

# Chapter 15

## Contrast Pattern Mining in Folk Music Analysis

Kerstin Neubarth and Darrell Conklin

**Abstract** Comparing groups in data is a common theme in corpus-level music analysis and in exploratory data mining. Contrast patterns describe significant differences between groups. This chapter introduces the task and techniques of contrast pattern mining and reviews work in quantitative and computational folk music analysis as mining for contrast patterns. Three case studies are presented in detail to illustrate different pattern representations, datasets and groupings of folk music corpora, and pattern mining methods: subgroup discovery of global feature patterns in European folk music, emerging pattern mining of sequential patterns in Cretan folk tunes, and association rule mining of positive and negative patterns in Basque folk music. While this chapter focuses on examples in folk music analysis, the concept of contrast patterns offers opportunities for computational music analysis more generally, which can draw on both musicological traditions of quantitative comparative analysis and research in contrast data mining.

### 15.1 Introduction

In his introduction to computational and comparative musicology, Cook (2004) outlined the potential of computational approaches to analysing large repertoires of music, and proclaimed an opportunity for re-evaluating comparative analysis in musicology. For ethnomusicology, Nettle (2005, 2010) re-assessed comparative

---

Kerstin Neubarth  
Canterbury Christ Church University, Canterbury, UK  
e-mail: kerstin.neubarth@canterbury.ac.uk

Darrell Conklin  
Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, San Sebastián, Spain  
IKERBASQUE, Basque Foundation for Science, Bilbao, Spain  
e-mail: darrell.conklin@ehu.eus

research, including quantified comparison (Nettl, 1973, 1975), as a methodological option among others rather than a defining feature of the discipline. Quantitative comparisons between groups or across time—based on music corpora, bibliographic data or compilations of context information—can support research on, for example, composers' and national styles (Trowbridge, 1986; VanHandel, 2009), a composer's choices (Lampert, 1982), a performer's repertoire selection (Kopiecz et al., 2009), or changes in musical taste, music practice and its social, political, economic or technological context (Alessandri et al., 2014; Carter, 1987; Forrest and Heaney, 1991; Hess, 1953; Rose and Tuppen, 2014). In many cases, recent studies can draw on, and are confronted with, larger datasets than their forerunners (e.g., Forrest and Heaney, 1991; Rose and Tuppen, 2014). Data mining provides concepts and methods for organizing and analysing large datasets, discovering underlying relations in data and describing interesting patterns (Klösgen, 1999; Witten et al., 2011). *Contrast data mining* focuses on finding differentiating characteristics between groups in labelled data or trends in time-stamped data (Bay and Pazzani, 2001; Dong and Li, 1999; Webb et al., 2003). This chapter introduces concepts and methods of contrast data mining and illustrates their application to music with examples from folk music analysis.

Computational analysis of folk music has been referenced in the context of computational and empirical musicology (Cook, 2004; Lincoln, 1970, 1974), modern methods for musicology (Marsden, 2009) and digital humanities (Fujinaga and Weiss, 2004), and folk music corpora have attracted attention in music information retrieval (Cornelis et al., 2010; Tzanetakis et al., 2007; van Kranenburg et al., 2010). From a data mining point of view, folk music has proven a fruitful domain for exploring, developing and testing computational methods in corpus-level analysis thanks to the availability of relatively large, coherent, musicologically curated and annotated digital music collections. From a musicological point of view, ethnomusicologists have long explored computational approaches for organizing (Elscheková, 1966; Suchoff, 1967, 1968), indexing (Hoshovs'kyj, 1965; Járdányi, 1965), analysing (Elscheková, 1966, 1999; Suchoff, 1971) and better understanding (Elscheková, 1965, 1966) folk music collections and repertoires. The potential of computational methods is seen in facilitating the fast, accurate and reliable processing of large amounts of data (Csébfalvy et al., 1965; Elscheková, 1965, 1999; Járdányi, 1965; Rhodes, 1965; Steinbeck, 1976; Suchoff, 1968), supporting flexible search of folk music collections (Járdányi, 1965; Steinbeck, 1976; Suchoff, 1971), enhancing transparency of the analysis (Elscheková, 1999; Jesser, 1991), preserving analytical data (Elscheková, 1966) and enabling the discovery of hidden patterns in folk music corpora (Keller, 1984; Suchoff, 1970).

Comparative analysis of folk music has investigated acoustic, stylistic, functional or behavioural traits in folk music and their convergence, distribution or variation, both heuristically and speculatively (Bohlman, 1988; Nettl, 2005, 2010; Schneider, 2006). The analytical interest in finding differences between repertoires and practices within folk music corpora using statistical and computational methods is reflected in research questions such as those suggested by one of the pioneers of computational folk music analysis:

*Historical Questions.* [...] Are there different habits and preferences in melodic range and mode at different periods of history, and what is their relative strength?

*Geographical Questions.* What are the characteristic differentiae of specific regions? [...]

*Typological Questions.* What are the prevailing melodic forms within a given area of study?

(Bronson, 1959, p. 165)

Quantitative comparisons underlie observations on contrasting features such as:

A strange contrast [of songs by the Yuman and Yaqui] to all tribes previously analysed is shown in the relative proportion of songs ending on the third and fifth above the keynote [...]. The percentage ending on the keynote is smaller than in the total number of songs previously analysed. This is a peculiarity of this group of Indians [...].

(Densmore, 1932, p. 38)

We can conclude that organization and general war songs are low, rapid, and of wide range. By contrast the love songs tend to be high, slow, and of medium range.

(Gundlach, 1932, p. 138)

In contrast to German melodies, Chinese songs hardly ever start with an upbeat [...].

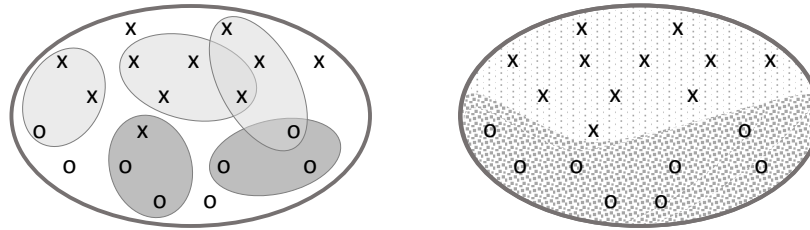
(Schaffrath, 1992, p. 108)

The average range of at least an eleventh [in Scottish melodies] is rather impressive. [...] In contrast to this, folksongs of the Shetlands seem to have less important [sic] ranges.

(Sagrillo, 2010, [p. 8])

Contrast data mining provides a coherent framework for relating early and more recent work in quantitative and computational folk music analysis.

More specifically, contrast data mining is the task of identifying significant differences between groups in data. In this chapter we focus on contrast data mining of folk music as a form of supervised descriptive pattern discovery (Herrera et al., 2011; Novak et al., 2009b). *Supervised* data mining is applied to labelled data instances: contrast data mining discovers contrasting characteristics of selected subpopulations in the data which are identified by group labels. In this respect supervised contrast mining differs from *unsupervised* techniques (e.g., clustering), in which groups are not predetermined but are identified during the mining. Supervised *descriptive* pattern discovery is primarily interested in finding individual rules which describe groups by characteristic local patterns: discovered contrast patterns make statements about parts of the data space (Hand et al., 2001), patterns tolerate some counter-examples (Lavrač et al., 2004), and patterns may overlap, describing different aspects of the same data instances (Klösgen, 1999; Lavrač et al., 2004). Exhaustive algorithms find all interesting patterns; heuristic algorithms apply strategies to reduce the search space, resulting in a subset of possible patterns (Herrera et al., 2011; Novak et al., 2009b). Descriptive patterns ideally are relatively simple and understandable (Herrera et al., 2011; Klösgen, 1996). In *predictive* data mining, on the other hand, induced models should be complete (i.e., cover all instances in the dataset) and ideally should be consistent (i.e., predict the correct group label for all instances) (Fürnkranz et al., 2012). Resulting models may be complex and possibly intransparent (Klösgen, 1996). Predictive methods generally infer one model out of a family of potential models

**supervised descriptive**

group labels  
 local patterns covering subsets of groups  
 patterns evaluated by interestingness  
 comprehensible patterns  
 exhaustive or heuristic search

**supervised predictive**

group labels  
 global models covering all examples of groups  
 models evaluated by predictive accuracy  
 potentially intransparent models  
 heuristic search

**Fig. 15.1** Schematic view and summary of supervised descriptive vs. predictive data mining

which best fits the complete dataset according to chosen heuristics (Hand et al., 2001). Figure 15.1 summarizes the fundamental differences between supervised descriptive and predictive data mining. The diagram constructs a small artificial data mining scenario, showing a dataset organized into two groups (labelled by x: 11 examples; and o: 8 examples). For the descriptive schema, solidly shaded areas refer to individual rules: three rules describing x examples (light regions) and two rules describing o examples (dark shading). For the predictive task the hatched areas represent global models constituted by sets of rules. Together the rules in a set provide a global representation of a group (Witten et al., 2011); individual rules can be difficult to interpret in isolation (Bay and Pazzani, 2001; Lavrač et al., 2004).

This chapter offers three main contributions to the area of computational music analysis. First, it generalizes the concept of pattern in inter-opus and corpus-level music analysis beyond melodic and polyphonic patterns: contrast patterns are primarily defined by their ability to distinguish groups of pieces within a music corpus. Two possible representations of contrast patterns are considered in detail: sequential patterns which capture succession relations between event features, and global feature patterns which describe music pieces by unordered sets of global features. Second, this chapter revisits existing research in quantitative folk music analysis spanning a century—from early work, even predating computational methods, through to modern approaches—and shows how this work can be viewed as contrast pattern mining. Contrast pattern mining provides a vocabulary which can highlight both shared analysis interests and approaches but also different choices in the methodological design. Thus, current and future research in computational music analysis can draw on, possibly critically, both musicological experience and the substantial existing research on theory, methods and algorithms for data mining. Third, this chapter shows how

---

**Contrast pattern mining task**

Contrast pattern mining is the task of discovering and describing patterns that differentiate groups in data.

