Antipattern discovery in Basque folk tunes

DARRELL CONKLIN
Department of Computer Science and Artificial Intelligence
Universidad del País Vasco UPV/EHU, San Sebastián, Spain, and
IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
conklin@ikerbasque.org

Abstract

This paper presents a new pattern discovery method for labelled folk song corpora. The method discovers general patterns that are rare or even entirely absent in a corpus, and among those the ones that are the most general or frequent in the background set. The method is applied to two parallel ontologies of a large corpus of Basque folk tunes.

1 Introduction

In recent years there has been a renewed interest in folk song analysis, due to interest in cultural heritage and advances in music informatics methods. The ability to analyse music content for different properties of songs such as place name, dance type, tune family, tonality, and social function, is an important part of the management and understanding of large corpora.

The objective of the project *Análisis Computacional de la Música Folclórica Vasca* is the development and application of data mining methods to Basque song collections through automated pattern discovery and classification, using data mining algorithms to discover predictive models that relate musical content and the class labels of songs. This project will open new paths in the study and analysis of essential elements of Basque music, supporting the study of the

evolution and origins of Basque melodies, and will lead to methods for automatic classification and analysis of Basque folk 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. Recently the entire collection has been compiled in four volumes (de Riezu, 1996) and digitised, a process overseen by the Euskomedia Foundation[1] (Usurbil, Spain) and the Eresbil Foundation[2] (Renteria, 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 levels of territorio (region), municipio (municipality), and nucleo (town).

Much research to date on pattern discovery in music has been concerned with discovering patterns that are frequent, salient, over-represented, etc. in an analysis piece or set of pieces. This paper presents a method for discovering patterns that by contrast are infrequent, rare, and under-represented. Such patterns might be used, for example, within predictive classification, where their occurrence might strongly suggest against membership in a class. The pattern discovery method is applied with illustration to the Cancionero Vasco.

2. Methods

This section describes the theory leading to a new method for discovering under-represented

[1] www.euskomedia.org
[2] www.eresbil.com

patterns in a corpus. The general methods of subgroup discovery are presented, followed by a general method for applying subgroup discovery to sequential patterns in music. It is then demonstrated how this method may be readily modified to find under- rather than over-represented patterns in a corpus.

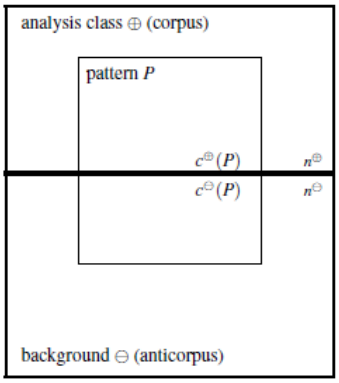


Figure 1: The schema for subgroup discovery, showing the major regions of objects involved. The top part of the outer box encloses the class of interest (in music called the corpus), below this the background (the anticorpus). The inner box contains the objects described by a pattern, and the top part of the inner box the subgroup described by a discovered pattern.

Notation	Meaning
P	a pattern
$\oplus$	corpus
$\ominus$	anticorpus
$c^{\oplus}(P)$	number of pieces with pattern P in the corpus
$c^{\ominus}(P)$	number of pieces with pattern P in the anticorpus
$n^{\oplus}$	number of pieces in the corpus
$n^{\ominus}$	number of pieces in the anticorpus

Table 1: Glossary of Notation.

2.1 Subgroup discovery

Figure 1 depicts the data mining scenario known as *subgroup discovery*, or alternatively *supervised descriptive rule discovery* (Novak et al., 2009), a relatively recent paradigm for data mining. Given an analysis class, subgroup discovery attempts to discover patterns predictive of the class not for any possible example, but only for a subset (subgroup) among them covering as few of the $\ominus$ objects as possible. An ideal pattern would have as its extension all and only the $\oplus$ examples. In practice this is not possible except in pattern description languages that permit disjunction (where a trivial disjunctive description of all examples may be possible).

In contrast to supervised predictive methods, subgroup discovery must therefore realise two tasks: identify the interesting subgroups, then (in fact, in parallel) describe them with comprehensible patterns. Thus the method is at the same time supervised (using labelled data) and descriptive (not having class prediction as the main objective).

