Data mining of Basque folk music

Darrell Conklin CONKLIN@IKERBASQUE.ORG
Department of Computer Science and Artificial Intelligence
University of the Basque Country UPV/EHU, San Sebastián, Spain
IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

1. Background

This project is concerned with the development and application of data mining methods for the study of Basque song collections. Pattern discovery and classification algorithms are used to discover models and rules that relate musical content and the class labels of songs.

The Cancionero Vasco is a collection of Basque dance and song melodies, compiled by the musicologist, composer, and priest Padre Donostia in 1912 as part of a competition held by the Basque government to gather musical folklore of the region. A total of 1902 songs of the Cancionero Vasco are hosted by the Euskomedia Foundation, Usurbil, Spain.

Songs in the Cancionero Vasco contain two important types of information: musical data (in MIDI format) that encodes the melody, and metadata collected by Donostia including the region of collection of the song, and its genre. In the Cancionero Vasco a total of 24 distinct genres are referenced, besides toponyms organised in hierarchical levels of territorio (region), municipio (municipality), and nucleo (town). Each song in the Cancionero Vasco is annotated with exactly one toponym (at various levels of the toponymic hierarchy) and one genre label. Inference is applied to infer more general genre and toponymic class membership from more specific classes.

In addition to the annotated songs, in the Cancionero Vasco there are 341 songs that were not annotated with a genre at the time of song collection. It is hypothesised that predictive data mining methods can be used to predict a possible class for each of these, and also for new songs collected by musicologists in the future.

In *5th International Workshop on Machine Learning and Music*, Edinburgh, Scotland, UK, 2012.

2. Methods

Three main data mining methods are applied to the Cancionero Vasco.

Predictive data mining. The multiple viewpoints method for symbolic music classification has been applied to the Cancionero Vasco. This is a statistical ensemble classification method that combines the class predictions of multiple n-gram models of derived features of the melodic surface.

Association rule mining. To determine if songs in the Cancionero Vasco collected in specific regions are over- or under-represented in certain genres, all region/genre pairs were enumerated and tested for statistical significance (Fisher's exact test). Selected pairs were oriented into association rules and grouped into various categories depending on whether the pair is over- or under-represented and on the confidence of the oriented rule.

Pattern discovery. The pattern discovery method reported by Conklin (2010) is used to discover maximally general melodic patterns that are surprisingly over-represented in a corpus as related to a background or anticorpus. To apply this method, all classes (regions and genres) are in turn provided as the corpus, with the remainder of songs as the anticorpus. Results are ranked by their p-value. An *antipattern* (*anticorpus pattern*) is a pattern that is absent or surprisingly under-represented within a corpus but occurs frequently in an anticorpus. By inverting the role of corpus and anticorpus, and modifying the p-value computations, the same pattern discovery method can be used to discover both patterns and antipatterns (Conklin, 2012).

3. Results

This section provides an overview on some results achieved with the Cancionero Vasco using the data mining methods described above.

Classification accuracy. Many of the 24 genres are sparsely populated (9 genres have less than 10 songs), and therefore the three largest genres of *danza* (495 tunes), *amorosa* (247 tunes), and *religiosa* (209 tunes) are considered. This corpus contains 951 songs, covering 0.61 of the labelled pieces of the entire collection. Using a collection of 12 viewpoints, the multiple viewpoint method obtains a stratified 10-fold cross-validation accuracy of 0.776 on the corpus, with high precision (probability of class correct given class predicted) of 0.889 on the *danza* class:

	Predicted		
	religiosa	*amorosa*	*danza*
religiosa	102	80	27
amorosa	37	180	30
danza	11	28	456
Precision	0.680	0.625	0.889

By contrast, an SVM classifier using a feature vector of 99 global features (McKay & Fujinaga, 2004) achieves a significantly lower accuracy of 0.655 on the corpus. Given the precision on the *danza*, the trained model was used to predict which of the 341 unlabelled pieces might be dances, with the top five predictions all showing strong evidence (through title, annotations, or melodic similarity) of being dances.

Association rules. For regions R and genres G occurring at least once in the corpus, a total of 9000 G/R pairs including negations ($\overline{G}$ and $\overline{R}$) were enumerated and evaluated for statistical significance and musicological meaning. The following are examples of G/R association categories:

category	rule	p-value	conf
$G \to R$	*artaxuriketak* $\to$ *Nafarroa*	4.3e-05	0.79
$R \to G$	*Araba* $\to$ *danza*	7.8e-12	0.89
$G \to \overline{R}$	*danza* $\to \overline{\text{Atharratze}}$	0.00857	0.99
$R \to \overline{G}$	*Lapurdi* $\to \overline{\textit{artaxuriketak}}$	0.01112	0.99

Patterns and antipatterns. As an example of a pattern, the melodic interval pattern $[-4, +2, +2]$ occurs in $5/6 = 0.83$ of songs from the genre *epitalamios* (wedding songs), but only in $365/1892 = 0.19$ of songs from other genres. It can be said that this pattern is distinctive of wedding songs, because its relative frequency in that class is much higher than in the background.

As an example of an antipattern, the pattern $[+2, +5]$ occurs in 275 pieces in the anticorpus, but only in 2 (of 98) songs from the genre *cuneras* (cradle songs). The confidence of the oriented negative association rule $[+2, +5] \to \overline{\textit{cuneras}}$ is therefore high ($273/275 = 0.99$).

4. Conclusions

In the next phase of the project the mined results will be integrated into a formal ontology of classes that currently expresses all the toponyms and genres in their hierarchical relations (Goienetxea et al., 2012). For this a representation of melodic patterns in description logic will be employed. Associations between patterns, between genres and regions, and between patterns and genres will all be represented using conjunctive concepts in the ontology.

For predictive data mining the next step is to assign genres to the entire unlabelled section of the Cancionero Vasco using semi-supervised learning, while permitting the labelling of songs with more than one genre. The method of *class association rules* will be used as another method for the prediction of genre from song content, in this case using patterns. Finally possible annotation errors will be identified using both association rules and trained statistical classifiers.

Acknowledgments

This research was partially supported by a grant *Análisis Computacional de la Música Folclórica Vasca* (2011-2012) from the Diputación Foral de Gipuzkoa, Spain. Thanks to all participants of this project: Iñaki Arrieta, Jon Bagüés, Arantza Cuesta, Izaro Goienetxea, Pello Leiñena, and Kerstin Neubarth.

References

Conklin, D. Discovery of distinctive patterns in music. *Intelligent Data Analysis*, 14(5):547–554, 2010.

Conklin, D. Antipatterns in Basque folk tunes. In Diáz Báñez, J., Escobar Borrego, F., and Ventura Molina, I. (eds.), *Las Fronteras entre los Géneros: Flamenco y otras Músicas de Tradición Oral, II International Workshop on Folk Music Analysis - FMA*, pp. 219–224, Universidad de Sevilla, 2012.

Goienetxea, I., Arrieta, I. Bagüés, J., Cuesta, A., Leiñena, P., and Conklin, D. Ontologies for representation of folk song metadata. Technical Report EHU-KZAA-TR-2012-01, Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, 2012. http://hdl.handle.net/10810/8053.

McKay, C. and Fujinaga, I. Automatic genre classification using large high-level musical feature sets. In *Proceedings of the International Conference on Music Information Retrieval*, pp. 525–530, Barcelona, Spain, 2004.