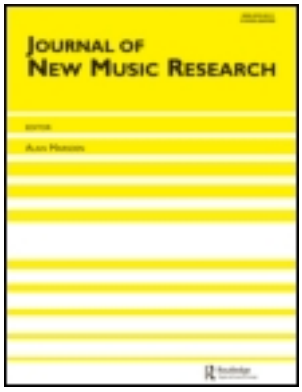


Journal of New Music Research
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/nnmr20

Antipattern Discovery in Folk Tunes

Darrell Conklin [a]
[a] University of the Basque Country UPV/EHU, Spain, and IKERBASQUE, Basque Foundation for Science , Spain
Published online: 20 Aug 2013.

To cite this article: Darrell Conklin (2013) Antipattern Discovery in Folk Tunes, Journal of New Music Research, 42:2, 161-169, DOI: 10.1080/09298215.2013.809125

To link to this article: http://dx.doi.org/10.1080/09298215.2013.809125

Journal of New Music Research, 2013
Vol. 42, No. 2, 161–169, http://dx.doi.org/10.1080/09298215.2013.809125

Routledge
Taylor & Francis Group

Antipattern Discovery in Folk Tunes

Darrell Conklin

University of the Basque Country UPV/EHU, Spain, and IKERBASQUE, Basque Foundation for Science, Spain

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 from a set of pieces, and among those the patterns that are frequent in a background set. Pattern discovery is performed with reference to a background ontology of folk tune genres. The method is applied to a large corpus of Basque folk tunes and results are evaluated as descriptive patterns and as negative association rules.

1. Introduction

In recent years there has been a renewal in folk song analysis, due to increasing interest in cultural heritage and advances in music informatics methods. The ability to correlate music content with metadata 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 work presented in this paper is part of a larger project on computational analysis of folk music, concerned with the development and application of data mining methods to build predictive models and associations between musical content and metadata of songs.

Sequential pattern mining is an active area of research in the general field of machine learning (Agrawal & Srikant, 1995; Ayres, Gehrke, Yiu, & Flannick, 2002; Mabroukeh & Ezeife, 2010). In music, sequential pattern mining has been applied to detect repetition of melodic material for the analysis of a single piece, and also to detect patterns distinctive to a set of pieces (Hsu, Liu, & Chen, 1998; Lartillot, 2004; Liu, Wu, & Chen, 2005; Lo, Lee, & Chang, 2008; Conklin, 2010b; Knopke & Jürgensen, 2009). Pattern distinctiveness may be measured either analytically by assuming some statistical background distribution (Conklin & Anagnostopoulou, 2001; Conklin, 2002, 2006; Honingh, Weyde, & Conklin, 2009) or empirically by using an explicit background set of pieces (Conklin, 2009, 2010a). In music, sequential pattern discovery methods have been used for music analysis (Conklin, 2010b) and also to find patterns that may be used to classify melodies (Sawada & Satoh, 2000; Shan & Kuo, 2003; Lin, Liu, Wu, & Chen, 2004; Conklin, 2009).

Much research to date on pattern discovery in music has been concerned with discovering patterns that are frequent, salient, and over-represented 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. These patterns might lead to musicological hypotheses about genres, in terms of melodic or rhythmic structures that generally occur frequently though not within certain genres. A collection of antipatterns may be used to characterize a class of songs, highlighting the features that are rare or even absent. They might be used, for example, within predictive classification, where their occurrence might strongly suggest against membership in a class. Furthermore, exceptions to strong antipatterns might be worth inspecting for possible erroneous or ambiguous genre labelling. The pattern discovery method developed in this paper is applied to a corpus of Basque folk tunes to reveal patterns that are rare in certain genres. These patterns are evaluated both as *descriptive* patterns, interesting and interpretable within a piece or collection of pieces, and as *predictive* patterns, as rules that predict the negation of a genre from the presence of a pattern.

An area of interest in data mining is the discovery of *negative association rules*, those where the antecedent of a rule probabilistically implies the negation of the consequent (Artamonova, Frishman, & Frishman, 2007; Wu, Zhang, & Zhang, 2004; Antonie & Zaïane, 2004). There has been some attention to *negative sequential patterns*, where one or more of the elements of the pattern, most notably the last element, is a negated feature set (Lin, Chen, Hao, Chueh, & Chang, 2008; Zheng, Zhao, Zuo, & Cao, 2009; Ouyang & Huang, 2007). In this paper a new idea is presented, where a sequential pattern is associated with a genre in which it is rare and under-represented. The other central contribution of this paper is to demonstrate how a background ontology can be used

Correspondence: Darrell Conklin, Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, San Sebastián, Spain, and IKERBASQUE, Basque Foundation for Science, Bilbao, Spain. E-mail: conklin@ikerbasque.org

Table 1. Glossary of notation.

