Antipattern Discovery in Ethiopian Bagana Songs

Darrell Conklin[1,2] and Stéphanie Weisser[3]

[1] Department of Computer Science and Artificial Intelligence
University of the Basque Country UPV/EHU, San Sebastián, Spain
[2] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[3] Université Libre de Bruxelles, Brussels, Belgium

Abstract. This paper develops and applies sequential pattern mining to a corpus of songs for the bagana, a large lyre played in Ethiopia. An important aspect of this repertoire is the unique availability of rare motifs that have been used by a master bagana teacher in Ethiopia. The method is applied to find antipatterns: patterns that are surprisingly rare in a corpus of bagana songs. In contrast to previous work, this is performed without an explicit set of background pieces. The results of this study show that data mining methods can reveal with high significance these antipatterns of interest based on the computational analysis of a small corpus of bagana songs.

1 Introduction

Sequences are a special form of data that require specific attention with respect to alternative representations and data mining techniques. Sequential pattern mining methods [2,1,19] can be used to find frequent and significant patterns in datasets of sequences, and also sequential patterns that contrast one data group against another [16,23]. In music, sequential pattern discovery methods have been used for the analysis of single pieces [9], for the analysis of a corpus of pieces [8], and also to find short patterns that can be used to classify melodies [21,22,18,7].

Further to standard pattern discovery methods, which search for frequent patterns satisfying minimum support thresholds [24], another area of interest is the discovery of rare patterns. This area includes work on rare itemset mining [14] and negative association rules [4]. For sequence data, rare patterns have not seen as much attention, but are related to *unwords* [15] in genome research (i.e. absent words that are not subsequences of any other absent word), and *antipatterns* [10] in music (patterns that are surprisingly rare in a corpus of music pieces). Antipatterns may represent structural constraints of a music style and can therefore be useful for classification and generation of new pieces.

The Ethiopian lyre *bagana* is played by the Amhara, inhabitants of the Central and Northern part of the country. The bagana is a large lyre, equipped with ten gut strings, most of them plucked with the fingers. According to Amhara tradition, the bagana is the biblical instrument played by King David and was

S. Džeroski et al. (Eds.): DS 2014, LNAI 8777, pp. 62–72, 2014.

brought to Ethiopia, together with the Ark of Covenant, by the legendary Emperor Menelik, mythical son of King Solomon and Queen of Sheba. The bagana belongs to the spiritual sphere of Amhara music, even though it is not played during liturgical ceremonies. Because of its mythical origin and connection to the divine, the bagana is highly respected, as the instrument of kings and nobles, played by pious men and women of letters [26].

The analysis of the learning process used by the most revered player, Alemu Aga, has shown that the first phase of this process is based on exercises composed of short motifs [25]. These exercises correspond, according to Alemu Aga, to motifs that are either frequently or rarely encountered in his real bagana songs. They are meant to familiarize the student with the playing technique, the numbered notational system (see below) as well as with the sound colour of the instrument, which is, due to its buzzing quality, unique in the Amhara musical systems.

The study of a bagana corpus provides a unique opportunity for evaluation of pattern discovery techniques, because there exist known rare motifs that also have functional significance. The aim of this paper is to explore whether sequential pattern discovery methods, specifically methods for the discovery of rare or absent patterns in music [10], can reveal the known rare patterns and possibly other rare patterns in a corpus of bagana songs.

Fig. 1. The Ethiopian lyre bagana

2 Bagana Background

This section provides some background on the Ethiopian bagana, presenting how the fingers are assigned to strings, the tuning and scales of the bagana, and finally the encoding of a corpus of bagana songs for computational analysis.

2.1 Bagana Notation

The bagana has 10 strings which are plucked by the left hand, with the fingers numbered from 1 (thumb) to finger 5 as described in Table 1.

Table 1. Fingering of the bagana, with finger numbers assigned to string numbers

string	1	2	3	4	5	6	7	8	9	10
fingering	1	r	$2'$	2	r	3	r	4	r	5

In Table 1, "r" (for "rest") indicates a string that is not played, but rather is used as a rest for the finger after it plucks the string immediately next to it. Strings 3 and 4 are both played by finger number 2 (string 3 being therefore notated as finger $2'$), otherwise the assignment of finger number to string number is fixed.

