Discovery of mediating association rules for folk music analysis

Kerstin Neubarth\textsuperscript{1,2}, Colin G. Johnson\textsuperscript{2}, Darrell Conklin\textsuperscript{3,4}

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

Abstract. Many folk music collections are organised by parallel ontologies of geographical regions and folk music genres. When folk music corpora are used in music data mining, genres and regions are generally analysed separately. This study proposes a descriptive rule mining method for discovering patterns which link genre and region rules. Given pairwise associations between music content features, genres and regions, the method identifies combinations of content–region, content–genre and region–genre rules such that the content–region rule may be redundant.

Keywords: association rule mining, descriptive rule mining, computational folk music analysis

1 Introduction

In computational folk music analysis, predictive and descriptive data mining have been applied to learn models and rules which reveal underlying structures in folk music corpora. Supervised mining tasks, where folk tunes are labelled, include classification into folk music genres (Conklin, 2013), regions of origin or collection (Hillewaere et al., 2009) and tune families (van Kranenburg et al., 2013), pattern discovery for genres and regions (Conklin and Anagnostopoulou, 2011) and subgroup discovery for regions (Taminau et al., 2009). Even when a corpus provides both genre and region metadata, the two tasks of genre and region mining have been treated separately.

This paper introduces a method to link genre and region rules. The method is applied to two folk music corpora. Previous work using these corpora showed that classification accuracy was substantially lower for region than for genre classification (Conklin, 2013), suggesting that in these repertoires music content may be more predictive of genres than of regions. The findings raise the question whether region classification may be partially mediated by genres, especially if regions and genres are strongly associated (Hillewaere, 2013). In the context of subgroup discovery, Taminau et al. (2009), manually inspecting discovered rules, speculated that the over-representation of 3/4 meter in Scandinavian tunes could reflect the dominance of triple-meter polskas in the Scandinavian sample.
The method proposed here strives to automate the discovery of such patterns: adapting the definition of redundant rules by Shah et al. (1999), the method finds triples of content–region, content–genre and region–genre rules such that the content–region rule may be redundant. In this way content–genre and region–genre rules can assist with the interpretation of content–region rules.

2 Data and method

The Cancionero Vasco contains 1902 tunes of five folk music genres, collected in seven Basque provinces. The Europa-6 corpus (Hillewaere et al., 2009) consists of 3367 tunes of nine folk dance types from six European countries. To represent musical content, 19 global features are computed from the MIDI files, selected from existing feature sets (McKay and Fujinaga, 2006; Eerola and Toiviainen, 2004; Müllensiefen, 2009; Steinbeck, 1982). Numeric features are discretised into categorical values, resulting in 63 content items for the Cancionero Vasco and 84 content items for Europa-6.

In a first step, pairwise association rule mining is applied to discover content–region, region–genre and content–genre rules. In this study positive and negative association rules are considered. A pairwise association rule is a rule $x \rightarrow y$ ($x \rightarrow \neg y$), where $x$ and $y$ are items. The support of the rule is the number of instances – here folk tunes – annotated with both items $x$ and $y$ (annotated with $x$ but not $y$), and the confidence of the rule is the proportion of instances annotated with $x$ which are also annotated with $y$ (which are not annotated with $y$). For each rule a $p$-value is calculated according to Fisher’s one-tailed exact test, with right tail for positive rules and left tail for negative rules.

In a second step, the discovered rules are mined for rule triples: given a content–region rule, the method searches for combinations of content–genre and region–genre rules such that the content–region rule may be redundant, i.e. one of the following constellations applies, where $C$ denotes a content item, $R$ a region item and $G$ a genre item:

1. $C \rightarrow R$
   $C \rightarrow G$
   $G \rightarrow R$

2. $R \rightarrow C$
   $R \rightarrow G$
   $G \rightarrow C$

3. $C \rightarrow \neg R$
   $C \rightarrow G$
   $G \rightarrow \neg R$

4. $R \rightarrow \neg C$
   $R \rightarrow G$
   $G \rightarrow \neg C$

That is, a content–region rule may be mediated by a genre if the content feature is over-represented in the region, over-represented in a certain genre, and that genre is over-represented in the region (cases 1 and 2); or if the content feature is under-represented in the region, over-represented in a certain genre, and that genre is under-represented in the region (case 3); or if the content feature is under-represented in the region, under-represented in a certain genre, and that genre is over-represented in the region (case 4).
Table 1. Number of discovered rules (CR: content–region rules; CG: content–genre rules; RG: region–genre rules) and of mediating rule triples for the four constellations.