**Given:**

- a dataset with  $N$  instances
- a target attribute which partitions the dataset into groups
- a pattern description language
- an evaluation measure and threshold

**Discover:** patterns which distinguish a group from other groups, returning patterns whose evaluation measure value is above some threshold, or a specified number of patterns ranked highly by the evaluation measure

---

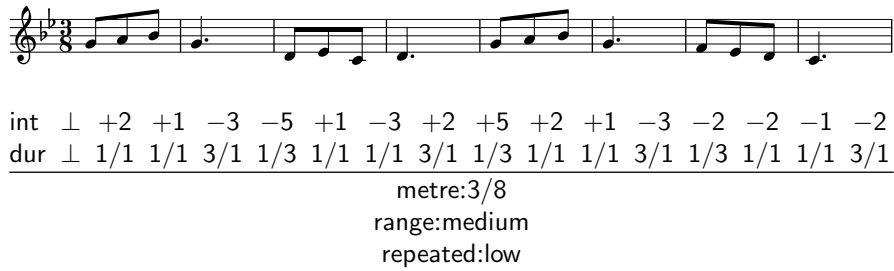
**Fig. 15.2** Definition of contrast pattern mining, adapting the definition of local pattern mining by Zimmermann and De Raedt (2009)

the concept of subsumption (Baader et al., 2003), the logical specialization relation between patterns, applies equally to global feature and sequential patterns. This provides a basis for navigating the search space during pattern discovery, and also for organizing and presenting the results of contrast pattern mining.

The task and terminology of contrast pattern mining are defined in Sect. 15.2. In the subsequent sections contrast pattern mining is applied to comparative analyses of folk music: Sect. 15.3 offers a systematic overview of existing work as contrast pattern mining; Sect. 15.4 presents three case studies illustrating different pattern representations, folk music corpora and contrast mining methods. Section 15.5 briefly looks beyond contrast pattern mining to other comparative approaches and looks ahead to possible directions for future work.

## 15.2 Contrast Pattern Mining

A recurring theme in exploratory data analysis is that of determining differences between groups. In inferential statistics, this is done by studying different samples and determining whether they significantly differ in their distribution of one or more variables. In contrast pattern mining (see Fig. 15.2 for a definition of the task) the aim is to find local patterns which capture differences between groups of data. This section introduces the contrast pattern mining task and particularly looks at how patterns are represented. The relevant notation and concepts of contrast pattern mining are introduced, and three contrast mining methods are reviewed. This theoretical background provides the context for rephrasing work in folk music analysis as contrast pattern mining in Sects. 15.3 and 15.4.



**Fig. 15.3** Lullaby “Itsasoan laño dago” (excerpt) from the *Cancionero Vasco*. Top: score fragment. Middle: examples of event feature representation; viewpoints refer to melodic interval in semitones (int) and duration ratio (dur). These features are undefined ( $\perp$ ) for the first event (Conklin and Witten, 1995). Bottom: examples of global features, numeric features discretized; the abbreviated attribute name repeated refers to the fraction of notes that are repeated melodically (McKay, 2010)

### 15.2.1 Patterns

In contrast pattern mining of music, *patterns* are predicates that map music pieces to boolean values. In this chapter, two types of pattern representation are considered: global feature patterns and sequential patterns based on event features. Here a *feature* is an attribute–value pair. A *global feature* represents a piece by a single value (see Fig. 15.3 bottom). A *global feature pattern* is a set of features, representing the logical conjunction of the features in the set. A pattern is *supported* by a piece of music in the corpus if all features in the feature set are true of the piece. Global features can be explicit in the metadata annotations of pieces (e.g., for attributes such as region, genre, tune family, collector), or can be derived directly from the score (e.g., for attributes such as average melodic interval, range). In classic contrast data mining, pattern descriptions are based on categorical attributes and continuous attributes are discretized, either in a pre-processing step (e.g., Bay, 2000; Kavšek and Lavrač, 2006) or dynamically during the data mining (e.g., Srikant and Agrawal, 1996). A *sequential pattern*, on the other hand, is a sequence of *event features* (see Fig. 15.3 middle): attribute–value pairs over contiguous events. Event features can be numeric (e.g., intervals, durations), categorical (e.g., contour, chord types) or binary (e.g., contour change, in scale/not in scale). Sequential patterns are by definition derived directly from the score. A piece *supports* a sequential pattern if the pattern occurs at least once within the piece. For both global feature patterns and sequential patterns, the absolute support or *support count* of a pattern  $X$ , denoted by  $n(X)$ , is the number of pieces in the dataset supporting the pattern.

	$G$	$\neg G$	
$X$	$n(X \wedge G)$	$n(X \wedge \neg G) = n(X) - n(X \wedge G)$	$n(X)$
$\neg X$	$n(\neg X \wedge G) = n(G) - n(X \wedge G)$	$n(\neg X \wedge \neg G) = n(\neg X) - n(\neg X \wedge G)$	$n(\neg X) = N - n(X)$
	$n(G)$	$n(\neg G) = N - n(G)$	$N$

**Fig. 15.4** Contingency table describing all relationships between a pattern  $X$  and a group  $G$

### 15.2.2 Contrast Patterns

As a supervised data mining technique, contrast pattern mining requires a dataset to be partitioned into labelled groups. Intuitively, groups arise from the values of a target attribute (see Fig. 15.2). For example, folk tunes may be grouped by their function into several genres, such as lullabies, wedding songs or laments. A piece in the dataset supports a group  $G$  if it is a member of group  $G$ . The number of pieces in the dataset supporting a group  $G$  gives the support count of the group,  $n(G)$ .

A pattern is a *contrast pattern* if its support differs significantly between groups in a dataset. The support count of pattern  $X$  in a group  $G$ , denoted by  $n(X \wedge G)$ , is the number of pieces in the dataset supporting both the pattern  $X$  and the group  $G$ . To assess whether or to what extent a pattern distinguishes a group from other groups, an evaluation measure compares the support of the pattern in the different groups. Many evaluation measures have been proposed, based on notions of, for example, generality, reliability, conciseness, peculiarity, surprisingness or utility (Geng and Hamilton, 2006). Contrast mining techniques commonly consider reliability (strength of the relation between a pattern and a group), generality (proportion of data instances supporting a pattern) and sometimes conciseness (simplicity of the description). Evaluation measures are used to prune the search space during the mining process, to filter or rank rules in a post-processing phase, or to provide additional information when presenting results.

Evaluation measures are usually computed from the  $2 \times 2$  contingency table which summarizes the occurrence of a pattern in a specific group of interest  $G$  against other groups (see Fig. 15.4): the marginal counts  $n(X)$  and  $n(G)$  refer to the support counts of pattern  $X$  and group  $G$ . The variable  $N$  indicates the total number of pieces in the dataset. The notations  $\neg X$  and  $\neg G$  denote the complements of pattern  $X$  and group  $G$ : the pieces not supporting  $X$  and  $G$  respectively. The inner cells of the contingency table contain the support counts for pairwise conjunctions of  $X$ ,  $\neg X$ ,  $G$  and  $\neg G$ . If  $n(X \wedge G)$ ,  $n(X)$ ,  $n(G)$  and  $N$  are known all other counts can be derived. From the absolute counts empirical probabilities are calculated as  $P(X) = n(X)/N$ ,  $P(G) = n(G)/N$  and  $P(X \wedge G) = n(X \wedge G)/N$ , and conditional probabilities are derived as  $P(X|G) = P(X \wedge G)/P(G)$  and  $P(G|X) = P(X \wedge G)/P(X)$ . Statistical tests, such as Fisher's exact test, assess observed counts in the inner cells of the contingency table against expected counts based on the pattern and group distribution across the full corpus reflected in the marginal counts: the lower the  $p$ -value calculated by the test the less likely are the observed counts.

Relations between contrast patterns and the groups they characterize can be expressed as *rules*, directed relations between a pattern  $X$  and a group  $G$ :  $X \rightarrow G$  (e.g., Novak et al., 2009b). The left-hand side of the rule is called the rule antecedent, the right-hand side of the rule is called the rule consequent. *Positive rules* describe patterns which are frequent or over-represented in a group: a rule  $X \rightarrow G$  generally captures that pieces supporting pattern  $X$  tend to be members of group  $G$  and thus group  $G$  may be distinguished from other groups by a high proportion of pieces supporting pattern  $X$ . Patterns which are infrequent, under-represented or even absent in a group can be expressed as *negative rules*. Several formalizations of negative rules exist, depending on whether negation is applied to the rule antecedent or consequent as a whole, to attribute–value pairs within patterns or to the implication between antecedent and consequent (Cornelis et al., 2006; Savasere et al., 1998; Wong and Tseng, 2005). In this chapter only negative rules with negated consequent,  $X \rightarrow \neg G$ , are considered. An intuitive interpretation of a rule  $X \rightarrow \neg G$  is that a pattern  $X$  tends to be found in pieces outside of group  $G$  and thus is rare or even absent in group  $G$ .

### 15.2.3 Methods for Contrast Pattern Mining

Specific methods of contrast pattern mining include subgroup discovery (Klösgen, 1996), emerging pattern mining (Dong and Li, 1999) and contrast set mining (Bay and Pazzani, 2001). At times, methods have been adapted from one contrast mining task to another, for example, subgroup discovery to perform contrast set mining (Novak et al., 2009a) or association rule mining to perform subgroup discovery (Kavšek and Lavrač, 2006). This section briefly summarizes three representative methods for discovering contrast patterns; examples of their application in folk music analysis will be presented in Sect. 15.4.

**Subgroup Discovery** The formulation of subgroup discovery is generally traced back to Klösgen (1996), although the term only appears in later publications (e.g., Klösgen, 1999; Wrobel, 1997). Here subgroup discovery is defined as the task of finding subgroups in a dataset which exhibit distributional unusualness with respect to a given target attribute. An additional condition requires subgroups to be sufficiently large. Several evaluation measures have been proposed, which trade off unusualness and generality of subgroups (Klösgen, 1996; Wrobel, 1997); the case study presented in Sect. 15.4.1 below uses *weighted relative accuracy*:

$$WRAcc(X \rightarrow G) = P(X) \times [P(G|X) - P(G)]. \quad (15.1)$$

The first term, coverage  $P(X)$ , measures the generality of the pattern; the second term, relative accuracy or added value  $P(G|X) - P(G)$ , measures the reliability of the rule  $X \rightarrow G$  as the gain between the probability of group  $G$  given pattern  $X$  and the default probability of group  $G$ . Subgroup discovery performs a one-vs-all comparison, in which data instances supporting a target group  $G$  are considered positive examples



and all other instances are considered negative examples, corresponding to a  $2 \times 2$  contingency table with columns indexed by  $G$  and its complement  $\neg G$  (see Fig. 15.4).

**Emerging Pattern Mining** Emerging patterns are conjunctions of global features (Dong and Li, 1999) or sequential patterns (Chan et al., 2003), whose support increases significantly from one dataset (or group) to another. In its original formulation (Dong and Li, 1999), emerging pattern mining corresponds to a one-vs-one comparison and can be represented in a  $2 \times 2$  contingency table with columns indexed by two groups  $G$  and  $G'$ . A contrast between the two groups is measured as the *growth rate* of a pattern  $X$ :

$$\text{GrowthRate}(X, G, G') = \frac{P(X|G)}{P(X|G')} \quad \text{with } P(X|G) > P(X|G'). \quad (15.2)$$

A pattern  $X$  is considered an emerging pattern if its growth rate is above a user-defined threshold  $\theta$  (with  $\theta > 1$ ). Compared to weighted relative accuracy in subgroup discovery, growth rate in emerging pattern mining does not take into account the generality of a pattern: emerging pattern mining focuses on the change in relative support from group  $G'$  to group  $G$ , while the absolute support levels can be low (Dong and Li, 1999).