Notation	Meaning
P	a pattern
$\oplus$	analysis class (corpus)
$\ominus$	background (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
$\mathcal{K}$	a knowledge base: a set of concept and instance definitions
$C \sqsubseteq C'$	concept C is subsumed by (implies) concept C'
$C(x)$	x is an instance of concept C
$C \sqcap D$	concept which is the conjunction of concepts C and D
$\mathcal{K} \models x$	statement x is a logical consequence of the knowledge base $\mathcal{K}$

during the pattern discovery process, thereby extending the earlier work of Conklin and Anagnostopoulou (2011) on pattern discovery in folk tune genres.

2. Methods

This section describes the theory leading to a new method for discovering antipatterns in a corpus of folk tunes. First, the corpus of Basque folk tunes is described. Next, a genre ontology for this corpus is presented along with the relevant background in description logics. Then the general methods of pattern discovery are presented, followed by a general method for applying subgroup discovery to sequential patterns in music. Following this, a new method for incorporating an ontology into the pattern discovery process is presented. Finally, it is shown how to find under- rather than over-represented patterns. Table 1 provides a list of the essential notation that is used in this paper.

2.1 Cancionero Vasco

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 published in four volumes (de Riezu, 1996) and digitized, 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 annotations made by Padre Donostia including the region of collection of the song, and its genre. In the total collection of 1902 tunes, 1561 are labelled with a genre, and 1630 are labelled with a toponym organized in levels of region, municipality, and town. These genres and toponyms have been organized into an ontology, using newly created internal classes, to describe hierarchical relations between genres and between toponyms (Goienetxea et al., 2012; Neubarth, Goienetxea, Johnson, & Conklin, 2012). In this paper only the portion containing the genre hierarchy will be described and used for pattern discovery.

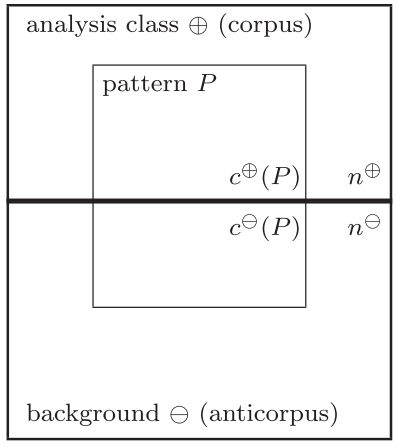


Fig. 1. The schema for subgroup discovery, showing the major regions of objects involved. The top part of the outer box encloses data labelled with the class of interest, below this the background. The inner box contains the objects described by a pattern, and the top part of the inner box the subgroup described by a discovered pattern.

2.2 Folk song ontology

An *ontology* is an encoding of concepts and their relations in a domain of knowledge. Following the terminology of Description Logic (Baader, Calvanese, McGuiness, Nardi, & Patel-Schneider, 2003), a knowledge base $\mathcal{K}$ is divided into two parts, a *TBox* which defines all of the concepts, and an *ABox* which declares instances of these concepts. The TBox contains a set of statements defining concepts and also some subsumption relationships. The ABox contains descriptions of individuals using concepts defined in the TBox. The two main inferences supported by description logic are: *subsumption*, inferring whether the set of instances of one concept are always a superset of another; and *instance checking*, inferring whether an individual is a member of a concept.

For the Cancionero Vasco the genre part of the TBox consists of axioms of the form $G \sqsubseteq G'$ which define the genres and their subsumption relations, for example, the axiom

$$\text{Danzas} \sqsubseteq \text{género}$$

declares the genre of dance songs and

$$\text{Canciones religiosas} \sqsubseteq \text{Canciones de creencias}$$

states that religious songs are a subclass of belief songs. Only direct subsumption relations need be expressed and other relations are inferred as logical consequences of the knowledge base $\mathcal{K}$, for example:

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

Table 2. The genre hierarchy used in the Cancionero Vasco. Numbers in brackets are the inferred song counts of the genres. 1561 of the 1902 songs in the Cancionero Vasco have a genre annotation.

género (1561)		
	Canciones del ciclo histórico (121)	
	Canciones narrativas (86)	
	Canciones de circunstancias (17)	
		Canciones de contrabando (1)
		Canciones de cortesía (6)
		Canciones biográficas (1)
		Canciones de mutil zaharra (9)
	Canciones de guerra-milicia (18)	
		Canciones carlistas (3)
	Canciones de entretenimiento (167)	
	Canciones festivas (82)	
		Canciones satíricas (72)
		Canciones de pelota (10)
	Canciones báquicas (56)	
	Canciones humorísticas (29)	
		Aldrebeskeriak (4)
	Canciones del ciclo vital (477)	
	Canciones cuneras (98)	
	Canciones infantiles (56)	
	Canciones de ronda-cuestación (23)	
	Canciones amorosas (247)	
	Epitalamios (6)	
	Canciones de oficios-de trabajo (46)	
		Artaxuriketak (38)
	Endechas-Elegías (1)	
		Elegías (1)
	Danzas (495)	
	Canciones de creencias (301)	
	Canciones morales (11)	
	Canciones religiosas (290)	
		Canciones de Navidad (81)

$$\mathcal{K} \models \text{Canciones satíricas} \sqsubseteq \text{Canciones de entretenimiento}.$$

The transitive closure of inferred subsumption relations between concept names appearing in TBox axioms can be expressed as a concept taxonomy: Table 2 shows such a taxonomy for the genres of the Cancionero Vasco.