Table 2. Tuning of the bagana, in two different scales, and the nearest Western tempered note corresponding to the degrees of the scales

finger	1	$2'$ or 2	3	4	5
string	1	3 or 4	6	8	10
scale tezeta	E/F	C	D	A	G
scale anchihoye	F	C	D♭	G♭	A
scale degree	$\hat{3}$	$\hat{1}$	$\hat{2}$	$\hat{5}$	$\hat{4}$

As with the other Amhara instruments, the bagana is tuned to a traditional pentatonic scale. Usually, the player chooses between the tezeta scale and the anchihoye scale. Tezeta is anhemitonic (without semitones) and is relatively close to Western tempered degrees (see Table 2). Anchihoye, however, is more complex and comprises two intervals smaller than the Western tempered tone.

To illustrate the notation in Table 2, for example, in the tezeta scale the ascending pentatonic scale (C, D, F, G, A) would be notated by the sequence (2, 3, 1, 5, 4). The scale degree notation is also useful to measure the diatonic interval between two strings, as will be used in Section 4.

Figure 2 shows the placement of the left hand on the 10 bagana strings, along with the information from Table 2: the finger numbering used, and the notes played by the fingers.

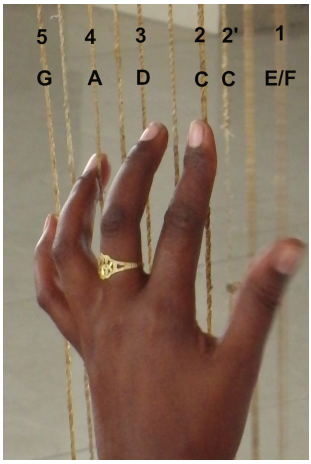

Fig. 2. Placement of left hand on the strings of the bagana

2.2 Bagana Corpus

Bagana songs, also called *yebagana mezmurotch* in Amharic, are based on a relatively short melody, repeated several times with different lyrics, except for the refrain (*azmatch*) for which the lyrics do not vary. Bagana songs are usually preceded by instrumental preludes, called *derdera* (pl. *derderotch*). The analyzed corpus comprises 29 melodies of bagana songs performed by 7 players (5 men, 2 women), and 8 derderotch. These 37 pieces were recorded by Weisser [25] between 2002 and 2005 in Ethiopia (except for 2 of them recorded in Washington DC). In this paper, no differentiation will be made between derdera and bagana songs. A total of 1903 events (finger numbers) are encoded within the 37 pieces (events per song: $\mu = 51$, $\sigma = 30$, min $= 13$, max $= 121$). Figure 3 shows an example of a fragment of a bagana song, encoded as a sequence of finger numbers, corresponding to the fingering of the song.

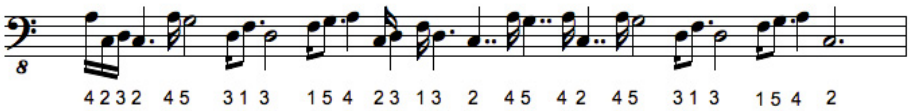

Fig. 3. A short fragment encoded in score and finger notation, from the beginning of the song Abatachen Hoy ("Our Father"), one of the most important bagana songs, as performed by Alemu Aga (voice not shown). Transcribed in 2006 (see [27]).

2.3 Rare Motifs

Table 3 shows four motifs that correspond, according to the bagana master Alemu Aga, to motifs that are rarely encountered in his real bagana songs and are used during practice to strengthen the fingers with unusual finger configurations [25]. The first two motifs in Table 3 are short bigram patterns. In Section 4 it will be explored whether these two rare patterns can be discovered from corpus analysis. The third and fourth motifs of Table 3 (bottom) correspond to longer pentagram patterns that form ascending and descending pentatonic scales and are also used for didactic purposes. Since most pentagram patterns will be rare in a small corpus, an additional question that will be explored in Section 4 is whether these two pentagram patterns are surprisingly rare.

