Proceedings of the Third International Workshop on Folk Music Analysis, June 6-7, 2013. Amsterdam, Netherlands. P. van Kranenburg, C. Anagnostopoulou, and A. Volk (Eds.).
Amsterdam: Meertens Institute, Utrecht: Department of Information and Computing Sciences, Utrecht University.

DESCRIPTIVE RULE MINING OF BASQUE FOLK MUSIC

Kerstin Neubarth[1,2] **Colin G. Johnson**[2] **Darrell Conklin**[3,4]

[1]Canterbury Christ Church University, Canterbury, United Kingdom
[2]School of Computing, University of Kent, Canterbury, United Kingdom
[3]Department of Computer Science and Artificial Intelligence,
University of the Basque Country UPV/EHU, San Sebastián, Spain
[4]IKERBASQUE: Basque Foundation for Science, Bilbao, Spain

1. INTRODUCTION

As early as in the 1950s, Bronson (1959) proposed a computational approach to address typological or geographical questions about folk music, such as: "What are the characteristic differentiae of specific regions? Are there rhythmical preferences? Modal preferences? And to what degree of intensity?" One way to represent such regional or typological preferences computationally is using rules. Within predictive data mining of folk music, in particular classification, most studies (e.g. Bohak & Marolt, 2009; Hillewaere et al., 2009; Conklin, 2013) focus on global classification performance and do not report individual rules, which in isolation may not have high accuracy but nevertheless could provide answers to Bronson's questions. For classification of tune families, a recent study (van Kranenburg et al., 2013) indicates the discriminative power of features, but does not systematically specify feature values and the partitioning of the feature space. Applications of descriptive mining to folk music include subgroup discovery (Taminau et al., 2009) and distinctive pattern discovery (Conklin & Anagnostopoulou, 2011). These studies highlight musical characteristics – global features or interval patterns – that are over-represented in a geographical region or in a genre relative to their distribution in the total corpus. Metadata of a folk music collection has been mined to extract qualified associations between folk music genres and geographical regions (Neubarth et al., 2012).

Here the method of the previous study (Neubarth et al., 2012) is extended to mine for associations between musical content (global features computed from midi files) and folk music genres or regions. Association rule mining (e.g. Srikant & Agrawal, 1995) is adapted to identify different categories of rules, covering both over- and under-representation of content characteristics in genres or regions. The aim of our research is to extract rules which reflect musicological observations such as "Western melodies are largely in triple metre", "ballads rarely have unmeasured rhythms" or "the pentatonic system is not found in children's songs" (Sadie, 2001). While such statements are prominent in folk music surveys, they have not been captured by existing approaches to computational folk music analysis. The method proposed here explicitly distinguishes different relations between music content and genres or regions.

2. DATA

As a corpus we use the *Cancionero Vasco*, an iconic collection of Basque folk songs and dances. Its musicologically curated metadata includes information on the genre of a folk tune (e.g. wedding song or work song) and the geographical location where it was collected. The corpus thus offers an opportunity to analyse both song types or genres and regions, for the same corpus.

The digitised collection contains 1902 midi files. Of the 1902 tunes in the corpus, 1561 are annotated with a folk music genre and 1630 are annotated with a location. The metadata vocabulary consists of 50 genre labels and 2968 geographical labels. Both genres and geographical regions are hierarchically organised (Goienetxea et al., 2012).

Music content features were selected from an existing feature set (McKay & Fujinaga, 2006), such that the selected subset reflects content characteristics in folk music surveys (Sadie, 2001). These features were computed with jSymbolic (McKay & Fujinaga, 2006). Another three features were additionally implemented: pitch class entropy, interval perplexity and duration perplexity. Numeric features were discretised, using Weka (Hall et al., 2009), and transformed into string content items; discretisation bins are based on musicological statements (Sadie, 2001) and the distribution of feature values in the corpus.

3. METHOD

To analyse content–genre and content–region associations of different categories, we proceeded in four steps: identification of association categories; translation of the categories into association constraints; mining for association rules which meet the defined constraints; and post-processing of hierarchical rules.

