Descriptive Subgroup Mining of Folk Music

Jonatan Taminau[1⋆], Ruben Hillewaere[1], Stijn Meganck[1], Darrell Conklin[2], Ann Nowé[1], and Bernard Manderick[1]

[1] Computational Modeling Lab (CoMo), Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium
[2] Department of Computing, City University London, United Kingdom

Abstract. Descriptive analysis of music corpora is important to musicologists who are interested in identifying the properties that characterize specific genres of music. In this study we present such an analysis of a large corpus of folk tunes, all labeled by their origin. Subgroup Discovery (SD) is a rule learning technique located at the intersection of predictive and descriptive induction. One of the advantages of using this technique is the intuitive and interpretable result in the form of a collection of simple rules. Classification accuracy is not the goal of this study. Instead, we discuss some of the highest scoring rules with respect to their descriptive power.

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 machine learning algorithms exist that can be applied to distinguish between pre-defined classes. Most of these algorithms have relatively poor performance on musical data and 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 [1] and Wrobel [2] 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.

⋆ To whom correspondence should be addressed. E-mail: jtaminau@vub.ac.be

in MML 2009: International Workshop on Machine Learning and Music, Bled, Slovenia, 1-6, 2009.

In many fields, human experts are not particularly familiar with all aspects of data mining and therefore have problems analysing complicated machine learning algorithms such as the so-called black-box algorithms like neural networks and support vector machines. The rules obtained by the SD algorithm can be interpreted by anyone familiar with the meaning of only the features.

The idea in this paper is to apply the SD algorithm to a corpus of folk tunes, which we call the *Europa-6* collection [3], a large collection of folk songs in MIDI format. Four well-known global feature sets were joined to represent every piece as a feature vector. These features form the background knowledge representation for the algorithm, i.e. the vocabulary used for the rules. The concrete goal is to come up with interpretable rules that describe subgroups in *Europa-6*, which might correspond with musical subgenres that could make the classification task harder. Another interesting outcome of these descriptive rules would be if there were certain features that consistently appear in the rules.

2 Methods

In this section we describe the Subgroup Discovery algorithm, the dataset of folk songs, and the features used in our analysis.

2.1 Subgroup Discovery (SD)

The rules we obtain using SD are of the following form "if $Cond$ then $Class$". $Cond$ is a conjunction of attribute-value pairs and $Class$ is the property of interest. For instance, we wish to retrieve rules like: "if pitchrange = L and silsignum = H then France". This hypothetical rule defines a subgroup of pieces with a small melodic range and a lot of significant silences, and pieces in this subgroup are most likely to be from France.

In this paper we use the CN2-SD algorithm which is detailed in [4]. This algorithm builds rules by growing the $Cond$ one attribute-value pair at a time. It uses a beam search which only keeps the best rules based on a selected quality measure. The quality heuristic used for rule selection is the Weighted Relative Accuracy (WRAcc) defined as:

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

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

To perform our experiments we used the component-based data mining software toolkit Orange [5]. Orange contains CN2 and CN2 Rules Viewer modules, well suited for our analysis. All runs were performed by selecting the WRAcc heuristic with exclusive covering, which means that covered examples are removed from the dataset for subsequent rule discovery. The beam width specifies the width of the search tree used in the original CN2 algorithm [6] and can be used to optimize the balance between on the one hand performance and on the

Origin	*# Pieces*
England	1013 (29.2%)
France	404 (11.6%)
Ireland	824 (23.8%)
Scotland	482 (13.9%)
S.E. Europe	127 (3.7%)
Scandinavia	620 (17.9%)
Total	3470

Table 1. The *Europa-6* collection: the number of pieces in each region.

other hand accuracy, since in a greater search space the chance to find better solutions is increased. We performed different runs with beam width set to respectively 25, 50 and 100. We report the results with the maximal beam width of 100 in this paper to ensure the best results, although we did not find significant improvement on the rule quality comparing beam widths of 50 and 100. Because of the relatively small number of features and samples in our corpus performance was not an issue.

On initial experiments with the Orange toolkit, we found that rules involving numeric attributes dominated the results and were very hard to interpret. Therefore we performed some preprocessing on the data, by discretizing all non-nominal features to either H (High) or L (Low) if the value was higher/lower than the mean of this feature in the entire corpus. This discretization has two consequences: the rules are easily interpreted by musicologists and we avoid overfitting the training data. Furthermore, after exploration of the initial results we decided to only output rules of length 2, which gives us more comprehensible results.

2.2 Dataset and music representation

The dataset we will use for our experiments is the *Europa-6* collection containing folk tunes from 6 different European regions, as detailed in Table 1. 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 four well-known global feature sets. These are:

Alicante: 28 global features, proposed by Ponce de Léon and Iñesta, applied to classification of 110 MIDI tunes in jazz/classical/pop genres [7]. From this set, we re-implemented a compact subset: the top 12 features chosen in [7]: Table 1;

Fantastic: 92 features computed by the program called Feature ANalysis Technology Accessing STatistics (In a Corpus), currently developed by Müllensiefen [8] (v0.9, downloaded on May 5 from [9]). For this study, we only include the global features based on a single melody, which reduces the set to 37 features;

Jesser: 39 pitch and duration statistics [10]. The pitch-based features are simple relative interval counts, like "amajsecond" (ascending major second). Similar features are present for all ascending and descending intervals in the range of the octave;

McKay: 62 global features, developed for the classification of orchestrated MIDI files, i.e. with instrumentation and dynamics [11]. Importantly, a superset of these features were used in the winning 2005 MIREX symbolic genre classification experiment which used orchestrated files for evaluation. The features were computed with McKay's software package jSymbolic (version 12.2.0) from [12].