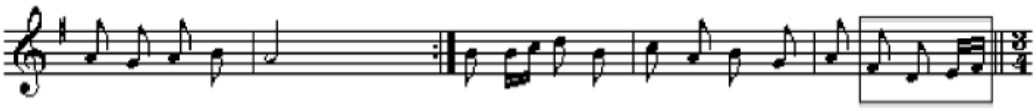
In general, in the supervised descriptive data mining scenario, the patterns discovered may not cover all examples, that is, they are agnostic about making predictions of examples that are not matched by the pattern. Therefore the results of data mining are evaluated according to the interest of patterns (usually some statistical measure of over-representation) rather than classification accuracy in the case of supervised predictive methods.

2.2 Sequential pattern mining in music

To extend supervised descriptive methods towards music, Conklin (2010a) presented the idea of using distinctive sequential patterns to describe subgroups. A sequential pattern in music is a sequence of features of notes, for example, [+2,+1] is a sequence of melodic intervals that

matches (for example) the note sequences [C, D, E♭] or [D, E, F].

There is a close connection with distinctive pattern discovery in music and gene set enrichment studies in bioinformatics (Al-Shahrour et al., 2004). There, genes of interest (which are selected by the scientist, for example by being over-expressed in some experiment) are probed with various gene ontology terms to find terms that are overrepresented within the selected set of interest. The analogy to pattern discovery in music is that pieces in a specified class are analogous to genes of interest, and patterns are analogous to gene ontology terms.



		class *wedding song*	
		no	yes
pattern [−4, +2, +2]	no	1527	1
	yes	$c^{\ominus}(P) = 365$	$c^{\oplus}(P) = 5$
		$n^{\ominus} = 1892$	$n^{\oplus} = 6$

Table 2: A contingency table for a Basque folk tune pattern. The p-value of the association between P and ⊕ is 0:0014.

Subgroup discovery introduces computational complexities for music, because the space of terms (patterns) used to describe subgroups is not fixed, as in the bioinformatics studies, and may be practically infinite and therefore the search for interesting patterns must be handled carefully. This applies even when a simpler representation of conjunctions of global piece features is used to describe subgroups (Taminau et al., 2009). The MGDP (maximally general distinctive pattern) algorithm (Conklin, 2010a) discovers associations between patterns and classes in an efficient way due to its structuring and pruning of the pattern search space. It uses two important concepts to manage the problem of a large pattern space: these are *distinctiveness* and *generality*.

To illustrate the concept of distinctiveness, Table 2 shows a contingency table for the melodic interval pattern [-4,+2,+2] that occurs in a subgroup of songs of the genre wedding song. The pattern appears in $c^{\oplus}$(P)=5 of $n^{\oplus}$=6 (83%) of wedding songs, but only in $c^{\ominus}$(P) = 365 of $n^{\ominus}$=1892 (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 higher than in the background. Any pattern with a relative probability above a threshold may be considered distinctive.

The Fisher's exact test, based on the cumulative hypergeometric distribution (Falcon and Gentleman, 2008), is used to further quantify the statistical significance of an association between a pattern and a class. Referring to Figure 1, the *p-value* of an association is the probability of laying the inner box, i.e. drawing a set of $c^{\oplus}$(P)+$c^{\ominus}$(P) pieces from the corpus of $n^{\oplus}$+$n^{\ominus}$ pieces, and finding $c^{\oplus}$(P) or more members of the class ⊕. Lower p-values indicate more surprising associations. The function is symmetric: the same probability results from drawing $n^{\oplus}$ pieces from the corpus and finding $c^{\oplus}$ (P) or more pieces containing the pattern P. In Figure 2, for example, the cumulative hypergeometric distribution gives the probability of finding 5 or more wedding songs in a sample of 370 pieces (or, symmetrically, 5 or more pieces containing the pattern [-4,+2,+2] in a sample of 6 pieces).

The concept of generality or subsumption is very important to structure the search and presentation space of patterns. A pattern is *subsumed* by another if all songs that contain the pattern also will contain the other (for example, the pattern [-4,+2,+2] is subsumed by the pattern [+2]). The MGDP algorithm discovers a set of patterns that are both distinctive and

among those the most general (not subsumed by any other distinctive pattern). For example, if the pattern [+2] is not distinctive it would not be reported. If the pattern [-4,+2,+2] is distinctive, then no more specific pattern (e.g., [4,+2,+2,+1]) would be reported, and in fact the entire search space under the pattern [-4,+2,+2] need not be explored.