The ABox portion of the knowledge base is a set of assertions *G(x)* where *G* is a genre concept and *x* is an identifier of a song. Thus a statement

$$\text{Canciones satíricas}(2409)$$

asserts that song with Euskomedia code 2409 is a satirical song. Songs are labelled with their original genre annotation as noted by Padre Donostia, which is not necessarily the lowest level of the genre hierarchy. Only one genre is assigned to each song, even though in description logic there is no such restriction. In description logic, instantiations that are not directly asserted are inferred as logical consequences of the knowledge base $\mathcal{K}$, for example:

$$\mathcal{K} \models \text{Canciones de entretenimiento}(2409).$$

Table 2 illustrates in brackets the inferred instance count of every genre. The ontology for the Cancionero Vasco is encoded using the OWL Web Ontology Language (W3C OWL Working Group, 2009), within the Protégé ontology editor and integrated development environment (protege.stanford.edu). These tools facilitate the development and evolution of the ontology, and more importantly, support automated reasoning to infer subclass and instance relationships that are not directly asserted in the knowledge base.

2.3 Supervised descriptive data mining

Supervised data mining methods work with data that are labelled with a special *class* attribute. For folk tunes the label may refer to, for example, the genre of the tune, its geographic area of collection (or origin), or the family of the tune (presumably deriving from the same ancestral tune). For *predictive* approaches the goal is to learn a model that predicts the class with high accuracy on unseen data. For *descriptive* approaches the goal is to group unlabelled data into meaningful clusters.

The intermediate data mining scenario known as *subgroup discovery* or alternatively as *supervised descriptive rule discovery* (Novak, Lavrač, & Webb, 2009) or *contrast data mining* (Dong & Bailey, 2012) is a relatively recent paradigm for data mining. As illustrated in Figure 1, data labelled with multiple classes is partitioned into two groups, an analysis class $\oplus$, and a background $\ominus$ which labels instances of all other

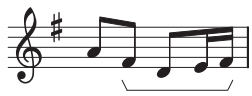
classes. The goal is to discover patterns predictive of the class not for any possible example, but only for interesting subsets (subgroups) among them, covering as few of the $\ominus$ instances as possible. In contrast to supervised predictive methods, subgroup discovery must therefore realize two tasks: identify the interesting subgroups, then (in fact, in parallel) describe them with comprehensible patterns. Thus the methods are 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 $\oplus$ data, that is, they are agnostic about making class predictions for 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 by classification accuracy as in the case of supervised predictive methods.

2.4 Sequential pattern mining

The previous section defined the problem of supervised descriptive data mining. In many settings, including music and bioinformatics, objects may be represented as *sequences*. Sequential pattern mining methods have been developed to find patterns that distinguish one set of sequences from another (Chan, Kao, Yip, & Tang, 2003; Ji, Bailey, & Dong, 2005; Deng & Zaïane, 2009). In bioinformatics, the area of *discriminative motif finding* also fits within the general area of supervised descriptive data mining (Vens, Rosso, & Danchin, 2011; Sumazin et al., 2005). Extending 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 (e.g. melodic intervals, melodic contours, duration ratios). Patterns can be viewed as unary predicates – functions from sequences of events to boolean values – and therefore they have natural interpretations as logical concepts. A piece *instantiates* a pattern if the pattern occurs one or more times in the piece: if the components of the pattern are instantiated by successive events in the piece. For example, $[+2, +1]$ is a sequence of melodic intervals that is instantiated by melodies containing (for example) the note sequences [C, D, E♭] or [D, E, F].

The task of subgroup discovery introduces computational complexities for sequential patterns, because the set of patterns used to describe subgroups may be practically infinite and therefore the search for interesting patterns must be handled carefully. 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, Figure 2 shows a contingency table for the melodic interval pattern $[-4, +2, +2]$ (denoted by P) that occurs in a subgroup of songs of the



pattern $[-4, +2, +2]$		genre Epitalamios: no	genre Epitalamios: yes
	no	1222	1
	yes	$c^{\ominus}(P) = 333$	$c^{\oplus}(P) = 5$
		$n^{\oplus} = 1555$	$n^{\oplus} = 6$

Fig. 2. Top: illustration of an instance of a Basque folk tune pattern (from Euskomedia code 1616). Bottom: a contingency table. The p-value of the association between $[-4, +2, +2]$ and Epitalamios is 0.0023.

genre Epitalamios. The pattern instantiates 83% ($c^{\oplus}(P) = 5$ of $n^{\oplus} = 6$) of wedding songs, but only 21% ($c^{\ominus}(P) = 333$ of $n^{\oplus} = 1555$) of songs from other genres. It may be suggested that this pattern is distinctive of wedding songs, because its relative frequency in that class is higher than in the background.