**Association Rule Mining** An association rule (Agrawal and Srikant, 1994) is a rule of the form  $A \rightarrow B$ , where  $A$  and  $B$  can be sets of attribute–value pairs. In *class association rule mining*, the consequent of the rule is restricted to a class or group in the dataset (Ali et al., 1997; Liu et al., 1998); then an association rule between a pattern  $X$  and a group  $G$  is of the form  $X \rightarrow G$ . The reliability of an association rule is generally evaluated by rule *confidence*:

$$c(X \rightarrow G) = P(G|X). \quad (15.3)$$

The generality of the rule is captured by its relative support,  $s(X \rightarrow G) = P(X \wedge G)$ . Support and confidence are computed from a  $2 \times 2$  contingency table with columns indexed by  $G$  and  $\neg G$  (see Fig. 15.4), thus comparing one group ( $G$ ) against all other groups ( $\neg G$ ). The task of association rule mining consists of finding all rules which meet user-defined support and confidence thresholds.

The methods summarized above differ mainly in their task or comparison strategy and in the evaluation measure used to assess candidate contrast patterns. Emerging pattern mining originally compares two groups by a one-vs-one strategy, while subgroup and class association rule discovery translate a multigroup mining task into a series of one-vs-all comparisons. In emerging pattern mining, growth rate builds on sensitivity  $P(X|G)$  to evaluate the distribution of a pattern in the two groups; association rule mining uses confidence  $P(G|X)$  to assess the relation between a pattern and a group, and weighted relative accuracy in subgroup discovery integrates added value  $P(G|X) - P(G)$  to measure rule reliability. Relative support  $P(X \wedge G)$  in association rule mining and pattern coverage  $P(X)$  as part of weighted relative accuracy in subgroup discovery also consider the generality of potential contrast

patterns. At an algorithmic level, implementations of these methods may differ in the search and pruning strategies employed to generate candidate contrast patterns and to filter redundant patterns, and in statistical techniques used to control false positives or false negatives (e.g., Atzmüller, 2015; Novak et al., 2009b; Webb et al., 2003).

### 15.3 Applications in Folk Music Analysis

Using the criteria and terminology introduced in the previous section, Table 15.1 summarizes 15 selected studies which analyse folk music corpora for contrasts between groups. The first nine of the listed studies use global feature representations; the remaining six studies mine for contrasting sequential patterns. The table includes both quantitative analyses which extract support counts of global feature or sequential patterns in different groups but do not explicitly quantify the contrast (Densmore, 1913, 1918, 1929; Edström, 1999; Grauer, 1965), and studies which directly adopt contrast data mining methods such as subgroup discovery (Taminau et al., 2009) and constrained association rule discovery (Neubarth et al., 2012, 2013a,b), or explicitly relate their method to emerging pattern mining or supervised descriptive rule discovery (Conklin, 2009, 2010a, 2013; Conklin and Anagnostopoulou, 2011).

**Datasets** The folk music corpora used by the cited studies range from regional repertoires through corpora covering larger areas to diverse styles across different continents: Cretan folk music (Conklin and Anagnostopoulou, 2011) and Basque folk music (Conklin, 2013; Neubarth et al., 2012, 2013a,b); European folk music (Neubarth et al., 2013b; Taminau et al., 2009) and North-American folk music (Densmore, 1913, 1918, 1929); or regional and cultural styles from around 250 areas across the world (Grauer, 1965; Lomax, 1962). Most regionally defined corpora represent a variety of folk music genres; on the other hand, Anagnostopoulou et al. (2013) focus on children's songs, and the European folk music corpus used in Taminau et al. (2009) and Neubarth et al. (2013b) is largely dominated by dance genres. The listed studies generally consider complete tunes, with two exceptions: Anagnostopoulou et al. (2013) take tune segments as data instances (505 segments derived from 110 tunes), and Edström (1999) extracts rhythmic patterns from the first four bars of refrains.

**Groups** Groupings in quantitative and computational folk music analyses often refer to geographical regions, ethnic groups and folk music genres or functions. The folk music styles suggested by Lomax (1959) and referenced in later analyses (Grauer, 1965; Lomax, 1962) are to some extent mapped onto geographical or cultural areas, such as Western European song style. Edström (1999) compares Swedish and German foxtrots in the context of constructing Swedishness. Regarding the analyses by Densmore, Table 15.1 includes both analyses of the song repertoires of different Native American tribes (Densmore, 1929) and of folk music genres among the songs of a tribe (Densmore, 1913). The third cited study by Densmore compares old and comparatively new songs within the music of the Teton–Sioux Indians (Densmore, 1918).

**Table 15.1** Contrast analysis of folk music: example studies. Top: studies using global feature representations. Bottom: studies using event feature representations

Study	Dataset		Groups		Description		Contrast mining		
	Repertoire	No.	Target attr.	No.	Attributes	No.	Strategy	Measure	Rules
Densmore 1913	North American	340	genre	10	content	18 <sup>g</sup>	one-vs-one	[narrative]	–
Densmore 1918	North American	600	style	2	content	18 <sup>g</sup>	one-vs-one	[narrative]	–
Densmore 1929	North American	1072	tribe	6	content	13 <sup>g</sup>	one-vs-all	[narrative]	–
Lomax 1962	cross-cultural	n.s.	style	5 <sup>e</sup>	perform.	37	one-vs-one	[visual]	–
Grauer 1965	cross-cultural	1700	style	4 <sup>e</sup>	perform.	37	one-vs-all	[narrative]	–
Taminau et al. 2009	European	3470	region	6	content	150	one-vs-all	WRAcc	pos
Neubarth et al. 2012	Basque	1902	genre <sup>h</sup>	31	region	272	one-vs-all	confidence	pos, neg
Neubarth et al. 2013a	Basque	1902	genre <sup>h</sup>	31	content	17	one-vs-all	confidence	pos, neg
Neubarth et al. 2013b	Basque	1902	genre	5	content	19	one-vs-all	confidence	pos, neg
	European folk dances	3367	region	7	content	19	one-vs-all	confidence	pos, neg
			genre	9	content	19	one-vs-all	confidence	pos, neg
			region	6	content	19	one-vs-all	confidence	pos, neg
Edström 1999	European	n.s.	nation	2	content	1	one-vs-one	[narrative]	–
Conklin 2009	European	195	region	2	content	5	one-vs-all	confidence	pos
Conklin 2010a	European and Asian	432	region	3	content	9	one-vs-all	growth rate	pos
Conklin & Anagnostopoulou 2011	Cretan	106	genre <sup>h</sup>	13	content	1	one-vs-all	growth rate	pos
Anagnostopoulou et al. 2013	European children's songs	505 <sup>s</sup>	region <sup>h</sup>	7	content	1	one-vs-all	growth rate	pos
Conklin 2013	Basque	1902	genre <sup>h</sup>	31	content	3	one-vs-all	growth rate	pos
						2	one-vs-all	<i>p</i> -value	neg

n.s. = not specified. <sup>s</sup> = tune segments. <sup>h</sup> = hierarchically structured attribute. <sup>e</sup> = examples reported. perform. = performance style. <sup>g</sup> = only global attributes counted.

**Description** To characterize groups within the datasets, the studies listed in Table 15.1 make use of metadata (Neubarth et al., 2012), global music content features which are extracted manually (Densmore, 1913, 1918, 1929) or automatically (Neubarth et al., 2013a,b; Taminou et al., 2009), or descriptors referring to the performance style of songs (Grauer, 1965; Lomax, 1962). Sequential patterns are either derived by computing the event feature sequence for predefined segments (Anagnostopoulou et al., 2013; Edström, 1999) or by discovering patterns of flexible length as part of the contrast mining process (Conklin, 2009, 2010a, 2013; Conklin and Anagnostopoulou, 2011). Many of the cited studies analyse one global or event feature at a time. The application of subgroup discovery to European folk music by Taminou et al. (2009) allows flexible conjunctions of two attribute–value pairs, while Grauer (1965) determines a fixed combination of four attribute–value pairs by inspecting individual songs of the target style; in a second step Grauer then considers the remaining 33 attributes for the covered songs. The study by Lomax (1962) presents style profiles using the complete set of 37 descriptor attributes, from which candidates for contrasting attributes can be suggested. Some sequential pattern studies extract several event features but treat each of these separately (Anagnostopoulou et al., 2013; Conklin, 2013); on the other hand, two of the listed analyses (Conklin, 2009, 2010a) mine for patterns using multiple features.

**Contrast Mining** Table 15.1 indicates the primary evaluation measure that the listed studies apply in the comparison. Analyses adopting contrast mining techniques, or explicitly referring to contrast data mining, use measures common in these techniques: weighted relative accuracy in subgroup discovery (Taminou et al., 2009), confidence in constrained association rule mining (Neubarth et al., 2012, 2013a,b), or growth rate from emerging pattern mining (Conklin, 2010a; Conklin and Anagnostopoulou, 2011). Conklin (2013) evaluates the  $p$ -value computed with Fisher’s exact test to assess candidate patterns. Cited earlier studies consider occurrences of patterns in different groups, but the comparison itself is mainly narrative (Densmore, 1913, 1918, 1929; Edström, 1999) or to some extent visual (Lomax, 1962). Occasionally Densmore’s textual description uses phrasings corresponding to growth rate (not illustrated in Table 15.1), for example: “The percentage of songs of a mixed form is more than twice as great in the Ute as in the Chippewa and Sioux” (Densmore, 1922, p. 53). Where group counts are included (Densmore, 1913, 1918, 1929), evaluation measures may be calculated post hoc (Neubarth, 2015). Most of the listed studies follow a one-vs-all strategy in comparing pattern distributions between groups. The publications by Densmore represent different comparison strategies: the analysis of Teton–Sioux music (Densmore, 1918) contrasts two chronologically ordered repertoires—old and relatively modern songs of the Teton–Sioux—presented as Group I and Group II (one-vs-one comparison); features of Pawnee music (Densmore, 1929) are presented against the cumulative support for the comparator groups (one-vs-all comparison); in the analysis of different genres among Chippewa music (Densmore, 1913) all groups are listed. Different comparison strategies, applied to the same dataset, may result in different contrast patterns (Neubarth, 2015).

## 15.4 Case Studies

In this section, three case studies will be presented in some detail to illustrate the different contrast mining methods applied to folk music. The first case study describes patterns by global features and discovers contrasting patterns as subgroups (Taminau et al., 2009). The second case study uses an event feature representation; candidate sequential patterns are evaluated as emerging patterns (Conklin and Anagnostopoulou, 2011). The third case study draws on two publications which apply constrained association rule mining to discover not only positive but also negative rules in folk music data (Conklin, 2013; Neubarth et al., 2013b); both global feature patterns and sequential patterns are considered.