Table 3. Rare motifs, from [25], page 50

	Motifs in numeric notation
First exercise	$(1, 4)$
Second exercise	$(1, 2)$
Third exercise	$(2, 3, 1, 5, 4)$
Fourth exercise	$(4, 5, 1, 3, 2)$

3 Antipattern Mining

In this work we apply data mining to discover antipatterns in the bagana corpus. The task is an instance of *supervised descriptive rule discovery* [20], a relatively new paradigm for data mining which unifies the areas of subgroup discovery [17], contrast set mining [6,12], and emerging pattern mining [13].

Referring to Figure 4, in the supervised descriptive mining paradigm, data may be partitioned into two sets, an analysis class $\oplus$ with $n^{\oplus}$ objects, and a background set $\ominus$ with $n^{\ominus}$ objects. The partitioning is flexible and the background set may contain instances labelled with multiple different classes. A *pattern* is a predicate that is satisfied by certain data objects. The number of occurrences of a pattern P in the set $\oplus$ is given by $c_P^{\oplus}$, and in the set $\ominus$ by $c_P^{\ominus}$. The goal is to discover patterns predictive of the $\oplus$ class, covering as few of the $\ominus$ objects as possible. If under- rather than over-represented patterns are desired, the reversal of the roles of the analysis and background classes $\ominus$ and $\oplus$ can naturally lead to the discovery of patterns frequent in $\ominus$ and rare or absent from $\oplus$ [10]. In this case the inner box of Figure 4 would be shifted downwards into the $\ominus$ region.

In the original studies of subgroup discovery and contrast data mining [17,6,13] objects and subgroups are described using attribute-value representations. Later work has shown that contrast data mining can be applied to sequence data: Ji et al. [16] consider *minimal distinguishing subsequence patterns* and Deng and Zaïane [11] consider *emerging sequences* (sequential patterns frequent in one group but infrequent in another).

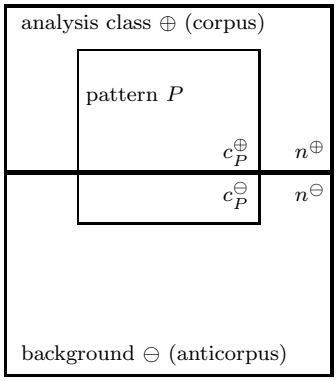

Fig. 4. The schema for contrast set mining, showing the major regions of objects involved. The top part of the outer box encloses data labelled with the class of interest, below this the background. The inner box contains the objects described by a pattern, and the top part of the inner box the contrast set described by a discovered pattern.

Music can be represented as sequences of events for the purposes of supervised descriptive data mining. In music the analysis and background set are called the *corpus* and *anticorpus* (Figure 4). In pattern discovery in music, the counting of pattern occurrences can be done in two ways [9]: either by considering *piece count* (the number of pieces containing the pattern one or more times, i.e. analogous to the standard definition of pattern support in sequential pattern mining) or by considering *total count* (the total number of positions matched by the pattern, also counting multiple occurrences within the same piece). The latter is used when a single music piece is the target of analysis. For the bagana, even though several pieces are available, total count is used, because we consider that a motif is frequent (or rare) if it is frequently (or rarely) encountered within any succession of events. Therefore in this study the set $\oplus$ (resp., $\ominus$) comprises all *suffixes* in the corpus (anticorpus), and practically $n^{\oplus}$ ($n^{\ominus}$) is therefore the total number of suffixes in the corpus (anticorpus), and $c_P^{\oplus}$ ($c_P^{\ominus}$) the total number of sequences in $\oplus$ ($\ominus$) for which the pattern P is a prefix. In this study overlapping pattern occurrences are excluded from the total count of a pattern.

For antipattern mining of bagana songs, unlike for previous antipattern mining studies with Basque folk songs [10], an interesting and important feature is that there is no naturally available anticorpus to contrast with the corpus. Therefore a different method was needed to reveal those patterns whose count within the corpus is significantly low, evaluated here using a binomial distribution of pattern counts along with a zero-order model of finger numbers to compute the pattern probabilities.