A one-tailed Fisher's exact test, based on the cumulative hypergeometric distribution (Falcon & Gentleman, 2008), is used to quantify the interest of an association between a pattern and a class. This test measures the probability that the frequency of a pattern in a class departs significantly from its frequency expected by chance. Referring to Figure 1, the p-value of an association is the probability of laying the inner box – drawing a set of $c^{\oplus}(P) + c^{\ominus}(P)$ pieces from $n^{\oplus} + n^{\oplus}$ pieces – and finding $c^{\oplus}(P)$ or more members of the class $\oplus$. Lower p-values indicate more distinctive patterns. 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 p-value (0.0023) for finding five or more wedding songs in a sample of 338 pieces (or, symmetrically, five or more pieces containing the pattern $[-4, +2, +2]$ in a sample of six pieces).

The concept of generality or subsumption is very important to structure the search and presentation space of patterns. A pattern P' *subsumes* (is more general than) a pattern P if all instances of P are also instances of P' in all possible corpora. Following description logic conventions, this relation is written $P \sqsubseteq P'$. A sound and complete algorithm for inferring subsumption between two patterns, which runs in the product of their lengths, simply aligns the two patterns together, while allowing end deletions in the subsumed (but not the subsuming) pattern. For example, the following subsumption relations hold among melodic interval patterns: $[+2] \sqsubseteq []$, $[+2, +1] \sqsubseteq [+2]$, $[+2, +1, +4] \sqsubseteq [+2, +1]$. Figure 3 shows a small subsumption taxonomy (a graph with transitive relations omitted) for illustration of these relations. The algorithm discovers a set of patterns that are distinctive, and among those the most general (not subsumed by any other distinctive pattern). For example, referring to Figure 3, if the pattern $[+2]$ is not distinctive it would not be reported. If the pattern $[+2, +1]$ is distinctive, then no more specific pattern (e.g. $[+2, +1, -3]$, $[+2, +1, +4]$) will be reported, and in

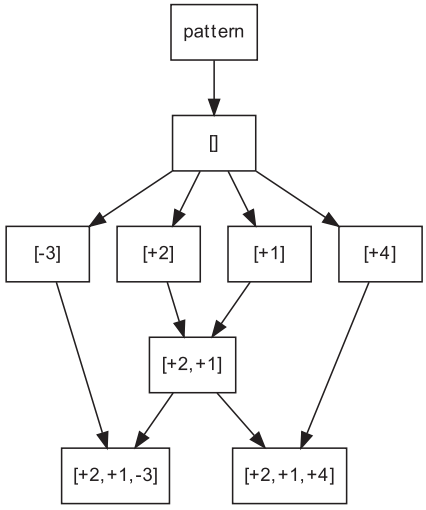


Fig. 3. A small subsumption network of melodic interval patterns.

fact the entire search space under the pattern $[+2, +1]$ need not be explored.

The MGDP algorithm performs a tree search over a specified pattern space (e.g. melodic interval patterns, duration patterns) and data mining parameters including a pattern distinctiveness threshold. Given a dataset containing multiple classes, the process is iterated, setting each class as the corpus and setting the rest of the pieces, irrespective of their classes (but excluding unlabelled songs), as the anticorpus (see Figure 1). As described by Conklin (2010a), the central notion of subsumption applies to any features or sets of features and therefore the MGDP algorithm does not depend on particular event features.

2.5 Pattern mining with ontologies

The notion of subsumption can be naturally extended to accommodate the hierarchical structure within genres, in addition to subsumption between patterns. A pattern/genre association can be considered as a composite concept $P \sqcap G$ which is the logical conjunction of a pattern P and a genre G, for example,

$$[+1, -3] \sqcap \text{Canciones infantiles}.$$

For the inference of subsumption relations between such associations, the following theorem holds

$$\mathcal{K} \models P \sqsubseteq P' \wedge \mathcal{K} \models G \sqsubseteq G' \Leftrightarrow \mathcal{K} \models P \sqcap G \sqsubseteq P' \sqcap G' \quad (1)$$

stating that in any knowledge base $\mathcal{K}$ subsumption between two associations $P \sqcap G$ and $P' \sqcap G'$ can be inferred by inspecting their corresponding pattern and genre components (proof in Appendix A). This is important as it provides a sound and complete decision procedure for testing two associations for subsumption. Figure 4 provides an illustrative example of this type of subsumption, showing four pattern/genre associations and their subsumption relationships. The MGDP method is naturally extended to use an ontology, given the definition of subsumption between associations. Referring to Figure 4, if

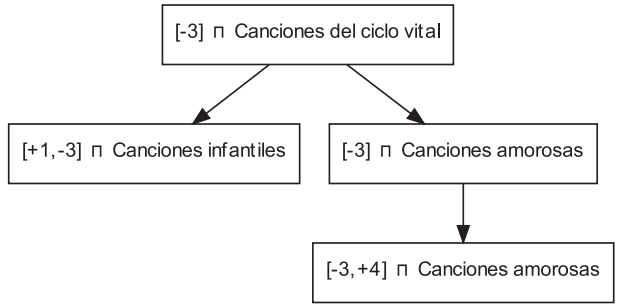


Fig. 4. An example of pattern/genre association subsumption, with genre subsumption based on the ontology of Table 2.

the pattern $[-3]$ is distinctive to the genre Canciones del ciclo vital (top node), then none of the more specific associations will be reported.