### *15.4.1 Case Study 1: Subgroup Discovery in European Folk Music*

The analysis of European folk music by Taminau et al. (2009) identifies global feature patterns which distinguish between folk songs of different geographical origin, through subgroup discovery. The authors explicitly set out to explore descriptive rule learning as an alternative approach to predictive classification, in order to find interpretable patterns. The following sections summarize the dataset, outline the data mining method and relate discovered subgroups.

#### **15.4.1.1 Dataset and Global Feature Representation**

The studied folk music corpus, called *Europa-6* (Hillewaere et al., 2009), contains 3470 folk music pieces from six European countries or regions; thus the dataset is partitioned into six groups: England (1013 pieces), France (404 pieces), Ireland (824 pieces), Scotland (482 pieces), South East Europe (127 pieces) and Scandinavia (620 pieces). All pieces are monophonic melodies, encoded in MIDI, quantized and with grace notes removed (Taminau et al., 2009). To represent melodies, global attributes are selected from existing attribute sets, resulting in a total of 150 global attributes: 12 attributes from the Alicante feature set (Ponce de León and Iñesta, 2004), 37 from the Fantastic feature set (Müllensiefen, 2009), 39 from the Jesser feature set (Jesser, 1991), and 62 from the McKay feature set (McKay, 2010). Numeric features are discretized in a pre-processing step, into categorical values low and high, using as a split point the attribute's mean value in the complete corpus. Consequently, melodies are represented as tuples containing 150 attribute–value pairs and the region.

#### **15.4.1.2 Contrast Pattern Mining by Subgroup Discovery**

Subgroup discovery is applied to find global feature patterns which are characteristic for a region compared to other regions. Subgroups are extracted for each region at a

**Table 15.2** Contrast patterns discovered in the *Europa-6* corpus (based on Taminau et al., 2009). The table lists pattern  $X$ , group  $G$ , coverage  $P(X)$ , prevalence  $P(G)$ , sensitivity  $P(X|G)$ , confidence  $P(G|X)$  and weighted relative accuracy  $WRAcc$ . Abbreviated attribute names: proportion of descending minor thirds (dminthird); proportion of dotted notes (dotted); proportion of melodic tritones (melTrit); interpolation contour gradients standard deviation (intcontgradstd). Bold  $WRAcc$  values mark the strongest subgroup among subsuming subgroups (details see text). Contrast patterns for South East Europe are omitted because of inconsistencies in the reported measures (Taminau et al., 2009)

$X$	$G$	$P(X)$	$P(G)$	$P(X G)$	$P(G X)$	$WRAcc$
mode:major, notedensity:low	England	0.36	0.29	0.54	0.43	<b>0.052</b>
mode:major	England	0.77	0.29	0.88	0.33	0.032
notedensity:low	England	0.50	0.29	0.63	0.37	0.038
dminthird:low, range:low	France	0.26	0.12	0.77	0.34	<b>0.059</b>
dminthird:low	France	0.52	0.12	0.83	0.19	0.036
range:low	France	0.43	0.12	0.93	0.25	0.058
dotted:low, compoundMetre:1	Ireland	0.23	0.24	0.62	0.65	0.093
dotted:low	Ireland	0.70	0.24	0.86	0.29	0.038
compoundMetre:1	Ireland	0.32	0.24	0.71	0.53	<b>0.093</b>
metre:3/4, melTrit:low	Scandinavia	0.14	0.18	0.62	0.78	0.086
metre:3/4	Scandinavia	0.15	0.18	0.63	0.75	<b>0.086</b>
melTrit:low	Scandinavia	0.94	0.18	0.96	0.18	0.004
metre:4/4, intcontgradstd:high	Scotland	0.17	0.14	0.62	0.52	<b>0.063</b>
metre:4/4	Scotland	0.38	0.14	0.77	0.28	0.054
intcontgradstd:high	Scotland	0.44	0.14	0.76	0.24	0.044

time, taking all instances annotated with the region under consideration as positive examples and all instances annotated with other regions as negative examples (one-vs-all comparison). The study uses the CN2-SD algorithm (Lavrač et al., 2004), which adapts the classification rule induction algorithm CN2 (Clark and Niblett, 1989) for subgroup discovery. In CN2-SD rule candidates are evaluated by weighted relative accuracy (see (15.1)) rather than predictive accuracy. Compared to classification rule induction, which seeks to create highly accurate rules, weighted relative accuracy trades off accuracy against coverage in order to find statistically interesting subgroups which “are as large as possible and have the most unusual distributional characteristics with respect to the target attribute” (Lavrač et al., 2004, p. 154). In the application to folk music (Taminau et al., 2009), rules are generated with a fixed length of two features in the antecedent in order to avoid overfitting the data and to increase the interpretability of discovered rules.

### 15.4.1.3 Discovered Contrast Patterns

The study by Taminau et al. (2009) presents the top contrast pattern for each of the geographical regions, ranked by weighted relative accuracy. To facilitate interpretation of these rules, additional evaluation measures are reported: the coverage and sensitivity for the pattern and for each of its global features individually as well as

the confidence of the rule. The information on the individual features allows us to analyse pattern subsumption: given two sets of global features, a more specific set  $X$  is *subsumed* by a more general set  $\hat{X}$  if all pieces in the corpus which support set  $X$  also support set  $\hat{X}$ . Syntactically, a subsumed global feature pattern  $X$  is a superset of a more general global feature pattern  $\hat{X}$  (see Fig. 15.5). For each group Table 15.2 first lists the two-feature pattern reported in Taminau et al. (2009), followed by the subsuming single-feature patterns. Bold values in the last column mark the highest weighted relative accuracy in each rule triple; if both the original specialized pattern and a more general single-feature pattern have the same measure value the more general pattern is marked, as the specialized pattern does not provide further distinctive information for the characterization of the region.

The results support several observations. Subgroup descriptions are simple expressions built from categorical (or discretized) attributes and their values. Different subgroups are characterized by different attributes; only the metre attribute appears in more than one subgroup. The rules, which link the global feature pattern with a region, are partial rules, which do not cover all instances of a group or pattern: for the originally reported rules, sensitivity  $P(X | G)$  ranges between 54% and 77%, and confidence  $P(G | X)$  ranges between 34% and 78%. The measure of weighted relative accuracy trades off confidence and coverage of the rule: for Scandinavia and Ireland, a more general subgroup reaches the same weighted relative accuracy, despite a lower rule confidence, because of the higher coverage  $P(X)$  of the rule antecedent. These subgroups could already be sufficiently characterized by a single feature. Indeed, Taminau et al. observe that for Scandinavia the second component of the rule, the low proportion of melodic tritones, does not increase the reliability of the rule as the probability of this component in the Scandinavia group ( $P(X | G) = 0.96$ ) is hardly higher than its probability in the total corpus ( $P(X) = 0.94$ ). The subgroup is mainly specified by the metre feature, presumably relating to the large proportion of triple-metre polskas among the Scandinavian tunes (Taminau et al., 2009). By comparison, for the Ireland subgroup Taminau et al. comment that the addition of the second component substantially increases the rule's confidence, with  $P(G | X)$  increasing from 0.29 to 0.65. In fact, in this case the second component, compound metre, dominates the subgroup, and adding the first component, low proportion of dotted rhythms, does not increase weighted relative accuracy: the main genre among the Irish tunes in the corpus is the jig (Neubarth et al., 2013b), typically in 6/8 metre. In the case of the French tunes in *Europa-6*, the low proportion of descending minor thirds only slightly increases the weighted relative accuracy of the subgroup (from  $WRAcc = 0.058$  to  $WRAcc = 0.059$ ). More characteristic is a low range, which may be related to the fact that all French tunes with known lyrics in the corpus are covered by this description; as sung melodies they would obey certain restrictions of the human voice compared to instrumentally performed tunes (Taminau et al., 2009).

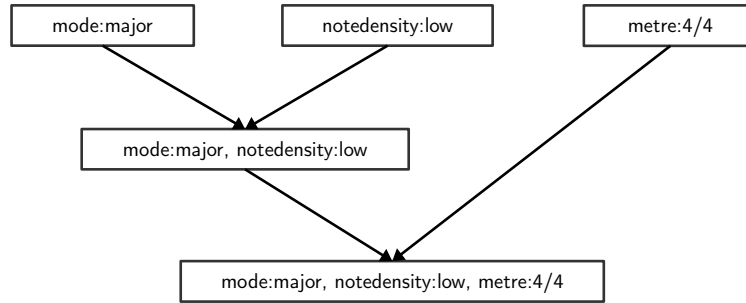


Fig. 15.5 Example of subsumption between global feature patterns

## 15.4.2 Case Study 2: Emerging Pattern Mining in Cretan Folk Tunes

As a second case study we present an example of sequential pattern discovery as emerging pattern mining (Conklin and Anagnostopoulou, 2011). Again, we summarize the data set, mining method and example results.

### 15.4.2.1 Dataset and Viewpoint Representation

For this study (Conklin and Anagnostopoulou, 2011), 106 Cretan folk tunes were selected from four printed sources which collate transcriptions of Cretan and Greek folk music. In preparation for computational analysis, the tunes were digitally encoded in score format and exported as MIDI files. The selected tunes represent eleven song types: four dance types and seven non-dance types. In addition, tunes are assigned to geographical regions: more specifically to one of five areas and more generally to Western or Eastern Crete. An idiosyncratic aspect of this dataset is that geographical area groups are not completely mutually exclusive: songs that are known to be sung in several areas of Crete were placed in all of the relevant area groups. In summary, the dataset is overlaid with four groupings (target attributes): type (11 groups), supertype (2 groups: dance vs. non-dance), area (5 groups) and superarea (2 groups: west vs. east).

Data instances are represented as sequences of intervals, using the viewpoint formalism of Conklin and Witten (1995): the melodic interval viewpoint calculates the interval between the current event and the previous event in semitones. As an example,  $[-4, +2]$  describes a sequence of a descending major third followed by an ascending major second, which is supported by, for example, the note (event) sequence [A, F, G]. With a single viewpoint, the viewpoint name is often omitted from the individual components in the description.



