Alignment Methods for Folk Tune Classification

Ruben Hillewaere, Bernard Manderick, and Darrell Conklin

Abstract This paper studies the performance of alignment methods for folk music classification. An edit distance approach is applied to three datasets with different associated classification tasks (tune family, geographic region, and dance type), and compared with a baseline n-gram classifier. Experimental results show that the edit distance performs well for the specific task of tune family classification, yielding similar results to an n-gram model with a pitch interval representation. However, for more general classification tasks, where tunes within the same class are heterogeneous, the n-gram model is recommended.

1 Introduction

With the growth of the Music Information Retrieval field and the expansion of data mining methods, folk music analysis has regained attention through the past decades. Folk music archives represent a cultural heritage, therefore they need to be categorized and structured to be more easily consulted and searched. The retrieval of similar tunes from a folk tune database has been the subject of several MIREX contests, and alignment methods have proven to be the most successful at this task (Urbano et al. 2011). Various melodic similarity measures have been investigated

R. Hillewaere (✉) · B. Manderick
Artificial Intelligence Lab, Department of Computing, Vrije Universiteit Brussel, Brussels, Belgium
e-mail: rhillewa@vub.ac.be; bmanderi@vub.ac.be

D. Conklin
Department of Computer Science and AI, Universidad del País Vasco UPV/EHU, San Sebastián, Spain

Ikerbasque, Basque Foundation for Science, Bilbao, Spain
e-mail: conklin@ikerbasque.org

M. Spiliopoulou et al. (eds.), *Data Analysis, Machine Learning and Knowledge Discovery*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-01595-8_40,

for the exploration of a folk song database, and they have been combined in the attempt to find an optimal measure (Müllensiefen and Frieler 2004).

Music classification has become a broad subfield of the computational music research area, with many challenges and possible approaches (Weihs et al. 2007). In a recent study, alignment methods have been applied to the specific task of *tune family classification* (van Kranenburg et al. 2013), a tune family being an ensemble of folk songs which are variations of the same ancestral tune. In that work, alignment methods with various features were compared with several global feature models, in which a melody is represented as a vector of global feature values. It was shown that alignment methods achieve remarkable accuracies for tune family classification in comparison with the global feature models, regardless which features were used to represent the data. An open question, however, is how alignment methods perform on other types of folk tune classification tasks where tunes within a class do not present detectable melodic similarity.

The n-gram model is another machine learning technique that can be applied to both music classification (Hillewaere et al. 2009) and music retrieval (Uitdenbogerd and Zobel 1999). In this study, we investigate the performance of a simple alignment method, the edit distance, versus an n-gram classifier for the following tasks:

(a) tune family classification, in order to verify that the edit distance achieves similar results to the alignment methods reported by van Kranenburg et al. (2013);
(b) two fundamentally different folk music classification tasks. The first task is *geographic region classification*, which we have thoroughly studied in our previous work (Hillewaere et al. 2009). The second task is folk tune *genre classification*, where the genres are the dance types of the tunes (Hillewaere et al. 2012).

Given the excellent results with alignment methods in the study by van Kranenburg et al. (2013), they might also perform well on the different classification tasks proposed in (b). However, we do not expect this to happen and hypothesize the high performance is due to high similarity within tune families and that n-gram models over the same representations will perform equally well.

Since folk music is orally transmitted, traditionally by people singing during their social activities or work, over time multiple variations arise in the tunes. This phenomenon has led to the notion of tune family, i.e. an ensemble of tunes that all derive from the same ancestral tune. This is a hypothetical concept, since we generally cannot trace the historical evolution of a folk song. Given this definition of a tune family, it is obvious that songs of the same tune family are very similar, although numerous musical variations between them are possible, we only cite a few (van Kranenburg et al. 2007): melodic contour, rhythmic changes, insertion and deletion of parts, range, and number of phrases. We illustrate this with a score example in Fig. 1, which shows the first phrases of three tunes belonging to the tune family called “Heer”. Clearly, the first two phrases are highly similar, and the third phrase repeats the same melodic motif. This is typical for the tune family concept,

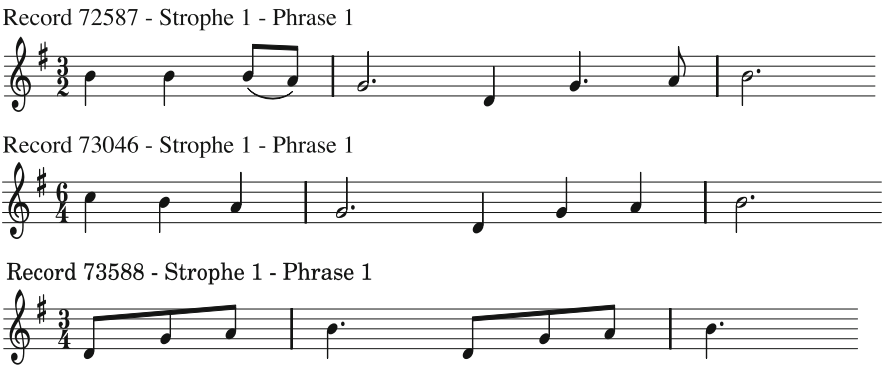


Fig. 1 Three tunes from the same tune family are very similar

and it is evident that tunes from the same geographic region or with the same dance type generally differ a lot more, which makes these classification tasks harder.