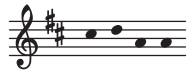
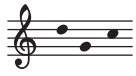
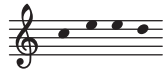
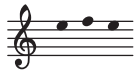
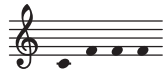
2.6 Antipatterns

An *antipattern* (*anti*corpus *pattern*) is a pattern that is absent or surprisingly rare within the analysis set but occurs frequently in the background. To discover such patterns it is tempting to try to enumerate the patterns occurring in the corpus from most specific (longest) to most general (shortest), rather than enumerating them from general to specific. Though highly specific patterns will indeed have low frequency in a corpus, this strategy is not computationally feasible because nearly all such patterns will be infrequent in a corpus and will therefore need to be visited during the search. Furthermore, a weakness of this strategy is that it cannot discover *jumping* antipatterns (Dong & Li, 1999): those that are completely absent from a corpus.

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 in the observed number *or fewer* pieces 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.

One issue that does not arise for positive patterns, which must be addressed for antipatterns, is the avoidance of antipatterns where the instances in the anticorpus depart significantly from a uniform distribution over classes. In such cases the antipattern could also be a (positive) pattern for just one of the classes in the anticorpus. As an extreme example, a jumping pattern (one that occurs only in one genre) will apparently be a jumping antipattern for any other genre, even though it is more appropriate to report it as a pattern rather than an antipattern. Therefore, to further measure the quality of an antipattern, a χ^2 goodness of fit test may be applied to the distribution of instances in the five top-level genres (bold in Table 2) in the

Table 3. A selection of antipatterns in the Cancionero Vasco. Top: melodic patterns; bottom: duration ratio patterns.

genre $\oplus$	antipattern P	$c^{\oplus}(P)$	$n^{\oplus}$	$c^{\ominus}(P)$	p-value	instance
Canciones del ciclo histórico	$[+1, -5, 0]$	0	121	60	0.007	
Canciones de entretenimiento	$[-7, +5]$	5	167	122	0.004	
Canciones de Navidad	$[+4, 0, -2]$	2	81	151	0.0099	
Canciones infantiles	$[+1, -1]$	13	56	741	7.8e-05	
Canciones de oficios-de trabajo	$[+5, 0, 0]$	2	46	301	0.0032	
Canciones de entretenimiento	$[3:2, 1:3, 1:1]$	6	167	124	0.009	
Canciones festivas	$[2:1, 3:2]$	2	82	162	0.0054	
Canciones infantiles	$[3:1]$	16	56	867	1.5e-05	
Canciones cuneras	$[3:1]$	29	98	854	2.3e-08	

anticorpus. Low p-values indicate skewed distributions and therefore these antipatterns are more reasonably presented as (positive) patterns.

3. Results

In this section the antipatterns method is evaluated in two ways. First, selected descriptive rules from the larger set of discovered antipatterns are presented and discussed. Second, the antipatterns are viewed as negative association rules, predicting the absence of a genre based on the presence of a pattern.

3.1 Individual patterns

The genres (Table 2) of the Cancionero Vasco were used as label dimensions for the discovery of antipatterns. Both melodic interval patterns and duration ratio patterns (e.g. 2:1 indicating an event with duration twice the length of its preceding event) are studied. A hypergeometric p-value threshold of 0.01 for distinctiveness is used, and additionally only antipatterns with a χ^2 goodness of fit (p-value threshold 0.001) to a uniform distribution in the anticorpus are presented.

Within melodic interval patterns, 226 antipatterns are found in the corpus: this is a decrease from the 254 found without considering maximally general patterns according to subsumption (Expression 1). This shows that the use of background ontologies can reduce the discovered pattern set to a more compact one, facilitating interpretation of the pattern set. Table 3 shows some selected discovered melodic interval and duration ratio antipatterns, along with their counts, hypergeometric p-value, and an example instance using concrete pitches or durations. The first melodic interval antipattern $[+1, -5, 0]$ is a jumping antipattern, having no occurrences in the indicated genre. The second antipattern $[-7, +5]$, representing a leap down of a perfect fifth following a leap up of a perfect fourth, is surprisingly rare in the genre Canciones de entretenimiento. Both of these antipatterns are at the highest level of the genre ontology, indicating the ability of the method to find antipatterns that are general in both content and genre. The next two antipatterns $[+4, 0, -2]$ and $[+1, -1]$ are associated at leaf genres of the ontology. The latter antipattern is frequent in the general corpus (occurring in nearly half of all pieces) but in a lower proportion of the indicated children song genre. The last melodic interval example in Table 3 is the antipattern $[+5, 0, 0]$ associated with a mid-level genre.