The MGDP algorithm operates by iteration, setting each class as the corpus $\oplus$, and setting the rest of the pieces, irrespective of their classes, as the anticorpus $\ominus$ (see Figure 1). Patterns are found within a corpus by tree search over a specified pattern space (e.g., melodic interval patterns, rhythmic patterns) and data mining parameters including a distinctiveness threshold. Statistics are computed for each MGDP found and the results sorted by increasing p-value. For folk songs, the class variable may be used to label any imaginable partitioning of the data, for example, by the genre of a song, its geographic area of collection (and/or origin, if known), or its tune family. The algorithm was applied with success to a corpus of Cretan folk music (Conklin and nagnostopoulou, 2011), which were labeled with 5 toponymic and 11 genre descriptors.

2.3 Antipatterns

An *antipattern* (*anticorpus pattern*) is a pattern that is absent or surprisingly rare within a corpus but occurs frequently in an anticorpus (i.e., is a general rather than a specific pattern) . To discover such patterns it is tempting to try to enumerate the space of patterns in the corpus from most specific (longest) to general (rather than general to specific), but this strategy is inefficient because nearly all conceivable patterns will be infrequent in a corpus. Furthermore, a weakness of this strategy is that it cannot discover jumping patterns (Dong and Li, 1999): those that are completely absent in a corpus.

class $\oplus$	pattern P	$c^{\oplus}(P)$	$n^{\oplus}$	$c^{\ominus}(P)$	p-value
amorosa	$[-3,+2,-4]$	4	247	156	3.9e-05
religiosa	$[+2,-2,-2,-1,-2,-2]$	2	209	128	2.5e-05
danza	$[0,+2,+2,0,0]$	3	494	36	0.0045
infantil	$[-3,-2]$	2	55	577	4.8e-07
de cuna	$[+5,-2]$	4	98	377	3.3e-06
narrativa	$[+9]$	1	86	211	0.00036
festiva-satirica	$[-4,+2,+2]$	4	72	366	0.00063
artaxuriketak	$[-3,-2]$	3	38	576	0.00078
zuberoa	$[-3,-4]$	3	80	301	5.7e-05
gipuzkoa	$[+1,+2,+2,0]$	2	175	118	9.3e-05
nafarroa beherea	$[0,+2,-4]$	0	53	134	0.0097
bizkaia	$[+4,+3]$	0	21	267	0.023

Table 3: Antipatterns in the Cancionero Vasco. Top: for genres; Bottom: for territorios.

An elegant solution is found by noting that there is a natural symmetry to patterns that are over-represented in an analysis class and those under-represented in the background. In fact the MGDP algorithm can naturally be used to discover antipatterns by reversing the roles of corpus and anticorpus. Furthermore, the p-value of an antipattern has a symmetric meaning: the probability that it occurs the observed number *or fewer* times in the corpus. Therefore, by switching the role of corpus and anticorpus, and modifying the p-value computations to compute the left rather than right tail of the cumulative hypergeometric distribution, the MGDP algorithm may be used to discover antipatterns.

3. Results

The 7 territorios and 24 genres in the Cancionero Vasco were used as label dimensions for the discovery of antipatterns. Therefore, for example, one territorio (e.g., bizkaia) would be taken as the corpus $\oplus$ and the songs in other territories as the anticorpus $\ominus$, and the MGDP algorithm

used to find patterns under-represented in the corpus. As described above, this is done by reversing the roles of the corpus and anticorpus, and finding patterns over-represented in the anticorpus.

Table 3 shows some examples of discovered antipatterns, derived from various different classes found within the Cancionero Vasco. Some antipatterns of Table 3 are surprising from a musicological sense, for example the religiosa antipattern which occurs in 128 songs in the anticorpus but only in 2 songs in the corpus. The antipattern represents a long scalic passage (e.g., [G,A,G,F,E,D,C]). The simple antipattern [+9], representing a leap of a major sixth, occurs in only 1 (of 86) narrativa songs, though in 211 pieces in the anticorpus. The antipattern [+4,+3] is a jumping pattern (it occurs in none of the pieces in the bizkaia territory), but despite this the p-value (0.023) is not significant, reflecting the fact that there are only 21 pieces in the corpus, making it statistically not so surprising as some of the other antipatterns. Interestingly, the wedding song pattern [-4,+2,+2] illustrated earlier in Table 2 is at the same time an antipattern for the class festiva-satirica.

