature vectors known to be inefficient for sound source classification.

Another fundamental choice is the class descriptors. A large number of useful attributes can be examined and reduced using a feature selection algorithm to retain only the attributes that are relevant for proper overall class discrimination.

2.2 Clustering the target classes

We wish to group together a number of M class probability densities p_i into a number of M_c clusters C_i within L levels of a hierarchical taxonomy. Thus we need appropriate probabilistic distances. Many such distances can be considered among which we chose the Bhattacharryya and divergence due to the resulting simplification in the following computations. The divergence distance J_D between two probability densities p_1 and p_2 is defined as

$$J_D(p_1, p_2) = \int_{\mathbf{x}} [p_1(\mathbf{x}) - p_2(\mathbf{x})] \log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} d\mathbf{x}. \quad (1)$$

The Bhattacharryya distance is defined as

$$J_B(p_1, p_2) = -\log \left(\int_{\mathbf{x}} [p_1(\mathbf{x}) p_2(\mathbf{x})]^{\frac{1}{2}} d\mathbf{x} \right). \quad (2)$$

While these distances admit analytical expressions whenever the class probability densities are Gaussian, computing such distances can be otherwise a difficult problem since it requires performing heavy numeric integrations [10]. In fact, in the Gaussian case, the distances can be expressed as functions of the means and covariance matrices according to

$$\begin{aligned} J_D(p_1, p_2) &= \tfrac{1}{2}(\mu_1 - \mu_2)^T(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2) \\ &+ \tfrac{1}{2} tr(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I_D), \end{aligned}$$

$$\begin{aligned} J_B(p_1, p_2) &= \tfrac{1}{8}(\mu_1 - \mu_2)^T[\tfrac{1}{2}(\Sigma_1 + \Sigma_2)]^{-1}(\mu_1 - \mu_2) \\ &+ \tfrac{1}{2} \log \frac{\left|\frac{1}{2}(\Sigma_1 + \Sigma_2)\right|}{|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}}}, \end{aligned}$$

where (μ_1, Σ_1) and (μ_2, Σ_2) are the mean vectors and the covariance matrices of the multivariate Gaussian densities describing respectively class 1 and class 2 in $\mathbb{R}^D$. Nevertheless, it would be highly sub-optimal, in our case, to assume that the original class observations follow Gaussian distributions since we deal with data with a non-linear structure. Fortunately, if this data is mapped from the original space to a Reproducing Kernel Hilbert Space (RKHS) [11], it is reasonable to assume it to be Gaussian in the RKHS [10]. Thus, a robust estimation of the needed probabilistic distances can be derived using analytical expressions provided that a proper estimation of the means and covariance matrices in the RKHS can be obtained. The strength of such an approach resides in that there is no need for knowing explicitly either the structure of the original probability densities or the non linear mapping to be used. Interested readers are referred to [10] for further details.

We then use agglomerative hierarchical clustering [9, 12] to produce "a hierarchy of nested clusterings" based on probabilistic distances in RKHS. The algorithm starts with as many clusters as original data objects ($M_c^1 = M$ at iteration 1), measuring the proximities $J(p_i, p_j)$ between all pairs of clusters and grouping together the closest pairs into new clusters to produce M_c^l new ones at iteration l, until all vectors lie in single cluster (at iteration M).

A convenient way to understand the result of such a procedure is to represent it as a graph (called *dendrogram*) which depicts the relations and proximities between the obtained nested clusters (see figure 1 for an example).

The relevance of the cluster tree can be evaluated by computing the *cophenetic correlation coefficient* [9]. The closer the cophenetic coefficient to 1, the more relevantly the cluster tree reflects the structure of the data.

Clustering is then obtained by cutting the dendrogram at a certain level or certain value of the vertical axis. By applying different cuts to the dendrogram we can obtain different clusterings (having a different number of clusters). The levels of the hierarchical taxonomy are to be induced from these alternative clusterings in such a way that the high levels are deduced from "coarse" clustering (low number of clusters) while the low levels are deduced from "finer" clustering (higher number of clusters).

3 TAXONOMIES OF MUSICAL INSTRUMENTS

Various taxonomies have been proposed for musical instrument classification on isolated notes roughly following the instrument families organization [1, 2, 3]. While some declinations are common to these studies, as for example the primary division of instruments into "sustained" and "pizzicati", other groupings are not unanimously shared, especially for the wind instruments.

It is worth to note that Peeters undertook a Multi-Dimensional Scaling (MDS) analysis based on the signal features in order to verify the consistency of the class tree he had assumed [3]. This provided objective justification of some of the choices made but could not be used to infer a taxonomy.

We here present an application of our algorithm to produce a hierarchical taxonomy of musical instruments. This taxonomy is to be induced from real world musical phrases and is expected to yield the organization that best matches the classification to be performed subsequently.