Within the duration ratio representation, 120 antipatterns are found: this is a decrease from the 132 found without maximal generality. Three are chosen to illustrate top, mid, and leaf level genres associated with antipatterns (Table 3). The general pattern [3:1], representing a note three times the duration of a preceding note, is relatively rare in the genre Canciones infantiles. Interestingly the exact same antipattern is associated with Canciones cuneras or cradle songs. Both patterns have low (significant) p-values.

3.2 Negative association rules

An interesting application of antipatterns arises by viewing them as negative association rules which predict the negation of a genre from the presence of a pattern. Exceptions to negative association rules may indicate possible erroneous or ambiguous labelling of the exceptions (Artamonova et al., 2007). Considering an association where the pattern P is an antipattern for the genre $\oplus$, the *confidence* of a negative

Table 4. A selection of negative association rules, highlighting one exception to the rule.

genre $\oplus$	pattern P	confidence	exception	annotated genre
Canciones de entretenimiento	$[-4, +2, -2]$	115/116	1691	Canciones humorísticas
Canciones humorísticas	$[1:1, 2:1, 1:1]$	941/947	2013	Canciones humorísticas
Canciones de entretenimiento	$[-3, -2, -1]$	77/78	1722	Canciones satíricas
Canciones de entretenimiento	$[1, 2, -2, -1, 1]$	71/72	2725	Canciones báquicas

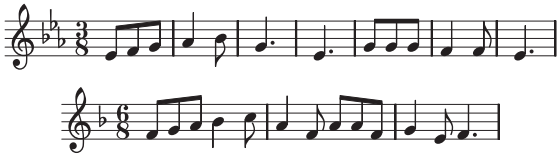


Fig. 5. Top: first phrase of 1722; bottom: first phrase of 1723.

association rule is given by the probability of the negated genre given the pattern:

$$\frac{c^{\ominus}(P)}{c^{\oplus}(P) + c^{\ominus}(P)}. \tag{2}$$

For example (see Table 4), the antipattern $[-4, +2, -2]$ occurs in just one (of 167) instance of the genre Canciones de entretenimiento, and occurs in 115 songs from other genres, giving a high confidence of 115/116 for the negative association rule. The one exceptional instance is in the song *Atso Zarra Belendrin* (Euskomedia code 1691). This song has the same lyrics as another called *Itsas-Dantza* (code 3062) which is labelled as a dance. Another example of an exception is the antipattern $[1:1, 2:1, 1:1]$ which occurs in six (of 29) instances of the genre Canciones humorísticas. One of the six exceptions to the antipattern is the song *Gure bordaltia zabal ordoki* (code 2013). Based on its lyrics, this song is unlikely to be in the class Canciones humorísticas but rather in the class Canciones amorosas.

Table 4 shows two further negative association rules, where the rule has just one exception, given by the song code indicated. The antipattern $[-3, -2, -1]$ matches the song *Baionako patroina* (Euskomedia code 1722) as an exception to the genre Canciones de entretenimiento. Melodic similarity (edit distance of melodic interval sequences) comparison to the entire database revealed the song *Premin Peruarena* (code 1723) as a similar melody (Figure 5 shows the first phrase of each tune). The melody 1723 is unlabelled, though based on an analysis of its lyrics it is more likely to be an instance of Canciones del ciclo vital. As a second example, an exception to the Canciones báquicas antipattern $[1, 2, -2, -1, 1]$ is the song *Hordi eta ezin freska* (code 2725), which is similar to melody *Oroitzen naiz, oroitzen* (code 2192). The melody 2192 is annotated as a member of Canciones amorosas.

In all four of the examples above, antipatterns have highlighted a possible erroneous or additional genre annotation to a song, based on the fact that the song is an exception to a confident negative association rule.

4. Conclusions

This paper has proposed a method for revealing maximally general patterns that are rare in certain genres within an ontology, and has illustrated the method on a corpus of Basque folk tunes. The results with antipattern discovery are promising and several directions for future work are planned. The topic of using pattern sets for music classification has been explored by several researchers (Shan & Kuo, 2003; Lin et al., 2004; Sawada & Satoh, 2000; Conklin, 2009). Distinctive antipatterns may strongly suggest against membership in a class and may be productively incorporated into pattern-based classification methods.

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. 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). Indeed, in the Cancionero Vasco there are cases where similar melodies have different annotated genres.

The results of descriptive data mining can be more difficult to evaluate than for predictive data mining due to lack of a clear performance measure. Antipatterns, those patterns that are rare within an analysis class, 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 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 and discovery of antipatterns over different representations of songs, for example at higher structural levels of phrases and sections.