### 15.4.2.2 Contrast Pattern Mining by Emerging Pattern Mining









The MGDGP method (Conklin, 2010a) is applied to find maximally general distinctive patterns: emerging patterns which differentiate between the groups in the dataset. In previous work (Conklin and Bergeron, 2008), a pattern was considered interesting with respect to a corpus of music pieces if its frequency in the corpus was higher than expected, where expected frequency was computed from some statistical background distribution (see also Chap. 16, this volume). When applying the MGDGP method for emerging pattern mining of a corpus organized into groups, the support of a pattern in a target group can be directly compared against its support in other groups, or more specifically in a one-vs-all approach: against its support in the rest of the corpus. Then pattern interest  $I(X)$  is defined as the growth rate (see (15.2)), with the background dataset  $G'$  consisting of all groups but  $G$  (i.e.,  $\neg G$ ) to adapt the measure to the one-vs-all comparison:  $I(X) = P(X|G)/P(X|\neg G)$ . A pattern is distinctive if its interest  $I(X)$  is greater than or equal to a specified threshold  $\theta$ , with  $\theta > 1$ . Thus a pattern is distinctive for a group  $G$  if it is at least  $\theta$  times more likely to occur in group  $G$  than in the other groups. For  $P(X|\neg G) = 0$ , pattern interest is infinite,  $I(X) = \infty$ , and pattern  $X$  is called a *jumping pattern*.

Among distinctive patterns the analysis is interested in maximally general patterns: patterns which are not subsumed by more general distinctive patterns. A pattern  $X$  is *subsumed* by a more general pattern  $\hat{X}$  if all instances supporting  $X$  also support  $\hat{X}$ . In particular, a single-viewpoint sequential pattern is subsumed by any of its subsequences, and, vice versa, a pattern subsumes any pattern extended by one or more components (see Fig. 15.6). For example, the interval pattern  $[-4, +2]$  subsumes the pattern  $[-4, +2, -3]$ , supported by the note sequence [A, F, G, E]. If both a pattern  $X$  and a more general pattern  $\hat{X}$  are distinctive (and no more general pattern subsuming  $\hat{X}$  is distinctive), only pattern  $\hat{X}$  is reported as a maximally general distinctive pattern: while  $X$  is distinctive, it is not maximally general. If, on the other hand,  $X$  is distinctive but  $\hat{X}$  is not distinctive,  $X$  is reported. In addition, a minimum support threshold can be applied to ensure a certain generality of discovered rules.

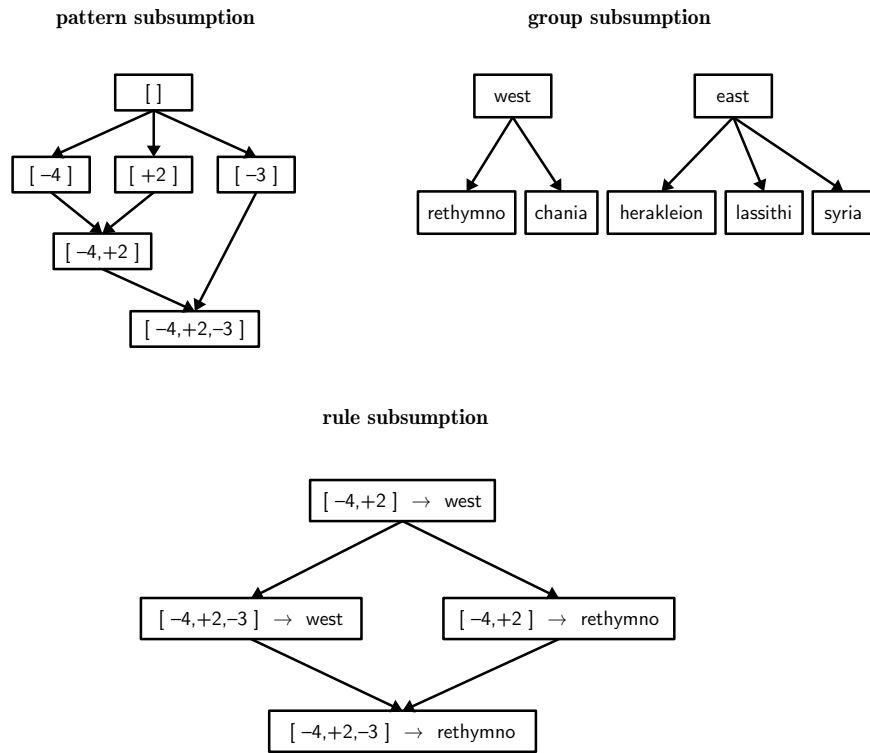
### 15.4.2.3 Discovered Contrast Patterns

Table 15.3 lists examples of discovered patterns which are distinctive (with a pattern interest threshold of 3) and maximally general. Only patterns with a minimum support count of 5 are presented (Conklin and Anagnostopoulou, 2011). From top to bottom the table includes two examples each for contrast mining by type, supertype, area and superarea. The results illustrate how local contrast patterns can overlap: of the dances described by the third and fourth rule in Table 15.3, twelve tunes support both patterns,  $[+4, -4]$  and  $[+4, +1, +2]$ . Two of the listed patterns are jumping patterns: the sequence of a descending fourth followed by a descending major second found in the dance syrtos, and the pattern of two intertwined falling thirds found in Western Crete.

**Table 15.3** Examples of maximally general distinctive patterns in Cretan folk tunes (Conklin and Anagnostopoulou, 2011). Columns indicate pattern  $X$ , group  $G$ , support count of the group  $n(G)$ , support count of the pattern in the group  $n(X \wedge G)$ , pattern interest  $I(X)$  and  $p$ -value according to Fisher’s exact test. The last column shows a schematic pitch sequence instantiating the pattern

$X$	$G$	$n(G)$	$n(X \wedge G)$	$I(X)$	$p$ -value	
$[-5, -2]$	syrtos	22	5	$\infty$	0.00032	
$[+1, +2, +3]$	malevisiotis	13	5	34.2	$8.6e-5$	
$[+4, -4]$	dance	51	16	19.1	$8.4e-6$	
$[+4, +1, +2]$	dance	51	19	11.4	$3.3e-6$	
$[-7, +4]$	lassithi	35	7	14.2	0.0017	
$[-4, +2, -3]$	rethymno	29	10	6.6	0.00028	
$[-4, +2, -3]$	west	64	14	$\infty$	0.00023	
$[+1, +2, -2, +2, -2]$	east	47	13	5.9	0.00081	

While only maximally general distinctive interval patterns are included in Table 15.3 and thus no subsumed distinctive patterns are reported, the rules for the west of Crete and for Rethymno (in Western Crete) are related by subsumption (see Fig. 15.6). In Conklin (2013), the MGD algorithm is extended to consider subsumption relations between groups and to exploit background ontologies in order to prune redundant rules. A rule linking a pattern and a group subsumes rules derived by specializing the pattern, specializing the group or both (see Fig. 15.6). If a more general rule is distinctive the search space underneath this rule can be pruned and specializations of the rule are not further explored. Thus, if the extended method was applied to the Cretan folk music corpus the pattern  $[-4, +2, -3]$  would no longer be reported for both west and rethymno but only for west: as the pattern is already distinctive for the super-area the specialized rule for rethymno would not be generated.



**Fig. 15.6** Examples of subsumption between patterns, groups and rules. Group (region) subsumption from Conklin and Anagnostopoulou (2011)

### 15.4.3 Case Study 3: Association Rule Mining in Basque Folk Music

The two case studies presented above yield positive rules: sufficiently frequent global feature patterns or sequential patterns which are more likely to occur in one group than in the other groups. In comparison, the third case study gives examples of mining not only for patterns which are over-represented but also for patterns which are under-represented in one group relative to the remaining corpus, considering both global feature patterns (Neubarth et al., 2013b) and sequential patterns (Conklin, 2013).

#### 15.4.3.1 Data

The *Cancionero Vasco* is a collection of Basque folk songs and dances, collated by the Basque folklorist and composer Padre Donostia. The digitized corpus, curated

by the Basque Studies Society, consists of 1902 tunes. The tunes are annotated with geographical information on the location of collection and with genre information, organizing the dataset into seven regions and five genres, which are further specified into sub-regions and sub-genres, giving a total of 272 regions and 31 genres used to label tunes in the *Cancionero Vasco*. The tunes were represented by global features (Neubarth et al., 2013b), with continuous attributes discretized during pre-processing, and by viewpoints (Conklin, 2013). Contrasting global feature patterns were mined for region and for genre as target attribute, only considering the top-level regions and genres (Neubarth et al., 2013b); for the discovery of contrasting sequential patterns (Conklin, 2013) the genre was selected as the target attribute, taking into account the complete genre ontology.

### 15.4.3.2 Contrast Pattern Mining by Association Rule Mining

Conklin (2013) introduced the term *antipatterns* to denote (sequential) patterns which are rare in a corpus, corresponding to a target group  $G$ , but frequent in an anticorpus, corresponding to the complement of the target group  $\neg G$ . This chapter shows that, in the context of contrast pattern mining, conjunctions of global features can be treated as patterns; thus the term *antipatterns* can equally be applied to global feature patterns under-represented in a corpus relative to an anticorpus. Then antipatterns can be expressed as negative class association rules,  $X \rightarrow \neg G$ , where  $X$  is a global feature pattern or a sequential pattern.

In the analysis of global feature patterns (Neubarth et al., 2013b), constrained association rule mining is applied to find both positive and negative rules. Candidate rules are primarily evaluated by rule confidence (see (15.3)):  $c(X \rightarrow G) = P(G|X)$  for positive rules and  $c(X \rightarrow \neg G) = P(\neg G|X) = 1 - P(G|X)$  for negative rules. A minimum confidence threshold ensures that the pattern occurs mainly in instances of group  $G$  (positive rules) or mainly in instances of groups other than  $G$  (negative rules), and that group  $G$  is distinguished from other groups by a high or low number of data instances supporting pattern  $X$ . The mining is restricted to pairwise associations between one attribute–value pair in the antecedent and the target attribute in the consequent in order to reveal the fundamental relationships in the data before considering attribute interactions and in order to reduce the search space by syntactic rather than evaluation measure constraints: for negative rules with small or even zero rule support the pruning strategies of positive rule mining based on minimum support (Agrawal and Srikant, 1994) can no longer be applied. For rules meeting the confidence threshold, a  $p$ -value is calculated according to Fisher’s one-tailed exact test, with right tail for positive and left tail for negative rules. The test measures the probability of finding at least (right tail) or at most (left tail) the observed number  $n(X \wedge G)$  of instances supporting both  $X$  and  $G$  given the marginal counts in the contingency table (see Fig. 15.4); the lower the  $p$ -value the less likely it is that the number of co-occurrences can be explained by chance alone.