3.1 Feature extraction and selection

A wide selection of more than 300 signal processing features is considered including some of the MPEG-7 de-

scriptors. Since these features have been extensively described in various previous work in the field of Music Information Retrieval (see [13] for example), in the following, we merely list the attributes which we examined in our study.

- *Temporal features* consist of autocorrelation coefficients, features obtained from the statistical moments, zero crossing rates, and amplitude modulation features.
- *Cepstral features* are mel-frequency cepstral coefficients as well as their first and second time derivatives.
- *Spectral features* include features obtained from the statistical moments, MPEG-7 audio spectrum flatness, spectral irregularity, spectral crest, spectral slope, spectral decrease, frequency cutoff, temporal variation of spectrum, and octave band signal intensities and their ratios providing a coarse description of the energy distribution of sound partials [14].
- *Perceptual features* are also utilized, namely loudness, sharpness and spread.

In order to fetch the most relevant features for optimal class discrimination, we use a simple feature selection algorithm, belonging to the family of "filter" algorithms, which is based on Fisher's Linear Discriminant Algorithm (LDA) [12]. The chosen method computes the relevance of each candidate feature using the weights estimated by the LDA.

3.2 Experimental parameters

Nineteen instruments from all instrument families are considered. Table 1 sums up the studied instruments giving their codes. Solo musical phrases played by each of these instruments were excerpted from commercial recordings. We had at least 4 different sources (different album, different artist) and at least 3 minutes available for each instrument.

All features described above were extracted on a frame basis. Unless otherwise specified, the default frame length is 32ms. Silence frames were detected and removed.

Instrument	Code	Instrument	Code
alto sax	As	oboe	Ob
bassoon	Bo	piano	Pn
double bass-pizzicato	Bs	tenor sax	Ts
double bass-bowed	Ba	soprano sax	Ss
bass clarinet	Cb	tuba	Tb
Bb clarinet	Cl	trombone	Tm
cello	Co	trumpet	Tr
flute	Fl	viola	Va
French horn	Fh	violin	Vl
classical guitar	Gt		

Table 1: Studied instruments and their codes.

Computing the probabilistic distances in RKHS (to be used for clustering) requires processing the EigenValue Decomposition of $n_k \times n_k$ Gram matrices [11], with n_k the number of training feature vectors of class C_k. Such an operation is computationally expensive ($O(n_k^3)$) since n_k is quite large. Hence, the training sets were divided into smaller sets of 1500 observations and the desired distances were obtained by averaging the distances approximated using as many reduced sets as possible. To measure these distances, one needs to choose a kernel function. We used the Radial Basis Function kernel.

3.3 Results

A total of 40 features were selected by the LDA approach from the original 304 candidates, namely:

- the 4 first mel-frequency coefficients (excluding the zero-th coefficient);
- the spectral centroid and the spectral asymmetry;
- the 15-th amplitude MPEG-7 spectral flatness coefficient;
- the frequency cutoff;
- Octave Band Signal Intensity (OBSI) coefficients 1, 2, 3 and 6, as well as Octave Band Signal Intensity Ratios (OBSIR) 1 to 6;
- spectral irregularity coefficient 5;
- the 4-th statistical moments measured both on the signal temporal waveform and amplitude envelope over 960-ms windows;
- the zero crossing rate measured over 32-ms windows and 960-ms windows;
- the Amplitude Modulation (AM) strength in the range 4-8Hz (tremolo) and the product of AM strength and AM frequency in the ranges 4-8 Hz (tremolo) and 10-40 Hz (graininess);
- relative specific loudness coefficients 1, 2, 5, 16, 18 and 21 as well as perceptual loudness and sharpness.

Based on these features, probabilistic distances in RKHS between each pair of considered classes were computed. Both the divergence and Bhattacharryya distances were obtained and fed to the agglomerative hierarchical clustering (described in section 2.2). A higher cophenetic coefficient was obtained with the Bhattacharryya distance compared to the one obtained with the divergence. Hence, more relevant clustering was obtained with the former. Its related dendrogram is depicted in figure 1. This can be already considered as a primary taxonomy. However, it is worth being processed so as to gain consistency and readability.

The processing of the tree consisted in applying 4 different cuts to the dendrogram, each cut inducing a level of hierarchy. The cuts were performed based on the *consistency coefficients* [9] for each dendrogram link so as to ignore the most inconsistent links. This resulted in the tree depicted in figure 2.

The obtained taxonomy does not follow the organization of instruments into traditional families. In fact, although some unions of the found solution are intuitive (for

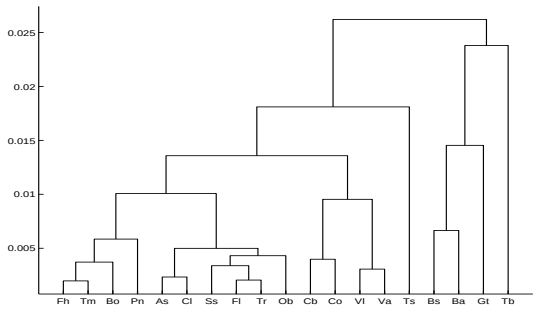

Figure 1: Dendrogram obtained with the Bhattacharryya distance. Vertical axis represents cluster distances.