Acknowledgements

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 Izaro Goienetxea for assistance with ontology building and pattern interpretation.

Special thanks to Kerstin Neubarth and the reviewers for valuable comments on the manuscript.

References

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In P. S. Yu & A. L. P. Chen (Eds.), *Proceedings of the Eleventh International Conference on Data Engineering* (pp. 3–14). Washington, DC: IEEE Computer Society.

Antonie, M., & Zaïane, O. (2004). Mining positive and negative association rules: an approach for confined rules. In J.-F. Boulicaut, F. Esposito, F. Giannotti & D. Pedreschi (Eds.), *Knowledge Discovery in Databases: PKDD 2004, Pisa, Italy* (pp. 27–38). Berlin: Springer.

Artamonova, I., Frishman, G., & Frishman, D. (2007). Applying negative rule mining to improve genome annotation. *BMC Bioinformatics, 8*, art. no. 261. doi:10.1186/1471-2105-8-261

Ayres, J., Gehrke, J., Yiu, T., & Flannick, J. (2002). Sequential pattern mining using a bitmap representation. In D. J. Hand, D. Keim & R. Ng (Eds.), *Proceedings of the International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada* (pp. 429–435). New York: Association for Computing Machinery.

Baader, F., Calvanese, D., McGuiness, D., Nardi, D., & Patel-Schneider, P. F. (Eds.) (2003). *The description logic handbook: Theory, implementation and applications*. Cambridge: Cambridge University Press.

Chan, S., Kao, B., Yip, C. L., & Tang, M. (2003). Mining emerging substrings. *Proceedings of the Eighth International Conference on Database Systems for Advanced Applications, DASFAA '03* (pp. 119–126). Washington, DC: IEEE Computer Society.

Conklin, D. (2002). Representation and discovery of vertical patterns in music. In C. Anagnostopoulou, M. Ferrand & A. Smaill (Eds.), *Music and artificial intelligence: Lecture notes in artificial intelligence 2445* (pp. 32–42). Berlin: Springer-Verlag.

Conklin, D. (2006). Melodic analysis with segment classes. *Machine Learning*, *65*(2–3), 349–360.

Conklin, D. (2009). Melody classification using patterns. In MML, (2009). *International Workshop on Machine Learning and Music*, Bled, Slovenia, pp. 37–41.

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., & Anagnostopoulou, C. (2001). Representation and discovery of multiple viewpoint patterns. In *Proceedings of the International Computer Music Conference, Havana* (pp. 479–485). Cuba: Instituto Cubano de la Musica.

Conklin, D., & 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.

Deng, K., & Zaïane, O. R. (2009). Contrasting sequence groups by emerging sequences. In *Proceedings of the 12th International Conference on Discovery Science, DS '09, Porto, Portugal* (pp. 377–384). Berlin: Springer-Verlag.

Dong, G., & Bailey, J., (Eds.). (2012). *Contrast Data Mining: Concepts, Algorithms, and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. London: Chapman and Hall/CRC.

Dong, G., & 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, San Diego, CA* (pp. 43–52). New York: ACM.

Falcon, S., & Gentleman, R. (2008). Hypergeometric testing used for gene set enrichment analysis. In F. Hahne, W. Huber, R. Gentleman & S. Falcon (Eds.), *Bioconductor case studies* (pp. 207–220). Berlin: Springer.

Goienetxea, I., Arrieta, I., Bagüés, J., Cuesta, A., Leiñena, P., & Conklin, D. (2012). 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. Retrieved from: http://hdl.handle.net/10810/8053

Honingh, A., Weyde, T., & Conklin, D. (2009). Sequential association rules in atonal music. In E. Chew, A. Childs & C.-H. Chuan (Eds.), *MCM 2009: International Conference on Mathematics and Computation in Music* (pp. 130–138). Berlin: Springer.

Hsu, J.-L., Liu, C.-C., & Chen, A. (1998). Efficient repeating pattern finding in music databases. In *Proceedings of the International Conference on Information and Knowledge Management, Bethesda, Maryland* (pp. 281–288). Washington, DC: Association of Computing Machinery.

Ji, X., Bailey, J., & Dong, G. (2005). Mining minimal distinguishing subsequence patterns with gap constraints. In *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining* (pp. 194–201). Washington, DC: IEEE Computer Society.

Knopke, I., & Jürgensen, F. (2009). A system for identifying common melodic phrases in the masses of Palestrina. *Journal of New Music Research*, *38*(2), 171–181.

Lartillot, O. (2004). A musical pattern discovery system founded on a modeling of listening strategies. *Computer Music Journal*, *28*, 53–67.

Lin, C.-R., Liu, N.-H., Wu, Y.-H., & Chen, A. (2004). Music classification using significant repeating patterns. In Y.-J. Lee, J. Li, K.-Y. Whang & D. Lee (Eds.), *Proceedings of the International Conference on Database Systems for Advanced Applications, DASFAA'04, Jeju Island, Korea* (pp. 506–518). Berlin: Springer.