Sequential antipattern candidates in the *Cancionero Vasco* are primarily evaluated by their  $p$ -value (Conklin, 2013). In addition to Fisher’s exact test in order to assess

**Table 15.4** Positive and negative rules discovered in the *Cancionero Vasco*. Top: global feature patterns and antipatterns discovered with a minimum confidence threshold of 0.5 (based on Neubarth et al., 2013b). Abbreviated attribute names referring to attributes of the McKay (2010) set: primary register (register); repeated notes (repeated); average melodic interval (intervals). Bottom: sequential antipatterns discovered with a maximum  $p$ -value threshold of 0.01 (Fisher’s test) and minimum  $p$ -value threshold of 0.001 ( $\chi^2$  test) (Conklin, 2013)

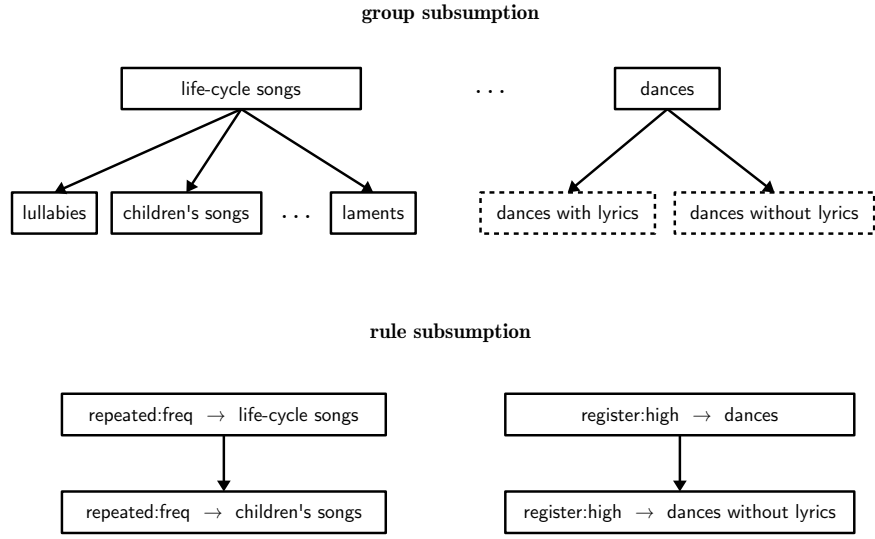
$X$		$G$ or $\neg G$	$n(X)$	$n(G)$	$n(X \wedge G)$	confidence	$p$ -value
register:high	→	dances	249	495	179	0.72	$\approx 0.0$
repeated:freq	→	life-cycle songs	7	477	5	0.71	0.012
intervals:narrow	→	$\neg$ Araba	945	27	4	0.99	$6.0e-4$
notedensity:low	→	$\neg$ Zuberoa	95	80	0	1.00	0.015
[+1, -5, 0]	→	$\neg$ historical songs	60	121	0	1.00	0.007
[+1, -1]	→	$\neg$ children’s songs	754	56	13	0.98	$7.8e-5$
[2/1, 3/2]	→	$\neg$ festive songs	164	82	2	0.99	0.005
[3/1]	→	$\neg$ lullabies	883	98	29	0.97	$2.3e-8$

the reliability of the association between an antipattern and a group, a  $\chi^2$  goodness of fit test is applied to assess the distribution of a candidate antipattern in the anticorpus. The anticorpus  $\neg G$  covers several groups. To be considered interesting, an antipattern candidate should be frequent across the groups of the anticorpus rather than being concentrated in only one group within the anticorpus; in the latter case it would be more appropriate to consider it as a pattern for that group instead of an antipattern for the target group (Conklin, 2013). Thus, antipatterns must satisfy two measure thresholds: a maximum  $p$ -value threshold in Fisher’s test and a minimum  $p$ -value threshold in the  $\chi^2$  test.

### 15.4.3.3 Discovered Contrast Patterns and Antipatterns

Table 15.4 shows selected results of mining the *Cancionero Vasco* for positive and negative rules. The first four rules relate to global feature patterns, with two examples each for positive and for negative associations, covering both genres and regions as target attributes. The second part of the table lists examples of sequential antipatterns for genres: two melodic interval antipatterns and two duration ratio antipatterns. The last global feature antipattern and the first interval antipattern are jumping antipatterns, which do not occur at all in the target group.

Considering not only the top-level genres and regions but the complete taxonomies and thus exploring rule subsumption (Neubarth et al., 2013a) could provide additional insight and help the interpretation of general rules (see Fig. 15.7). For example, high register is particularly related to dances without lyrics (165 out of the 179 tunes) and thus presumably instrumental performance of these tunes. The primary register attribute of the McKay (2010) set is calculated as the average MIDI pitch of a piece; a high register corresponds to an average pitch of 73 or higher, that is, above C5 (see example in Fig. 15.8). The positive association between tunes with a high proportion of repeated notes and the genre of life-cycle songs is supported by repetitive melodies



**Fig. 15.7** Examples of group (genre) subsumption and rule subsumption for patterns discovered in the *Cancionero Vasco*. In the group taxonomy solid boxes refer to metadata annotations, dashed boxes refer to specialized groups of dances derived by evaluating the presence of lyrics

in children’s songs, which include songs associated with children’s games such as skipping songs, clapping songs or counting-out rhymes (see also Martí i Pérez et al., 2001). Figure 15.9 shows a children’s song from the *Cancionero Vasco*, in which 70.1% of notes repeat the previous note, against a mean value of 19.2% in the complete corpus. For German folk songs, Steinbeck (1982) also observes a higher proportion of repeated notes in children’s tunes than in other folk music samples.



**Fig. 15.8** Untitled dance without lyrics (excerpt) from the *Cancionero Vasco*



**Fig. 15.9** Children’s song “Arri, arri, mandua” from the *Cancionero Vasco*

## 15.5 Conclusions

Contrast pattern mining aims to discover and describe patterns which differentiate subpopulations or groups in a dataset. This chapter has introduced the task and representative methods of contrast pattern mining and has reviewed work in quantitative and computational folk music analysis which can be phrased as mining for contrast patterns. Three case studies have been presented in detail, illustrating a variety of folk tune corpora, different groupings or partitionings of datasets, global feature and sequential representations of patterns, and different contrast pattern mining methods.

The analysis task of contrast pattern mining is related to classical methods in descriptive and inferential statistics, which describe the distributions of variables and differences between group means and variances for selected continuous variables. Statistical analysis has also been used in quantitative folk music analysis, generally restricted to analysing single variables (e.g., Freeman and Merriam, 1956; Gundlach, 1932; Kolinski, 1982; Segrillo, 2010; Steinbeck, 1982). Methods of descriptive and inferential statistics are productively applied for testing variable distributions across the complete dataset, rather than for knowledge discovery (i.e., the discovery of previously unknown relations). In classic contrast pattern mining, on the other hand, pattern descriptions are based on discrete attributes, and discovered contrast patterns are local descriptive patterns which constitute binary predicates on music pieces in a corpus and are assessed by evaluating pattern support in different groups.

Contrast patterns are traditionally specified as combinations (sets or sequences) of attribute–value pairs. Previous work in computational folk music analysis has sometimes used attribute selection or classification to evaluate the discriminative power of single attributes or sets of attributes in the context of predictive model induction (e.g., Bohak and Marolt, 2009; Hillewaere, 2013; van Kranenburg and Karsdorp, 2014; van Kranenburg et al., 2013); the methods reveal attributes which can separate groups of music pieces but reported results do not give insight into which values or value ranges of the attributes are characteristic for which group.

For sequential pattern discovery in inter-opus music analysis, two ways of determining pattern frequencies have been proposed (e.g., Conklin and Bergeron, 2008): the piece count of a pattern is the number of pieces in the corpus which contain the pattern one or more times; the total count of a pattern is the number of all pattern occurrences including repetitions of a pattern within the same piece. The definition of pattern support in this chapter, and in the case studies presented in this chapter, corresponds to piece count; it follows general approaches in contrast data mining and classic sequential pattern mining (Agrawal and Srikant, 1995; Mooney and Roddick, 2013; Novak et al., 2009b). Some quantitative studies in folk music research have also considered total counts, or proportional frequencies, of event features (e.g., Gundlach, 1932; Segrillo, 2010). Beyond folk music, examples of mining contrasting, or distinctive, melodic or polyphonic patterns using total counts include an analysis of Brahms' string quartet op. 51 no. 1 against the string quartets op. 51 no. 2 and op. 67 (Conklin, 2010b) and the analysis of Beethoven's piano sonatas in Chap. 17 of this volume.

As the name contrast data mining suggests, the analysis focuses on differences, rather than similarities, between groups. The notion of similarity has attracted considerable attention in music information retrieval and computational music research, including computational folk music research (e.g., Müllensiefen and Frieler, 2004; Scherrer and Scherrer, 1971); generally the comparison tends to be between pairs of individual folk tunes rather than at the level of groups of tunes. An early exception is an example study by Suchoff (1970), who compares support counts of interval sequences in folk music repertoires (Serbo-Croatian women's songs and Romanian winter solstice songs) against samples of 16th-century vocal and instrumental music and finds "interesting similarities [...] which seem to point toward folk-based or folk-styled characteristics of the art melodies" (p. 202). Similarity or overlap between folk music repertoires has also been studied in the context of clustering (e.g., Juhász, 2006; Juhász and Sipos, 2010). However, clustering itself is an unsupervised data mining method: information on groups is introduced manually by partitioning the mining input or during post-processing; groups are referenced by prototypical instances rather than generalized patterns (Juhász, 2006) or filtered visualization rather than symbolic description (Toivainen and Eerola, 2001).

Compared to the unsupervised task of clustering, contrast pattern mining is a supervised data mining task, which evaluates instances from labelled groups within a dataset. The three case studies presented in this chapter have all used predefined groupings or partitionings of a folk music corpus based on metadata, for example, into folk music genres or regions of collection. Obviously there is no one "correct" grouping of a corpus appropriate for all analysis tasks, and there is no reason the grouping cannot use any available ontology, be specified flexibly by an analyst considering metadata or music content characteristics (e.g., Jesser, 1991) or result from previous computational analysis such as clustering (e.g., Steinbeck, 1982).

While this chapter has focused on applications in folk music analysis, quantitative comparative analysis and contrast pattern mining are not confined to this repertoire. Existing musicological research has applied quantitative methods to corpora of, for example, early music (Trowbridge, 1986), performance criticism of Beethoven piano sonatas (Alessandri et al., 2014), 19th-century music (VanHandel, 2009) and popular music (Huron and Ommen, 2006), as well as comparisons between folk and art music (Suchoff, 1970). Recent example studies in music data mining and music information retrieval include, for example, subgroup discovery in Haydn and Mozart string quartets (Taminau et al., 2010), corpus analysis of ragtime (Volk and de Haas, 2013) or visualization of contrapuntal module occurrences in different style periods (Antila and Cumming, 2014). Contrast pattern mining could be used where support counts can be determined for features or feature sets which are derived from metadata, such as instruments, genres or styles in music publication titles (Hess, 1953; Rose and Tuppen, 2014); extracted from music scores or digitized music content, such as tempo indications, tonal material or melodic patterns (Albrecht and Huron, 2014; Post and Huron, 2009; Suchoff, 1970); or compiled from contextual information, such as dates, locations and repertoire of concert performances (Kopiez et al., 2009) or performers, compositions and release dates of music recordings (Alessandri et al., 2014).