<table>
<thead>
<tr>
<th></th>
<th>Pairwise rules</th>
<th>Mediating rule triples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CR</td>
<td>CG</td>
</tr>
<tr>
<td><strong>Cancionero Vasco</strong></td>
<td>27</td>
<td>51</td>
</tr>
<tr>
<td><strong>Europa-6</strong></td>
<td>160</td>
<td>167</td>
</tr>
</tbody>
</table>

Table 2. Selected rule triples for mediated content–region rules. For each rule are indicated: rule syntax, left-hand side (lhs) and right-hand side (rhs) of the rule, coverage (lhs support), rule support, confidence, and p-value.

<table>
<thead>
<tr>
<th>Template</th>
<th>lhs</th>
<th>rhs</th>
<th>coverage</th>
<th>support</th>
<th>confidence</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C \rightarrow R$</td>
<td>3/4 meter</td>
<td>Sweden</td>
<td>513</td>
<td>390</td>
<td>0.76</td>
<td>9.1e-241</td>
</tr>
<tr>
<td>$C \rightarrow G$</td>
<td>3/4 meter</td>
<td>polska</td>
<td>513</td>
<td>266</td>
<td>0.52</td>
<td>1.6e-230</td>
</tr>
<tr>
<td>$G \rightarrow R$</td>
<td>polska</td>
<td>Sweden</td>
<td>339</td>
<td>269</td>
<td>0.79</td>
<td>2.4e-185</td>
</tr>
<tr>
<td>$R \rightarrow C$</td>
<td>Araba</td>
<td>wide range</td>
<td>27</td>
<td>21</td>
<td>0.78</td>
<td>1.3e-5</td>
</tr>
<tr>
<td>$R \rightarrow G$</td>
<td>Araba</td>
<td>dance</td>
<td>27</td>
<td>24</td>
<td>0.89</td>
<td>7.8e-12</td>
</tr>
<tr>
<td>$G \rightarrow C$</td>
<td>dance</td>
<td>wide range</td>
<td>495</td>
<td>334</td>
<td>0.67</td>
<td>0.0</td>
</tr>
<tr>
<td>$C \rightarrow \neg R$</td>
<td>high register</td>
<td>Navarra</td>
<td>249</td>
<td>175</td>
<td>0.70</td>
<td>1.0e-5</td>
</tr>
<tr>
<td>$C \rightarrow \neg G$</td>
<td>high register</td>
<td>dance</td>
<td>249</td>
<td>179</td>
<td>0.72</td>
<td>0.0</td>
</tr>
<tr>
<td>$G \rightarrow \neg R$</td>
<td>dance</td>
<td>Navarra</td>
<td>495</td>
<td>317</td>
<td>0.64</td>
<td>0.0</td>
</tr>
<tr>
<td>$R \rightarrow \neg C$</td>
<td>Ireland</td>
<td>simple meter</td>
<td>798</td>
<td>565</td>
<td>0.71</td>
<td>0.0</td>
</tr>
<tr>
<td>$R \rightarrow G$</td>
<td>Ireland</td>
<td>jig</td>
<td>798</td>
<td>504</td>
<td>0.63</td>
<td>8.3e-154</td>
</tr>
<tr>
<td>$G \rightarrow \neg C$</td>
<td>jig</td>
<td>simple meter</td>
<td>793</td>
<td>789</td>
<td>0.99</td>
<td>0.0</td>
</tr>
</tbody>
</table>

3 Results and conclusions

The method was run with minimum item support 3 and minimum confidence 0.5. Maximum p-values were calculated as Bonferroni correction at significance level 0.1, to adjust for multiple comparisons: 0.0002, 0.0003 and 0.0021 (Cancionero Vasco) and 0.0002, 0.0001 and 0.0017 (Europa-6) for content–region, content–genre and region–genre rules respectively. Table 1 lists the number of pairwise association rules discovered during the first step and the number of mediating rule triples discovered during the second step, for both corpora. Of the 19 content features, four occur in mediating rule triples for the Cancionero Vasco and ten in mediating rule triples for Europa-6. Table 2 gives one example for each of the constellations. Examples 1 and 4 refer to Europa-6, Examples 2 and 3 to the Cancionero Vasco. The first example corresponds to the hypothesis of Taminau et al. (2009) that in Europa-6 the association of 3/4 meter with Scandinavia, in particular Sweden, may be mediated by the genre of polskas.

Discovering mediating rules can thus support the further analysis of content–region rules. The method goes beyond previous work on descriptive rule mining.
for folk music analysis (e.g. Taminau et al., 2009; Conklin and Anagnostopoulou, 2011) by mining for positive and negative rules of different directions and by linking discovered rules which describe overlapping areas of a folk music corpus.

Acknowledgements

We thank Fundación Euskomedia and Fundación Eresbil (Basque Country, Spain) for providing the Cancionero Vasco for study.

References