Lin, N., Chen, H., Hao, W., Chueh, H., & Chang, C. (2008). Mining strong positive and negative sequential patterns. *WSEAS Transactions on Computers*, *7*(3), 119–124.

Liu, N.-H., Wu, Y.-H., & Chen, A. L. P. (2005). An efficient approach to extracting approximate repeating patterns in music databases. In L. Zhou, B. C. Ooi & X. Meng (Eds.), *Proceedings of the International Conference on Database Systems for*

Advanced Applications, DASFAA'05, Beijing, China (pp. 240–252). Berlin: Springer.

Lo, Y., Lee, W., & Chang, L. (2008). True suffix tree approach for discovering non-trivial repeating patterns in a music object. *Multimedia Tools and Applications*, *37*(2), 169–187.

Mabroukeh, N. R., & Ezeife, C. I. (2010). *A taxonomy of sequential pattern mining algorithms. ACM Computing Surveys, 43*(1), 3:1–3:41.

Neubarth, K., Goienetxea, I., Johnson, C. G., Conklin, D., & (2012). Association mining of folk music genres and toponyms. In *ISMIR, 2012: 13th International Society for Music Information Retrieval Conference*, Porto, Portugal, pp. 7–12.

Novak, P. K., Lavrač, N., & 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.

Ouyang, W., & Huang, Q. (2007). Mining negative sequential patterns in transaction databases. In *2007 International Conference on Machine Learning and Cybernetics, Hong Kong* (Vol. 2, pp. 830–834). Hoboken, NJ: IEEE.

Sawada, T., & Satoh, K. (2000). Composer classification based on patterns of short note sequences. In *Proceedings of the AAAI-2000 Workshop on AI and Music, Austin, Texas* (pp. 24–27). American Association for Artificial Intelligence : Menlo Park, CA.

Selfridge-Field, E. (2006). Social cognition and melodic persistence: Where metadata and content diverge. In *ISMIR, 2006: 7th International Conference on Music Information Retrieval* (pp. 272–275). Canada: Victoria.

Shan, M.-K., & Kuo, F.-F. (2003). Music style mining and classification by melody. *IEICE Transactions on Information and Systems*, *E88D*(3), 655–659.

Sumazin, P., Chen, G., Hata, N., Smith, A. D., Zhang, T., & Zhang, M. Q. (2005). DWE: Discriminating Word Enumerator. *Bioinformatics*, *21*(1), 31–38.

Vens, C., Rosso, M., & Danchin, E. (2011). Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics*, *27*(9), 1231–1238.

W3C OWL Working Group (27 October 2009). *OWL 2 Web Ontology Language: Document Overview*. W3C Recommendation. Retrieved from: url http://www.w3.org/TR/owl2-overview/.

Wu, X., Zhang, C., & Zhang, S. (2004). Efficient mining of both positive and negative association rules. *ACM Transactions on Information Systems (TOIS)*, *22*(3), 381–405.

Zheng, Z., Zhao, Y., Zuo, Z., & Cao, L. (2009). Negative-GSP: An efficient method for mining negative sequential patterns. In *AusDM '09 Proceedings of the 8th Australasian Data Mining Conference* (pp. 63–67). Sydney, NSW: Australian Computer Society.

Appendix A: Proof of association subsumption

Theorem 1, repeated below, expresses the soundness (forward direction) and completeness (reverse direction) of subsumption inference of two associations $P \sqcap G$ and $P' \sqcap G'$:

$$\mathcal{K} \models P \sqsubseteq P' \wedge \mathcal{K} \models G \sqsubseteq G' \Leftrightarrow \mathcal{K} \models P \sqcap G \sqsubseteq P' \sqcap G'$$

The proof of soundness (the forward direction) is achieved by assuming that the right-hand side is false when the left-hand side is true, and considering any song x such that $P(x)$ and $G(x)$, with the left-hand side implying $P'(x)$ and $G'(x)$, in contradiction to the assumption. □

The proof of completeness (the reverse direction), relies on the additional fact that patterns and genres cannot subsume one another, that is for any pattern P and genre G, $\mathcal{K} \models \neg(P \sqsubseteq G)$ and $\mathcal{K} \models \neg(G \sqsubseteq P)$. Adding this fact, the reverse direction of the theorem becomes

$$\mathcal{K} \models P \sqcap G \sqsubseteq P' \sqcap G' \wedge \neg(P \sqsubseteq G) \wedge \neg(G \sqsubseteq P) \Rightarrow \mathcal{K} \models P \sqsubseteq P' \wedge \mathcal{K} \models G \sqsubseteq G'$$

and the proof is made by considering any song x with its instantiation of the four concepts P, P', G, and G', e.g. one possible instantiation would be $P(x)$, $\neg P'(x)$, $\neg G(x)$, $G'(x)$. A truth table with 16 rows yields the validity of the formula under all possible truth assignments. □