3.1 Patterns and Expectation

In this study of the bagana corpus, a *pattern* is a contiguous sequence of events, represented by finger numbers. To contrast the occurrences of a pattern between

a set $\oplus$ and a set $\ominus$, the *empirical background probability* of a pattern $P = (e_1, \ldots, e_\ell)$ may be computed simply as $c_P^{\ominus}/n^{\ominus}$, if a large background set $\ominus$ is available [13]. Without a large background corpus, the background probability of the pattern must be estimated analytically, for example using a zero-order model of the corpus:

$$b_P = \prod_{i=1}^{\ell} c_{e_i}^{\oplus}/n^{\oplus}$$

where $c_{e_i}^{\oplus}$ is the total count of event e_i, and $n^{\oplus}$ is the total number of events in the corpus. The background probability b_P therefore gives the probability of finding the pattern in exactly ℓ contiguous events.

A useful quantity derived from the background probability is the *expected total count*. Letting X be the random variable that models the total count of a pattern P, the expected total count is:

$$\mathbb{E}(X) = b_P \times t_P$$

where t_P is the maximum number of non-overlapping positions that can be possibly matched by the pattern, approximated here simply by $n^{\oplus}/\ell$.

In this study a zero-order analytic model of the corpus is used to permit the detection of over- or under-representation in bigram or longer patterns. A first- or higher-order analytic model would not be able to detect bigram patterns because the expected total count of a pattern would be equivalent to its actual count.

3.2 Antipatterns and Statistics

An *antipattern* is a pattern that is rare, or even absent, in a corpus. For data mining, this definition is not operational because almost any sequence of events is an antipattern, that is, most possible event sequences will never occur in a corpus, with their count rapidly falling to zero with increasing length. Most of these patterns are not interesting because it is expected that their total count is zero. Therefore we want to know which are the *significant* antipatterns: those that are *surprisingly* rare or absent from a corpus.

Antipatterns are evaluated according to a p-value, which gives the probability of finding an equal or fewer number of occurrences than the number observed. Low p-values are desired, because it means such patterns are surprisingly rare in the corpus. The p-value of finding $c_P^{\oplus}$ or fewer occurrences in the corpus is modelled using the binomial distribution:

$$\mathbb{P}\big(X \leq c_P^{\oplus}\big) = \sum_{i=0}^{c_P^{\oplus}} \mathbb{B}\big(i, t_P, b_P\big) \tag{1}$$

where $\mathbb{B}\big(i, t_P, b_P\big)$ is the binomial probability of finding exactly i occurrences of the pattern P, in t_P possible placements of the pattern, and b_P is the background probability of the pattern. Low p-values indicate patterns that are statistically surprising and therefore potentially interesting.

3.3 Discovery Algorithm

The antipattern discovery task is stated simply as: given a corpus, and a significance level α, find all patterns P with a p-value (Equation 1) of at most α:

$$\mathbb{P}\big(X \leq c_P^{\oplus}\big) \leq \alpha \tag{2}$$

Furthermore, for presentation we consider only those significant antipatterns that are *minimal*, that is, those that are not contained within any other significant antipattern [10].

The discovery of minimal significant antipatterns can be efficiently solved by a refinement search of pattern space [8,10], using a method similar to the SPAM algorithm [5]. A depth-first search starts at the most general (empty) pattern, and the search at a particular node of a search tree is continued only while the pattern is not significant. In this work only the S-step refinement operator [5], which extends a sequential pattern on the right hand side by one element, is used: an I-step is not necessary because events here have only one feature (the finger number).

The complexity of the antipattern discovery algorithm is determined by the significance level α, because with low α the search space must be more deeply explored before a significant pattern is reached. Nevertheless, similar to the statistical significance pruning method of Bay and Pazzani [6] who evaluate contrast sets using a χ^2 statistic, it is possible to compute the minimal p-value (Equation 1) achievable on a search path. This can lead to the pruning of entire paths that will not visit a pattern meeting the significance level of α.