Contrast patterns are not restricted to the specific representations of melodic patterns used in the reported case studies: conjunctions of global features and sequences of event features. The general approach is very flexible, and numerous other pattern representations used in symbolic music analysis could also be considered, for example, vertical viewpoint patterns (Conklin, 2002), geometric polyphonic patterns (Meredith et al., 2002, see also Chaps. 13 and 17, this volume) and chord sequence patterns (Conklin, 2010a). Theoretically, any representation that defines a computable boolean predicate can be used for a pattern representation (Atzmüller, 2015), for example: patterns described by a conjunction of global and event features, feature set sequential patterns (Conklin and Bergeron, 2008), sequential patterns with gaps (Comin and Parida, 2008), patterns with variables (Angluin, 1980), and even patterns described by grammars or rewrite rules (Sidorov et al., 2014). With more powerful pattern representations would come additional challenges. First, the classical tradeoff between expressiveness and tractability in knowledge representation (Levesque and Brachman, 1987) would naturally be encountered, and compromises on the search of the pattern space would need to be accepted. Second, the results of contrast pattern mining may be more difficult to interpret as one moves to further levels of abstraction from the underlying music representation. Logical subsumption between patterns, as well as consideration of domain ontologies, offers one strategy to restrict the search space, organize the data mining output and facilitate the interpretation of discovered patterns.

**Acknowledgements** This research is partially supported by the project Lrn2Cre8 which is funded by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 610859.

## References

- Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB 1994)*, pages 487–499, Santiago de Chile, Chile.
- Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the 11th International Conference on Data Engineering (ICDE 1995)*, pages 3–14, Taipei, Taiwan.
- Albrecht, J. D. and Huron, D. (2014). A statistical approach to tracing the historical development of major and minor pitch distributions, 1400–1750. *Music Perception*, 31(4):223–243.
- Alessandri, E., Eiholzer, H., and Williamon, A. (2014). Reviewing critical practice: an analysis of *Gramophone*'s reviews of Beethoven's piano sonatas, 1923–2010. *Musicae Scientiae*, 18(2):131–149.
- Ali, K., Manganaris, S., and Srikant, R. (1997). Partial classification using association rules. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97)*, pages 115–118, Newport Beach, CA, USA.

- Anagnostopoulou, C., Giraud, M., and Poulakis, N. (2013). Melodic contour representations in the analysis of children's songs. In *Proceedings of the 3rd International Workshop on Folk Music Analysis (FMA 2013)*, pages 40–42, Amsterdam, Netherlands.
- Angluin, D. (1980). Finding patterns common to a set of strings. *Journal of Computer and System Sciences*, 21:46–62.
- Antila, C. and Cumming, J. (2014). The VIS framework: analyzing counterpoint in large datasets. In *Proceedings of the 15th International Society of Music Information Retrieval Conference (ISMIR 2014)*, pages 71–76, Taipei, Taiwan.
- Atzmüller, M. (2015). Subgroup discovery: advanced review. *WIREs Data Mining and Knowledge Discovery*, 5(1):35–49.
- Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P. F., editors (2003). *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press.
- Bay, S. D. (2000). Multivariate discretization of continuous variables for set mining. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000)*, pages 315–319, Boston, MA, USA.
- Bay, S. D. and Pazzani, M. J. (2001). Detecting group differences: mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3):213–246.
- Bohak, C. and Marolt, M. (2009). Calculating similarity of folk song variants with melody-based features. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 597–601, Kobe, Japan.
- Bohlman, P. V. (1988). *The Study of Folk Music in the Modern World*. Indiana University Press.
- Bronson, B. H. (1959). Toward the comparative analysis of British-American folk tunes. *The Journal of American Folklore*, 72(284):165–191.
- Carter, T. (1986/1987). Music publishing in Italy, c.1580–c.1625: some preliminary observations. *Royal Musical Association Research Chronicle*, 20:19–37.
- Chan, S., Kao, B., Yip, C. L., and Tang, M. (2003). Mining emerging substrings. In *Proceedings of the 8th International Conference on Database Systems for Advanced Applications (DASFAA 2003)*, pages 119–126, Kyoto, Japan.
- Clark, P. and Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, 3(4):261–283.
- Comin, M. and Parida, L. (2008). Detection of subtle variations as consensus motifs. *Theoretical Computer Science*, 395(2–3):158–170.
- Conklin, D. (2002). Representation and discovery of vertical patterns in music. In *Proceedings of the 2nd International Conference on Music and Artificial Intelligence (ICMAI 2002)*, pages 32–42, Edinburgh, UK.
- Conklin, D. (2009). Melody classification using patterns. In *2nd International Workshop on Machine Learning and Music at ECML/PKDD 2009 (MML 2009)*, Bled, Slovenia.
- Conklin, D. (2010a). Discovery of distinctive patterns in music. *Intelligent Data Analysis*, 14:547–554.
- 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. (2013). Antipattern discovery in folk tunes. *Journal of New Music Research*, 42(2):161–169.
- Conklin, D. and Anagnostopoulou, C. (2011). Comparative pattern analysis of Cretan folk songs. *Journal of New Music Research*, 40(2):119–125.
- Conklin, D. and Bergeron, M. (2008). Feature set patterns in music. *Computer Music Journal*, 32(1):60–70.
- Conklin, D. and Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73.
- Cook, N. (2004). Computational and comparative musicology. In Clarke, E. and Cook, N., editors, *Empirical Musicology: Aims, Methods, Prospects*, pages 103–126. Oxford University Press.
- Cornelis, C., Yan, P., Zhang, X., and Chen, G. (2006). Mining positive and negative association rules from large databases. In *Proceedings of the IEEE International Conference on Cybernetics and Intelligent Systems*, pages 613–618, Bangkok, Thailand.
- Cornelis, O., Lesaffre, M., Moelants, D., and Leman, M. (2010). Access to ethnic music: advances and perspectives in content-based music information retrieval. *Signal Processing*, 90:1008–1031.
- Csébfalvy, K., Havass, M., Járdányi, P., and Vargyas, L. (1965). Systematization of tunes by computers. *Studia Musicologica Academiae Scientiarum Hungaricae*, 7:253–257.
- Densmore, F. (1913). *Chippewa Music II*. Smithsonian Institution Bureau of American Ethnology Bulletin 53.
- Densmore, F. (1918). *Teton Sioux Music*. Smithsonian Institution Bureau of American Ethnology Bulletin 61.
- Densmore, F. (1922). *Northern Ute Music*. Smithsonian Institution Bureau of American Ethnology Bulletin 75.
- Densmore, F. (1929). *Pawnee Music*. Smithsonian Institution Bureau of American Ethnology Bulletin 93.
- Densmore, F. (1932). *Yuman and Yaqui Music*. Smithsonian Institution Bureau of American Ethnology Bulletin 110.
- Dong, G. and Li, J. (1999). Efficient mining of emerging patterns: discovering trends and differences. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99)*, pages 43–52, San Diego, CA, USA.
- Edström, O. (1999). From *schottis* to *bonnjazz* – some remarks on the construction of Swedishness. *Yearbook for Traditional Music*, 31:27–41.
- Elscheková, A. (1965). General considerations on the classification of folk tunes. *Studia Musicologica Academiae Scientiarum Hungaricae*, 7:259–262.
- Elscheková, A. (1966). Methods of classification of folk tunes. *Journal of the International Folk Music Council*, 18:56–76.
- Elscheková, A. (1999). Musikvergleich und Computertechnik. *Croatian Journal of Ethnology*, 36(2):105–115.
- Forrest, J. and Heaney, M. (1991). Charting early morris. *Folk Music Journal*, 6(2):169–186.

- Freeman, L. C. and Merriam, A. P. (1956). Statistical classification in anthropology: an application to ethnomusicology. *American Anthropologist*, 58:464–472.
- Fujinaga, I. and Weiss, S. F. (2004). Music. In Schreibman, S., Siemens, R., and Unsworth, J., editors, *A Companion to Digital Humanities*, pages 97–107. Blackwell.
- Fürnkranz, J., Gamberger, D., and Lavrač, N. (2012). *Foundations of Rule Learning*. Springer.
- Geng, L. and Hamilton, H. J. (2006). Interestingness measures for data mining: a survey. *ACM Computing Surveys*, 38(3):1–32.
- Grauer, V. A. (1965). Some song-style clusters – a preliminary study. *Ethnomusicology*, 9(3):265–271.
- Gundlach, R. H. (1932). A quantitative analysis of Indian music. *The American Journal of Psychology*, 44(1):133–145.
- Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. The MIT Press.
- Herrera, F., Carmona, C. J., González, P., and del Jesus, M. J. (2011). An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, 29(3):495–525.
- Hess, A. G. (1953). The transition from harpsichord to piano. *The Galpin Society Journal*, 6:75–94.
- Hillewaere, R. (2013). *Computational models for folk music classification*. PhD thesis, Faculty of Sciences, Vrije Universiteit Brussel, Belgium.
- Hillewaere, R., Manderick, B., and Conklin, D. (2009). Global feature versus event models for folk song classification. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 729–733, Kobe, Japan.
- Hoshovs'kyj, V. (1965). The experiment of systematizing and cataloguing folk tunes following the principles of musical dialectology and cybernetics. *Studia Musicologica Academiae Scientiarum Hungaricae*, 7:273–286.
- Huron, D. and Ommen, A. (2006). An empirical study of syncopation in American popular music, 1890–1939. *Music Theory Spectrum*, 28(2):211–231.
- Járdányi, P. (1965). Experiences and results in systematizing Hungarian folk-songs. *Studia Musicologica Academiae Scientiarum Hungaricae*, 7:287–291.
- Jesser, B. (1991). *Interaktive Melodieanalyse. Methodik und Anwendung computergestützter Analyseverfahren in Musikethnologie und Volksliedforschung: typologische Untersuchung der Balladensammlung des DVA*. Studien zur Volksliedforschung 12. Peter Lang.
- Juhász, Z. (2006). A systematic comparison of different European folk music traditions using self-organizing maps. *Journal of New Music Research*, 35(2):95–112.
- Juhász, Z. and Sipos, J. (2010). A comparative analysis of Eurasian folksong corpora, using self organising maps. *Journal of Interdisciplinary Music Studies*, 4(1):1–16.
- Kavšek, B. and Lavrač, N. (2006). APRIORI-SD: adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, 20:543–583.