Association categories were identified through a qualitative analysis of 25 reference articles on European folk music (Sadie, 2001). Statements linking music content and genres or regions were extracted and grouped according to similar meaning. This resulted in nine association categories. The categories were given an interpretation in terms of over- and under-representation. For example, a content feature can be over-represented in a genre or region with respect to its occurrence in other genres or regions (category *Primarily*), or a content feature can be under-

represented in a genre or region with respect to other content features in the same genre or region (category *Uncommon*). The category interpretations were translated into association constraints: specific combinations of rule templates (Klemettinen et al., 1994) and rule evaluation measures (Geng & Hamilton, 2006; Lenca et al., 2008).

During the mining, item sets are formed as pairs between a content item and a genre or region item. Pairs are evaluated against the constraints of each category. If a candidate pair meets the constraints, a rule is added to the results. For each rule a p-value is calculated according to Fisher's one-tailed exact test. The lower the p-value, the less expected is the number of encountered co-occurrences between a content item and genre or region, given their distributions in the entire corpus.

As the items are hierarchically organised, the discovered rules are partly redundant. In a post-processing step we thus prune or group more specific rules relative to their parent rules, depending on whether they confirm, specify or deviate from the parent rule (Liu et al., 2000).

The method outputs a structured list of qualified associations, which provide an overview of the corpus and highlight content–genre or content–region patterns for further musicological exploration.

4. RESULTS

Traditional association rule mining using the support/confidence framework (e.g. Srikant & Agrawal, 1995) usually yields a large set of rules in the following form:

> Araba $\rightarrow$ high pitch class entropy;
> $s = 26$; $c = 0.96$

In the corpus, 96% of the Araba tunes have high pitch class entropy (confidence c), and there are 26 such tunes (support s).

Association categories provide an additional qualification based on natural language and support different views, e.g. focusing on musical characteristics of a region or focusing on the regional distribution of content features, for example:

> Araba, high pitch class entropy: *Usually*
> (template $R \rightarrow C$; $c = 0.96$; $p = 0.00003$)
>
> Navarra, low pitch class entropy: *Primarily*
> (template $C \rightarrow R$; $c = 0.52$; $p = 0.006$)

The first rule describes a region and an aspect of its musical character (template $R \rightarrow C$): within the *Cancionero Vasco*, tunes from Araba usually have high pitch class entropy. The second rule highlights a content characteristic and its regional occurrence (template $C \rightarrow R$): more than half of the tunes with low pitch class entropy in the corpus (52%) is concentrated in one of the seven provinces, i.e. tunes with low pitch class entropy are primarily found among the songs collected in Navarra.

By exploiting the hierarchical organisation of the item vocabulary, the post-processing allows to distinguish general rules, contributing, specialised and deviating sub-rules, so that resulting rule sets are structured, for example:

> life-cycle songs, narrow intervals: *Present*
> (template $G - C$; $s = 251$; $p = 0.0998$)
>
> Contributing:
>
> > love songs, narrow intervals: *Present*
> > (template $G - C$; $s = 116$; $p = 0.8038$)
>
> Specialised:
>
> > lullabies, narrow intervals: *Usually*
> > (template $G \rightarrow C$; $c = 0.65$; $p = 0.00099$)
>
> Deviating:
>
> > wedding songs, narrow intervals: *Absent*
> > (template $G \rightarrow \neg C$; $c = 1.0$; $p = 0.0529$)

Average narrow intervals are present in life-cycle songs, and within life-cycle songs are found in love songs. While life-cycle songs *may* move in narrow intervals, lullabies *usually* move in narrow intervals: about two thirds of the lullabies in the corpus ($c = 0.65$) have average narrow intervals. In the *Cancionero Vasco* tunes with average narrow intervals are absent among wedding songs, which form a sub-genre of life-cycle songs: all the wedding songs in the corpus ($c = 1.0$) have average melodic intervals other than narrow ($G \rightarrow \neg C$).