4 Results and Discussion

The pattern discovery method described in Section 3 was used to find all minimal antipatterns at the significance level of $\alpha = 0.01$ (Equation 2). The method revealed exactly ten significant antipatterns (Table 4): five patterns and their retrogrades (reversal). The third column shows the total count of the pattern, and in brackets their piece count (number of pieces containing the pattern one or more times). Interestingly, all minimal antipatterns are bigrams. The two patterns (4,1) and (2,1) presented at the top part of Table 4 (with their retrogrades, which are also significant) are the *most significant* antipatterns discovered, and correspond to the retrogrades of two of the didactic rare motifs (Table 3).

The second column of Table 4 presents the undirected diatonic interval formed by the pattern, in the tezeta scale (Table 2). Interestingly, all of the discovered antipatterns form a melodic interval of a major third or greater (P4, M3, and P5).

It is worth noticing that these results prove the rarity of the interval of fifths (perfect in tezeta, diminished and augmented in anchihoye). According to authoritative writings in ethnomusicology [3], the perfect fifth and the cycle of fifths play a founding role in anhemitonic pentatonic scales such as tezeta. The rarity of the actual fifths in the songs is therefore significant, and it can be speculated

Table 4. Bagana antipatterns discovered at $\alpha = 0.01$. Top: known rare bigram patterns; middle: novel rare patterns; bottom: pentagram patterns of Table 3

P	diatonic interval	$c_P^{\oplus}$	$\mathbb{E}(X)$	p-value
(4,1)	P4	2 (2)	48	6.3e-19
(1,4)		21 (14)	48	7.5e-06
(2,1)	M3	6 (5)	50	1.8e-15
(1,2)		13 (7)	50	2.6e-10
(4,3)	P5	2 (2)	30	3.8e-11
(3,4)		3 (3)	30	4.0e-10
(5,2)	P5	17 (10)	38	6.6e-05
(2,5)		16 (9)	38	2.7e-05
(3,5)	P4	5 (4)	26	5.7e-07
(5,3)		11 (8)	26	0.00077
(2,3,1,5,4)	ascending scale	8 (7)	0.11	1
(4,5,1,3,2)	descending scale	4 (3)	0.11	1

that this interval is a mental reference that is never (and does not necessarily need to be) performed. Similarly, the (3,5), a perfect fourth, is the inversion, i.e. the interval to be added to another one to constitute an octave. Intervals and their inversions are usually connected in several musical cultures, including Western art music.

For completeness with the results of Weisser [25], at the bottom of Table 4 are the two pentagram patterns from Table 3. As expected, these patterns are not frequent, but surprisingly they are not significant according to their p-value (Equation 1) (therefore they are not found by the pattern discovery method). In fact, they occur in the corpus many times more than their expected total count.

From a small corpus of bagana songs, antipattern discovery is able to find the two published rare bigram motifs. The results suggest several directions for future studies. The novel antipatterns (3,4), (2,5), and (3,5) (with their retrogrades) found by the method (Table 4) may have new implications for the study of didactic and distinctive motifs of the bagana. The study of these patterns is left for future work. Further, in this study only melodic aspects have been considered, and not rhythmic aspects. Though good results have been obtained with melodic information only, rhythmic patterns could also be of interest, especially when linked with melodic aspects, although data sparsity for this corpus may become an issue. A transcription and encoding of rhythmic information for the corpus from bagana recordings is in progress. It is also planned to explore positive pattern as well as antipattern discovery, partitioning the corpus in different interesting ways, for example according to performer and the scale employed in a performance. Finally, the use of antipatterns as structural constraints during the process of generating new bagana song instances will be explored.

Acknowledgements. This research is 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. Special thanks to Kerstin Neubarth, Dorien Herremans, and Louis Bigo for valuable comments on the manuscript.