Combining these four sets gives us a total of 150 features with few semantic overlap. After preprocessing this matrix which was used as input for the Orange toolkit.

3 Results

In this section we discuss our results of running the CN2-SD algorithm on the *Europa-6* dataset. Based on the WRAcc quality rankings from Orange, we select the best rule for each class and list these 6 rules in Table 2. It has to be noted that this quality measure is very hard to interpret since it is a local measure, not a representative measure with respect to the whole dataset. Therefore we have presented more informative probability values for each rule. These are:

- $p(A)$ the probability of the first rule component in the corpus.
- $p(A|\mathcal{C})$ the probability of the first rule component given class $\mathcal{C}$.
- $p(B)$ the probability of the second rule component in the corpus.
- $p(B|\mathcal{C})$ the probability of the second rule component given class $\mathcal{C}$.
- $p(A, B)$ the probability of the condition ($Cond$).
- $p(A, B|\mathcal{C})$ the probability of the condition given class $\mathcal{C}$.
- $p(\mathcal{C}|A, B)$ the probability of $\mathcal{C}$ given the condition.

Scandinavia: This rule defines a subgroup of pieces, all in 3/4 meter and containing a relatively low number of melodic tritones. This rule has a confidence of 0.78, meaning that the class is Scandinavia in 78% of the pieces that fulfill the condition. 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).

Ireland: This rule, found in 62% of Ireland tunes, refers to pieces with a relatively low number of dotted rhythms, and in a compound meter, meaning the numerator of the time signature is 6, 9, or 12, etc. Unlike the Scandinavia rule, the addition of the second condition B of the rule substantially increases the rule's specificity to Ireland.

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

Table 2. Best rule for each class. Each rule is a conjunction of two attribute-value pairs. We have used the $A\_$, $F\_$, $J\_$ or $M\_$ prefixes to represent the Alicante, Fantastic, Jesser or McKay features respectively. Two features are shown as an abbreviation for clarity: AvgVarIOI stands for AverageVariabilityofTimeBetweenAttacksForEachVoice and CompoundMeter corresponds to CompoundOrSimpleMeter.

Scotland: In 62% of the Scotland section are pieces in 4/4 meter, having a large standard deviation of the interpolation contour as defined in [8]. The latter means that the melodic contour over the span of a piece varies a lot.

S.E. Europe: With confidence 0.96 in the corpus, a piece with meter 7/8 and high variation in inter-onset intervals is a S.E. European folk tune. The unusual meter in pieces, using a diversity of rhythms, is predictive of the S.E. Europe class.

France: A relatively low number of descending minor thirds, combined with a low melodic range, defines this subgroup. The majority (77%) of the France pieces have this pattern. Interestingly, all of the France pieces from which we know that they have lyrics, are captured by this rule. This might indicate that this subgroup contains sung pieces, probably because of the low melodic range rule component, a voice having a smaller tessiture than for example a violin or a flute.

England: This subgroup found in England pieces is defined by pieces in major mode with a low density of notes, i.e., many long note durations. The mode was computed by estimating the probability of every major and minor key and picking the most probable key.

The only musical property that appears more than once in these rules is the meter, which therefore seems to be a good feature to characterize the subgroups in folk music of these 6 regions.

4 Conclusions and Future Work

We have presented an initial study on the discovery of descriptive rules for musical subgroups in which the CN2-SD algorithm was applied on the *Europa-6*

dataset, modeled by 150 global features coming from different studies reported in the literature. Discretization of the non-nominal features was necessary to avoid overfitting and improved the interpretability of the rules. For the same reason we only generated rules of length 2, even though allowing longer rules can improve their quality. The proposed approach only generated a limited number of rules with decent quality but the results contain some musically relevant information.

As future research, we would like to try this approach on a dataset with better annotation, so that we can automatically validate the discovered subgroups. An interesting alternative approach to SD that we want to explore is to look at in-class subgroups, which means that we would focus on one region at the time and try to identify and describe subgenres using only data from that region.

References

1. Klosgen, W.: Explora: A multipattern and multistrategy discovery assistant. Advances in Knowledge Discovery and Data Mining (1996) 249–271
2. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Proceedings of the First European Conference on Principles of Data Mining and Knowledge Discovery, MIT Press (1997) 78–87
3. Hillewaere, R., Manderick, B., Conklin, D.: Global feature versus event models for folk song classification. In: ISMIR 2009: 10th International Society for Music Information Retrieval Conference, Kobe, Japan (2009)
4. Lavrac, N., Kavsek, B., Flach, P., Todorovski, L.: Subgroup discovery with CN2-SD. Journal of Machine Learning Research (2004) 153–188
5. Demsar, J., Zupan, B., Leban, G.: Orange: From experimental machine learning to interactive data mining. White paper, Faculty of Computer and Information Science, University of Ljubljana (2004)
6. Clark, P., Niblett, T.: The CN2 induction algorithm. Machine Learning Journal **3**(4) (1989) 261–283
7. Ponce de Léon, P.J., Iñesta, J.M.: Statistical description models for melody analysis and characterization. In: Proceedings of the 2004 International Computer Music Conference. (2004) 149–156
8. Müllensiefen, D.: Fantastic: Feature analysis technology accessing statistics (in a corpus): Technical report v0.9. Technical report (2009)
9. `http://doc.gold.ac.uk/isms/mmm`
10. Jesser, B.: Interaktive Melodieanalyse. Peter Lang, Bern (1991)
11. McKay, C., Fujinaga, I.: Automatic genre classification using large high-level musical feature sets. In: Proceedings of the International Conference on Music Information Retrieval, Barcelona, Spain (2004) 525–530
12. `http://jmir.sourceforge.net/jSymbolic.html`