5. CONCLUSIONS

Recent supervised approaches to computational folk music analysis have focused on comparing predictive methods (e.g. for region, genre or tune family classification, Conklin, 2013; Hillewaere et al., 2009; van Kranenburg et al., 2013). By contrast, we deliberately chose a descriptive approach, within a knowledge discovery paradigm, in order to extract understandable rules in response to musicological questions such as those formulated by Bronson (1959). Knowledge discovery advocates a high level of interaction between the computational design and appreciation of the application domain (e.g. Fayyad et al., 1996). In our study a qualitative analysis of folk music surveys informs the task specification, selection and discretisation of content features, definition of association categories and presentation of results.

This research extends earlier work on descriptive mining of folk music. In a previous study, categorised genre–region associations were extracted from the metadata of the *Cancionero Vasco* (Neubarth et al., 2012); here we demonstrate that the method can also identify musical characteristics of folk music genres and regions and thus link music content and metadata. The subgroup discovery study by Taminau et al. (2009) was restricted to one form of content–region rules, which corresponds to mining rules with one template ($C \rightarrow R$) and one evaluation measure (weighted relative accuracy). The association mining approach presented here provides additional flexibility to discover rules of different categories by combining several rule templates and evaluation measures. Finally, our analysis goes beyond previous research (Conklin & Anagnostopoulou, 2011; Neubarth et al., 2012) by taking into account the hierarchical structure of the metadata.

6. ACKNOWLEDGEMENTS

We thank Fundación Euskomedia and Fundación Eresbil, Spain, for providing the *Cancionero Vasco* for study.

7. REFERENCES

Bohak, C. & Marolt, M. (2009). Calculating similarity of folk song variants with melody-based features. In *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, (pp. 597–601), Kobe, Japan.

Bronson, B. H. (1959). Toward the comparative analysis of British-American folk tunes. *The Journal of American Folklore*, *72*(284), 165–191.

Conklin, D. (2013). Multiple viewpoint systems for music classification. *Journal of New Music Research*, *42*(1), 19–26.

Conklin, D. & Anagnostopoulou, C. (2011). Comparative pattern analysis of Cretan folk songs. *Journal of New Music Research*, *40*(2), 119–125.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge discovery and data mining: towards a unifying framework. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, (pp. 82–88), Portland, Oregon, USA.

Geng, L. & Hamilton, H. J. (2006). Interestingness measures for data mining: a survey. *ACM Computing Surveys*, *38*(3), 1–32.

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

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The Weka data mining software: An update. *SIGKDD Explorations*, *11*(1), 10–18.

Hillewaere, R., Manderick, B., & 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)*, (pp. 729–733), Kobe, Japan.

Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., & Verkamo, A. I. (1994). Finding interesting rules from large sets of discovered association rules. In *Proceedings of the 3rd International Conference on Information and Knowledge Management*, (pp. 401–407), Gaithersburg, Maryland.

Lenca, P., Meyer, P., Vaillant, B., & Lallich, S. (2008). On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid. *European Journal for Operational Research*, *184*(2), 610–626.

Liu, B., Hu, M., & Hsu, W. (2000). Multi-level organization and summarization of the discovered rules. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2000)*, (pp. 208–217), Boston, MA.

McKay, C. & Fujinaga, I. (2006). jSymbolic: A feature extractor for MIDI files. In *Proceedings of the International Computer Music Conference*, (pp. 302–305), New Orleans, USA.

Neubarth, K., Goienetxea, I., Johnson, C. G., & 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)*, (pp. 7–12), Porto, Portugal.

Sadie, S. (Ed.). (2001). *New Grove Dictionary of Music and Musicians*. London: Macmillan.

Srikant, R. & Agrawal, R. (1995). Mining generalized association rules. In *Proceedings of the 21st VLDB Conference*, (pp. 407–419), Zurich, Switzerland.

Taminau, J., Hillewaere, R., Meganck, S., Conklin, D., Nowé, A., & Manderick, B. (2009). Descriptive subgroup mining of folk music. In *2nd International Workshop on Machine Learning and Music (MML 2009)*, Bled, Slovenia.

van Kranenburg, P., Volk, A., & 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.