References

1. Adamo, J.M.: Data Mining for Association Rules and Sequential Patterns. Springer (2001)
2. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Proceedings of the Eleventh International Conference on Data Engineering, Washington, DC, pp. 3–14 (1995)
3. Arom, S.: Le "syndrome" du pentatonisme Africain. Musicae Scientiae 1(2), 139–163 (1997)
4. Artamonova, I., Frishman, G., Frishman, D.: Applying negative rule mining to improve genome annotation. BMC Bioinformatics 8, 261 (2007)
5. Ayres, J., Gehrke, J., Yiu, T., Flannick, J.: Sequential pattern mining using a bitmap representation. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, pp. 429–435 (2002)
6. Bay, S., Pazzani, M.: Detecting group differences: Mining contrast sets. Data Mining and Knowledge Discovery 5(3), 213–246 (2001)
7. Conklin, D.: Melody classification using patterns. In: MML 2009: International Workshop on Machine Learning and Music, Bled, Slovenia, pp. 37–41 (2009)
8. Conklin, D.: Discovery of distinctive patterns in music. Intelligent Data Analysis 14(5), 547–554 (2010)
9. Conklin, D.: Distinctive patterns in the first movement of Brahms' String Quartet in C Minor. Journal of Mathematics and Music 4(2), 85–92 (2010)
10. Conklin, D.: Antipattern discovery in folk tunes. Journal of New Music Research 42(2), 161–169 (2013)
11. Deng, K., Zaïane, O.R.: Contrasting sequence groups by emerging sequences. In: Gama, J., Costa, V.S., Jorge, A.M., Brazdil, P.B. (eds.) DS 2009. LNCS, vol. 5808, pp. 377–384. Springer, Heidelberg (2009)
12. Dong, G., Bailey, J. (eds.): Contrast Data Mining: Concepts, Algorithms, and Applications. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC (2012)
13. Dong, G., Li, J.: 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 1999, San Diego, pp. 43–52 (1999)
14. Haglin, D.J., Manning, A.M.: On minimal infrequent itemset mining. In: Stahlbock, R., Crone, S.F., Lessmann, S. (eds.) Proceedings of the 2007 International Conference on Data Mining, Las Vegas, Nevada, USA, pp. 141–147. CSREA Press (2007)
15. Herold, J., Kurtz, S., Giegerich, R.: Efficient computation of absent words in genomic sequences. BMC Bioinformatics 9, 167 (2008)
16. Ji, X., Bailey, J., Dong, G.: Mining minimal distinguishing subsequence patterns with gap constraints. Knowledge and Information Systems 11(3), 259–296 (2007)
17. Klösgen, W.: Explora: A Multipattern and Multistrategy Discovery Assistant. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (eds.) Advances in Knowledge Discovery and Data Mining, pp. 249–271. MIT Press, Cambridge (1996)

18. Lin, C.-R., Liu, N.-H., Wu, Y.-H., Chen, A.L.P.: Music classification using significant repeating patterns. In: Lee, Y., Li, J., Whang, K.-Y., Lee, D. (eds.) DASFAA 2004. LNCS, vol. 2973, pp. 506–518. Springer, Heidelberg (2004)
19. Mooney, C.H., Roddick, J.F.: Sequential pattern mining – approaches and algorithms. ACM Comput. Surv. 45(2), 19:1–19:39 (2013)
20. Novak, P.K., Lavrač, N., Webb, G.I.: Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. Journal of Machine Learning Research 10, 377–403 (2009)
21. Sawada, T., Satoh, K.: 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 (2000)
22. Shan, M.K., Kuo, F.F.: Music style mining and classification by melody. IEICE Transactions on Information and Systems E88D(3), 655–659 (2003)
23. Wang, J., Zhang, Y., Zhou, L., Karypis, G., Aggarwal, C.C.: CONTOUR: an efficient algorithm for discovering discriminating subsequences. Data Min. Knowl. Disc. 18, 1–29 (2009)
24. Webb, G.I.: Discovering significant patterns. Machine Learning 71(1), 131 (2008)
25. Weisser, S.: Etude ethnomusicologique du bagana, lyre d'Ethiopie. Ph.D. thesis, Université Libre de Bruxelles (2005)
26. Weisser, S.: Ethiopia. Bagana Songs. Archives Internationales de Musique Populaire (AIMP)-VDE Gallo, Liner notes of the CD (2006)
27. Weisser, S.: Transcrire pour vérifier: le rythme des chants de bagana d'Éthiopie. Musurgia XIII(2), 51–61 (2006)