- Keller, M. S. (1984). The problem of classification in folksong research: a short history. *Folklore*, 95(1):100–104.
- Klösigen, W. (1996). Explora: a multipattern and multistrategy discovery assistant. In Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P., editors, *Advances in Knowledge Discovery and Data Mining*, pages 249–271. MIT Press.
- Klösigen, W. (1999). Applications and research problems of subgroup mining. In Raś, Z. W. and Skowron, A., editors, *Foundations of Intelligent Systems*, pages 1–15. Springer.
- Kolinski, M. (1982). Reiteration quotients: a cross-cultural comparison. *Ethnomusicology*, 26(1):85–90.
- Kopiecz, R., Lehmann, A. C., and Klassen, J. (2009). Clara Schumann’s collection of playbills: a historiometric analysis of life-span development, mobility, and repertoire canonization. *Poetics*, 37:50–73.
- Lampert, V. (1982). Bartók’s choice of theme for folksong arrangement: some lessons of the folk-music sources of Bartók’s works. *Studia Musicologica Academiae Scientiarum Hungaricae*, 24(3/4):401–409.
- Lavrač, N., Kavšek, B., Flach, P., and Todorovski, L. (2004). Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188.
- Levesque, H. J. and Brachman, R. J. (1987). Expressiveness and tractability in knowledge representation and reasoning. *Computational Intelligence*, 3:78–92.
- Lincoln, H. B. (1970). The current state of music research and the computer. *Computers and the Humanities*, 5(1):29–36.
- Lincoln, H. B. (1974). Use of the computer in music research: a short report on accomplishments, limitations, and future needs. *Computers and the Humanities*, 8(5/6):285–289.
- Liu, B., Hsu, W., and Ma, Y. (1998). Integrating classification and association rule mining. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD-98)*, pages 80–86, New York, USA.
- Lomax, A. (1959). Folk song style. *American Anthropologist*, 61(6):927–954.
- Lomax, A. (1962). Song structure and social structure. *Ethnology*, 1(4):425–451.
- Marsden, A. (2009). “What was the question?”: music analysis and the computer. In Crawford, T. and Gibson, L., editors, *Modern Methods for Musicology: Prospects, Proposals and Realities*, pages 137–147. Ashgate.
- Martí i Pérez, J., Cunningham, M., Pelinski, R., Martínez García, S., de Larrea Palacín, A., and Aiats, J. (2001). Spain. II: Traditional and popular music. In Sadie, S., editor, *New Grove Dictionary of Music and Musicians*, volume 24, pages 135–154. Macmillan, 2nd edition.
- McKay, C. (2010). *Automatic music classification with jMIR*. PhD thesis, Schulich School of Music, McGill University, Montreal, Canada.
- Meredith, D., Lemström, K., and Wiggins, G. (2002). Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, 31(4):321–345.
- Mooney, C. H. and Roddick, J. F. (2013). Sequential pattern mining: approaches and algorithms. *ACM Computing Surveys*, 45(2):Article 19.

- Müllensiefen, D. (2009). Fantastic: Feature ANalysis Technology Accessing Statistics (In a Corpus): Technical report v1.5. Technical report, Goldsmiths College, University of London, London, UK.
- Müllensiefen, D. and Frieler, K. (2004). Optimizing measures of melodic similarity for the exploration of a large folk song database. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, pages 274–280, Barcelona, Spain.
- Nettl, B. (1973). Comparison and comparative method in ethnomusicology. *Anuario Interamericano de Investigacion Musical*, 9:143–161.
- Nettl, B. (1975). The state of research in ethnomusicology, and recent developments. *Current Musicology*, 20:67–78.
- Nettl, B. (2005). *The Study of Ethnomusicology*. University of Illinois Press.
- Nettl, B. (2010). *Nettl's Elephant: On the History of Ethnomusicology*. University of Illinois Press.
- Neubarth, K. (2015). Densmore revisited: contrast data mining of Native American music. In *5th International Workshop on Folk Music Analysis (FMA 2015)*, Paris, France. (to appear).
- Neubarth, K., Goienetxea, I., Johnson, C. G., and Conklin, D. (2012). Association mining of folk music genres and toponyms. In *Proceedings of the 13th International Society of Music Information Retrieval Conference (ISMIR 2012)*, pages 7–12, Porto, Portugal.
- Neubarth, K., Johnson, C. G., and Conklin, D. (2013a). Descriptive rule mining of Basque folk music. In *Proceedings of the 3rd International Workshop on Folk Music Analysis (FMA 2013)*, pages 83–85, Amsterdam, Netherlands.
- Neubarth, K., Johnson, C. G., and Conklin, D. (2013b). Discovery of mediating association rules for folk music analysis. In *6th International Workshop on Music and Machine Learning at ECML/PKDD 2013 (MML 2013)*, Prague, Czech Republic.
- Novak, P. K., Lavrač, N., Gamberger, D., and Krstačić, A. (2009a). CSM-SD: methodology for contrast set mining through subgroup discovery. *Journal of Biomedical Informatics*, 42(1):113–122.
- Novak, P. K., Lavrač, N., and Webb, G. (2009b). Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10:377–403.
- Ponce de León, P. J. and Iñesta, J. M. (2004). Statistical description models for melody analysis and characterization. In *Proceedings of the 30th International Computer Music Conference (ICMC 2004)*, pages 149–156, Miami, Florida, USA.
- Post, O. and Huron, D. (2009). Western Classical music in the minor mode is slower (except in the Romantic period). *Empirical Musicology Review*, 4(1):2–10.
- Rhodes, W. (1965). The use of the computer in the classification of folk tunes. *Studia Musicologica Academiae Scientiarum Hungaricae*, 7:339–343.
- Rose, S. and Tuppen, S. (2014). Prospects for a big data history of music. In *Proceedings of the 1st Digital Libraries for Musicology Workshop (DLfM 2014)*, pages 82–84, London, UK.

- Sagrillo, D. (2010). Computer analysis of Scottish national songs. In *1st International Conference on Analytical Approaches to World Music (AAWM 2010)*, Amherst, MA, USA.
- Savasere, A., Omiecinski, E., and Navathe, S. (1998). Mining for strong negative associations in a large database of customer transactions. In *Proceedings of the 14th International Conference on Data Engineering (ICDE 1998)*, pages 494–502, Orlando, FL, USA.
- Schaffrath, H. (1992). The retrieval of monophonic melodies and their variants: concepts and strategies for computer-aided analysis. In Marsden, A. and Pople, A., editors, *Computer Representations and Models in Music*, pages 95–109. Academic Press.
- Scherrer, D. K. and Scherrer, P. H. (1971). An experiment in the computer measurement of melodic variation in folksong. *The Journal of American Folklore*, 84(332):230–241.
- Schneider, A. (2006). Comparative and systematic musicology in relation to ethnomusicology: a historical and methodological survey. *Ethnomusicology*, 50(2):236–258.
- Sidorov, K., Jones, A., and Marshall, D. (2014). Music analysis as a smallest grammar problem. In *Proceedings of the 15th International Society for Information Retrieval Conference (ISMIR 2014)*, pages 301–306, Taipei, Taiwan.
- Srikant, R. and Agrawal, R. (1996). Mining quantitative association rules in large relational tables. *ACM SIGMOD Record*, 25(2):1–12.
- Steinbeck, W. (1976). The use of the computer in the analysis of German folksongs. *Computers and the Humanities*, 10(5):287–296.
- Steinbeck, W. (1982). *Struktur und Ähnlichkeit. Methoden automatisierter Melodienanalyse*. Kieler Schriften zur Musikwissenschaft 25. Bärenreiter.
- Suchoff, B. (1967). Computer applications to Bartók’s Serbo-Croatian material. *Tempo*, 80:15–19.
- Suchoff, B. (1968). Computerized folk song research and the problem of variants. *Computers and the Humanities*, 2(4):155–158.
- Suchoff, B. (1970). Computer-oriented comparative musicology. In Lincoln, H. B., editor, *The Computer and Music*, pages 193–206. Cornell University Press.
- Suchoff, B. (1971). The computer and Bartók research in America. *Journal of Research in Music Education*, 19(1):3–16.
- Taminau, J., Hillewaere, R., Meganck, S., Conklin, D., Nowé, A., and Manderick, B. (2009). Descriptive subgroup mining of folk music. In *2nd International Workshop on Machine Learning and Music at ECML/PKDD 2009 (MML 2009)*, Bled, Slovenia.
- Taminau, J., Hillewaere, R., Meganck, S., Conklin, D., Nowé, A., and Manderick, B. (2010). Applying subgroup discovery for the analysis of string quartet movements. In *Proceedings of the 3rd International Workshop on Music and Machine Learning at ACM Multimedia (MML 2010)*, pages 29–32, Florence, Italy.
- Toivaiainen, P. and Eerola, T. (2001). A method for comparative analysis of folk music based on musical feature extraction and neural networks. In *Proceedings of the*

- VII International Symposium on Systematic and Comparative Musicology/III International Conference on Cognitive Musicology*, pages 41–45, Jyväskylä, Finland.
- Trowbridge, L. M. (1985/1986). Style change in the fifteenth-century chanson: a comparative study of compositional detail. *The Journal of Musicology*, 4(2):146–170.
- Tzanetakis, G., Kapur, A., Schloss, W., and Wright, M. (2007). Computational ethnomusicology. *Journal of Interdisciplinary Music Studies*, 1(2):1–24.
- van Kranenburg, P., Garbers, J., Volk, A., Wiering, F., Grijp, L., and Veltkamp, R. C. (2010). Collaborative perspectives for folk song research and music information retrieval: The indispensable role of computational musicology. *Journal of Interdisciplinary Music Studies*, 4(1):17–43.
- van Kranenburg, P. and Karsdorp, F. (2014). Cadence detection in Western traditional stanzaic songs using melodic and textual features. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, pages 391–396, Taipei, Taiwan.
- van Kranenburg, P., Volk, A., and Wiering, F. (2013). A comparison between global and local features for computational classification of folk song melodies. *Journal of New Music Research*, 42(1):1–18.
- VanHandel, L. (2009). National metrical types in nineteenth century art song. *Empirical Musicology Review*, 4(4):134–145.
- Volk, A. and de Haas, W. B. (2013). A corpus-based study on ragtime syncopation. In *Proceedings of the 14th International Society of Music Information Retrieval Conference (ISMIR 2013)*, pages 163–168, Curitiba, Brazil.
- Webb, G. I., Butler, S., and Newlands, D. (2003). On detecting differences between groups. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, pages 256–265, Washington, DC, USA.
- Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, third edition.
- Wong, T.-T. and Tseng, K.-L. (2005). Mining negative contrast sets from data with discrete attributes. *Expert Systems with Applications*, 29(2):401–407.
- Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *Proceedings of the 1st European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'97)*, pages 78–87, Trondheim, Norway.
- Zimmermann, A. and De Raedt, L. (2009). Cluster-grouping: from subgroup discovery to clustering. *Machine Learning*, 77(1):125–159.