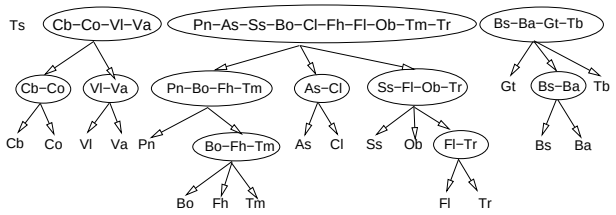

Figure 2: Obtained hierarchical taxonomy of musical instruments.

example, cello, violin and viola are put in the same cluster) many others may be surprising.

At the top level, bowed double bass and pizzicato double bass are grouped together with guitar and tuba, indicating that the "sustained/non-sustained" property has not been considered by our algorithm as useful for the classification. Indeed, since this property seems not to be captured by the selected features, it will not be "seen" by the classifiers to be used, hence it is not optimal to take it into account in the taxonomy. Additionally, the presence of tuba in the same cluster implies that features related to instrument register play an important role in the classification.

Most wind instruments are grouped together except the tuba, the bass clarinet- which is associated with cello, violin and viola- and tenor sax, which is left alone. The fact that our tenor sax excerpts are exclusively jazz excerpts while for all other instruments the sounds originate from both jazz and classic music might explain this exception. Finally, also surprising is that the piano lies in the same cluster as most wind instruments.

Going down in the hierarchy, interesting clusters are found. Alto sax is grouped with Bb clarinet, flute with trumpet, and bassoon with French horn and trombone. These arrangements do not really surprise us as they reflect the confusions which we have often noted in our previous experiments on instrument recognition using musical phrases. It appears that the instruments that are frequently confused are put by the algorithm in the same clusters.

4 CONCLUSIONS

In this paper, we have suggested a technique for inferring automatic taxonomies of musical descriptions expected to maximize the performance of subsequent classification. Our approach exploits robust probabilistic distances and agglomerative hierarchical clustering algorithms to produce class organizations in an unsupervised fashion.

We have tested this method in the context of musical instrument classification using signal processing features automatically selected from a high number of state-of-the-art features. The obtained arrangement of instruments is substantially different from usual taxonomies following instrument families organization. This suggests that the latter is probably not an optimal solution for automatic classification.

Future work will consider hierarchical classification experiments based on the induced taxonomies. Furthermore, we will attempt to address the feature selection problem in parallel to clustering so as to produce taxonomy-context dependent features.

ACKNOWLEDGEMENTS

This work was partly supported by the MusicDiscover project of the "ACI-Masse de données".

REFERENCES

[1] Keith Dana Martin. *Sound-Source Recognition : A Theory and Computational Model*. PhD thesis, Massachusets Institue of Technology, jun 1999.

[2] Antti Eronen. Automatic musical instrument recognition. Master's thesis, Tampere University of Technology, April 2001.

[3] Geoffroy Peeters. Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization. In *115th AES convention*, New York, USA, October 2003.

[4] Slim Essid, Ga el Richard, and Bertrand David. Instrument recognition in polyphonic music. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA, March 2005.

[5] F. Pachet and D. Cazaly. A taxonomy of musical genres. In *Content-Based Multimedia Information Access Conference (RIAO)*, Paris, France, April.

[6] C. McKay and I. Fujinaga. Automatic genre classification using large high-level musical feature sets. In *5th International Conference on Music Information Retrieval (ISMIR)*, Barcelona, Spain, October 2004.

[7] Tao Li and Mitsunori Ogihara. Music genre classification with taxonomy. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA, March 2005.

[8] Stephen McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff. Perceptual scaling of synthesized musical timbres: common dimensions, specificities and latent subject classes. *Psychological reserach*, 58:177–192, 1995.

[9] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern recognition*. Academic Press, 1998.

[10] S. Zhou and R. Chellappa. From sample similarity to ensemble similarity: probabilistic distance measures in reproducing kernel hilbert space. *IEEE transactions on pattern analysis and machine intelligence*. to be published.

[11] B. Sholkopf and A. J. Smola. *Learning with kernels*. The MIT Press, Cambridge, MA, 2002.

[12] Richard Duda and P. E. Hart. *Pattern Classification and Scence Analysis*. Wiley- Interscience. John Wiley & Sons, 1973.

[13] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, IRCAM, 2004.

[14] Slim Essid, Ga el Richard, and Bertrand David. Musical instrument recognition based on class pairwise feature selection. In *5th International Conference on Music Information Retrieval (ISMIR)*, Barcelona, Spain, October 2004.