4. Conclusions

The results with antipattern discovery are promising and several directions for future work are planned.

The topic of using a collection of patterns for classification was reviewed by Conklin (2009). Distinctive patterns may be used as boolean features as input to standard feature vector classifiers. In this sense, pattern discovery can be viewed as a feature generation problem. Distinctive antipatterns may strongly suggest against membership in a class.

Clearly some way to visualise results are necessary, and for this purpose it is planned to reincorporate discovered patterns back into a formal ontology of classes and patterns, using a description logic formalism encoded in the web ontology language OWL, and a ontology visualisation tool. Antipatterns may be represented as description logic concepts using a method similar to that described by Hirsh and Kudenko (1997).

Though this study has shown that labels for folk songs may be used productively in a pattern discovery setting, in general the labelling of folk songs always raises some questions. The semantics of geographic location labels can be unclear and open to interpretation. In the Cancionero Vasco the labels refer to the place of collection of the tune, which is not necessarily the same as the home area of the performer, or the area where the tune was learned. The genre labels may have an ambiguous relation to song content in cases where the same tune is used for different social functions (Selfridge-Field, 2006).

Antipatterns, those patterns that are rare within an analysis corpus, are arguably even harder to interpret than frequent patterns. This is because one cannot simply highlight the occurrences within a list of pieces that contain the pattern and inspect their musical context. One can inspect the few rare example pieces for obvious wider deviations from the style (or data anomalies) but in cases where the antipattern has a zero corpus count even this method cannot be applied.

Future explorations include the use of antipatterns for motivic analysis of single pieces (Conklin, 2010b) and discovery of antipatterns over different representations of songs, for example at higher structural levels of phrases and sections.

Acknowledgments

The Fundación Euskomedia and Fundación Eresbil are graciously thanked for participating in the project and providing the Cancionero Vasco for study. 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 K. Neubarth and the reviewers for valuable comments on the manuscript.

References

Al-Shahrour, F., Díaz-Uriarte, R., and Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580.
Conklin, D. (2009). Melody *classification using patterns*. In MML 2009: *International Workshop on Machine Learning and Music*, pages 37–41, Bled, Slovenia.
Conklin, D. (2010a). Discovery of distinctive patterns in music. *Intelligent Data Analysis*, 14(5):547–554.
Conklin, D. (2010b). Distinctive patterns in the first movement of Brahms' String Quartet in C Minor. *Journal of Mathematics and Music*, 4(2):85–92.
Conklin, D. and Anagnostopoulou, C. (2011). Comparative pattern analysis of Cretan folk songs. *Journal of New Music Research*, 40(2):119–125.
de Riezu, P. J. (1996). Cancionero vasco P. Donostia. *Revista Internacional de los Estudios Vascos*, 41:189–190.
Dong, G. and Li, J. (1999). Efficient mining of emerging patterns: discovering trends and differences. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pages 43–52.
Falcon, S. and Gentleman, R. (2008). Hypergeometric testing used for gene set enrichment analysis. In Hahne, F., Huber, W., Gentleman, R., and Falcon, S., editors, *Bioconductor Case Studies*, pages 207–220. Springer.
Hirsh, H. and Kudenko, D. (1997). Representing sequences in description logics. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Innovative Applications of Artificial Intelligence Conference*, pages 384–389, Providence, Rhode Island.
Novak, P. K., Lavra˘c, N., and Webb, G. I. (2009). Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10:377–403.
Selfridge-Field, E. (2006). Social cognition and melodic persistence: Where metadata and content diverge. In ISMIR 2006, *7th International Conference on Music Information Retrieval*, pages 272–275, Victoria, Canada.
Taminau, J., Hillewaere, R., Meganck, S., Conklin, D., Nowe, A., and Manderick, B. (2009). Descriptive subgroup mining of folk music. In MML 2009: *International Workshop on Machine Learning and Music*, pages 1–6, Bled, Slovenia.