Descriptive Mining of Folk Music
A testcase

Jonatan Taminau[a] Ruben Hillewaere[a] Stijn Meganck[a] Darrell Conklin[b]
Ann Nowé[a] Bernard Manderick[a]

[a] *Computational Modeling Lab, Vrije Universiteit Brussel, Belgium*
[b] *Department of Computing, City University of London, United Kingdom*

1 Introduction

Descriptive analysis of music corpora is important to musicologists who are interested in identifying the properties that characterize specific genres of music. In recent years, the machine learning community has worked on providing predictive tools for genre classification. With music pieces represented as feature vectors, several existing machine learning algorithms can be applied to distinguish between pre-defined classes but most don't offer any interpretable explanation for the classification. Therefore, instead of concentrating on classification, we want to explore the use of descriptive analysis, in particular the technique of Subgroup Discovery, in the context of music.

Subgroup Discovery (SD), first introduced by Klosgen [4] and Wrobel [10], is a rule learning technique located at the intersection of predictive and descriptive induction. The task of SD is defined as: given a population of individuals and a specific property or annotated class of those individuals we are interested in, find population subgroups that are statistically most interesting with respect to this property of interest.

The goal of SD is not to construct a good classification model consisting of a set of rules, but to construct individual rules that identify interesting subsets of related samples. This way each rule can be regarded separately as providing some knowledge about the data.

In this paper we apply a SD algorithm to a corpus of folk tunes, called the *Europa-6* collection [2]. Four well-known global feature sets were joined to represent every piece as a feature vector. The concrete goal is to come up with interpretable rules that describe subgroups in *Europa-6*, which might correspond with musical subgenres.

2 Methods

In this paper we use the CN2-SD algorithm which is detailed in [5]. This algorithm builds rules of the form "if $Cond$ then $Class$" by growing a conjunction of attribute-value pairs one at a time. The quality heuristic used for rule selection is the Weighted Relative Accuracy (WRAcc) defined as:

$$\text{WRAcc}(\text{if } Cond \text{ then } Class) = p(Cond)(p(Class|Cond) - p(Class)) \tag{1}$$

This heuristic provides a tradeoff between generality $p(Cond)$ and relative accuracy $p(Class|Cond) - p(Class)$.

The *Europa-6* collection contains folk tunes from 6 different European regions: England, France, Ireland, Scotland, South-East Europe and Scandinavia. This data consists of purely monophonic melodies in a clean quantized MIDI format, containing time and key signatures, but without any tempo or performance indications such as grace notes, trills, staccato and dynamics. To represent our data, we have chosen to join the relevant features of four global feature sets: Alicante[8], Fantastic[7], Jesser[3] and McKay[6], for a total of 150 features.

3 Results

To perform our experiments we used the component-based data mining software toolkit Orange [1], which includes a widget for CN2-SD. On initial experiments, we found that rules involving numeric attributes dominated the results and were very hard to interpret, therefore we discretized all non-nominal features to either H (High) or L (Low) if the value was higher/lower than the median of this feature in the entire corpus. This discretization has two consequences: the rules are easily interpretable by musicologists and we avoid overfitting the training data. Furthermore, for clarity we only discuss rules of length 2. Longer rules did obtain slightly higher scores based on WRAcc, but rendered the interpretation more difficult.

Based on the WRAcc quality rankings from Orange, we select the best rule for each class and list these 6 rules in the following table:

$Class$ $\mathcal{C}$	$Cond$ A B	$p(A)$	$p(A\|\mathcal{C})$	$p(B)$	$p(B\|\mathcal{C})$	$p(A,B)$	$p(A,B\|\mathcal{C})$	$p(\mathcal{C}\|A,B)$
Scandinavia	J_meter = 3/4 M_MelodicTritones = L	0.15	0.63	0.94	0.96	0.14	0.62	0.78
Ireland	J_dotted = L M_CompoundMeter = 1	0.70	0.86	0.32	0.71	0.23	0.62	0.65
Scotland	J_meter = 4/4 F_int.cont.grad.std = H	0.38	0.77	0.44	0.76	0.17	0.62	0.52
S.E. Europe	J_meter = 7/8 M_AvgVarIOI = H	0.01	0.21	0.39	0.61	0.01	0.21	0.96
France	J_dminthird = L M_Range = L	0.52	0.83	0.43	0.93	0.26	0.77	0.34
England	F_mode = major M_NoteDensity = L	0.77	0.88	0.50	0.63	0.36	0.54	0.43

Because of lack of space, we will only discuss one rule in detail, complete results and more detailed explanations can be found in [9].

Scandinavia: This rule defines a subgroup of pieces, all in 3/4 meter and containing a relatively low number of melodic tritones. It is likely that the restriction of this rule to 3/4 meter reflects the fact that the Scandinavian portion of the corpus is dominated by polska or hambo melodies. The tritone component does not add much to the rule (the low tritone probability is 0.94 in the corpus, and 0.96 in the Scandinavia class).

References

[1] J. Demsar, B. Zupan, and G. Leban. Orange: From experimental machine learning to interactive data mining. White paper, Faculty of Computer and Information Science, University of Ljubljana, 2004.

[2] Ruben Hillewaere, Bernard Manderick, and Darrell Conklin. Global feature versus event models for folk song classification. In *ISMIR 2009: 10th International Society for Music Information Retrieval Conference*, Kobe, Japan, 2009.

[3] B. Jesser. *Interaktive Melodieanalyse*. Peter Lang, Bern, 1991.

[4] W. Klosgen. Explora: A multipattern and multistrategy discovery assistant. *Advances in Knowledge Discovery and Data Mining*, pages 249–271, 1996.

[5] N. Lavrac, B. Kavsek, P. Flach, and L. Todorovski. Subgroup discovery with cn2-sd. *Journal of Machine Learning Research*, pages 153–188, 2004.

[6] C. McKay and I. Fujinaga. Automatic genre classification using large high-level musical feature sets. In *Proceedings of the International Conference on Music Information Retrieval*, pages 525–530, Barcelona, Spain, 2004.

[7] D. Müllensiefen. Fantastic: Feature analysis technology accessing statistics (in a corpus): Technical report v0.9. Technical report, 2009.

[8] P.J. Ponce de Léon and J.M. Iñesta. Statistical description models for melody analysis and characterization. In *Proceedings of the 2004 International Computer Music Conference*, pages 149–156, 2004.

[9] J. Taminau, R. Hillewaere, S. Meganck, D. Conklin, A. Nowe, and B. Manderick. Descriptive subgroup mining of folk music. In *Second International Workshop on Machine Learning and Musicl*, pages 1–6, 2009.

[10] S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *Proceedings of the First European Conference on Principles of Data Mining and Knowledge Discovery*, pages 78–87. MIT Press, 1997.