To verify our hypothesis, an edit distance method is applied to three folk music datasets with three different classification tasks, which will be described in the next section. For each of the folk tune collections, the pieces are encoded in melodic and rhythmic representations: as strings of pitch intervals, and as strings of duration ratios. These basic representations have been chosen to compare the predictive power of models based on melodic information versus rhythmic information. For each data collection and for each representation, pairwise edit distances are computed and the classification is done with a one nearest neighbour algorithm, which is similar to the approach used by van Kranenburg et al. (2013). A tenfold cross validation scheme is used to assess the performances in terms of classification accuracies.

2 Data Sets

In our experiments we use three folk tune datasets in MIDI format with different associated classification tasks, which we detail in this section.

2.1 TuneFam-26

This dataset of 360 songs is the tune family dataset used in the study of van Kranenburg et al. (2013). The source of this dataset is a larger collection called "Onder de groene linde", which is hosted at the Meertens Institute in Amsterdam. It contains over 7,000 audio recordings of folk songs that were tape-recorded all over

the country. The Witchcraft project digitized over 6,000 songs, both transcriptions of those audio recordings and from written sources.

In this large digitized database there are over 2,000 tune families, and a part of the Witchcraft project is to develop methods to retrieve melodies belonging to the same tune families. Therefore, the 360 songs were selected as to be representative, and they were grouped into 26 tune families by domain experts (van Kranenburg et al. 2013). This dataset, called the Annotated Corpus, is what we refer to as TuneFam-26.[1]

2.2 *Europa-6*

This is a collection of folk music from six geographic regions of Europe (England, France, South Eastern Europe, Ireland, Scotland and Scandinavia), for which the classification task is to assign unseen folk songs to their region of origin. Li et al. (2006) studied this problem with factored language models, and they selected 3,724 pieces from a collection of 14,000 folk songs transcribed in the ABC format.

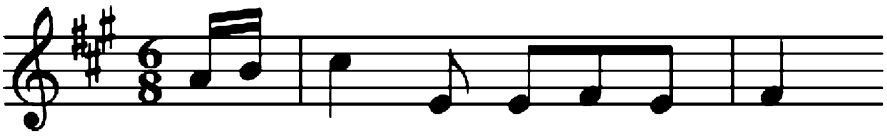
Their collection was pruned to 3,367 pieces by filtering out duplicate files, and by removing files where the region of origin was ambiguous. In order to end up with core melodies that fit for our research purpose, a preprocessing in two steps was carried out: the first step ensures that all pieces are purely monophonic by retaining only the highest note of double stops which occurred in some of the tunes, and in the second step we removed all performance information such as grace notes, trills, staccato, etc. Repeated sections and tempo indications were also ignored. Finally, by means of abc2midi we generated a clean quantized MIDI corpus, and removed all dynamic (velocity) indications generated by the style interpretation mechanism of abc2midi. In our previous work, we have shown that n-gram models outperform global feature models on this corpus (Hillewaere et al. 2009).

With a total of 3,367 pieces, Europa-6 is a larger dataset than TuneFam-26, and another contrast is that this dataset not only contains sung music, but also folk dances for example.

2.3 *Dance-9*

The corpus Dance-9 is a large subset of Europa-6: 2,198 folk tunes subdivided into nine dance type categories, the largest ones being jigs, reels and polskas. The associated classification task is to predict the dance type of an unseen tune. Several

[1]We would like to thank Peter van Kranenburg for sharing the Annotated Corpus and for the kind correspondence.



pitch	69	71	73	64	64	66	64	66
melodic interval	⊥	+2	+2	-9	0	+2	-2	+2
duration ratio	⊥	1	4	0.5	1	1	1	2
melodic interval string	‘ffqjfdf’							
duration ratio string	‘ltilllm’							

Fig. 2 Excerpt of the Scottish jig “With a hundred pipers”, illustrating the event feature sequences and the string representation

string methods have been compared with global feature models and event models using this data set (Hillewaere et al. 2012).

To construct Dance-9 we only selected the ABC-files with an unambiguous dance type annotation. Furthermore, we discarded all dance types that occurred insufficiently by putting a minimal threshold of 50 tunes, because it would lead to a highly unbalanced dataset. To the remaining 2,198 pieces, the same preprocessing steps as for Europa-6 have been applied, as is for the conversion to MIDI.

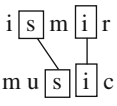
3 Methods

3.1 Music Representation

A piece of music can be viewed as an ordered sequence of events, and every event is represented by an event feature of one’s choice. In our case of monophonic folk music, the music events are note objects, with pitch and duration as basic event features. In this paper, we use two derived event features to represent the pieces: the first is the melodic interval in semitones between the current and the previous note, and the second is the duration ratio, i.e. the duration of the current note divided by the duration of the previous note. The obtained event features can be mapped onto characters, choosing a distinct character for each distinct feature value. Thereby the event feature sequence is mapped onto an ASCII symbol string, in which case we talk about a string representation. This is illustrated in Fig. 2 on a short excerpt of a jig.

3.2 Edit Distance

Alignment methods compute a distance measure between two sequences of symbols, by estimating the minimal cost it takes to transform one sequence into the other. In its simplest form this transformation is carried out by means of edit operations, such as substitution, insertion and deletion. Therefore, this method is often referred to as "edit distance", which is in fact the Levenshtein distance. For example, the edit distance between the strings "ismir" and "music" is equal to 4, since the optimal alignment between them is given by



which means four edit operations are needed: two substitutions ("i" to "m" and "r" to "c"), one insertion (the "u") and one deletion (the "m"). For the purpose of our current research, we have used WEKA's implementation of the edit distance (http://www.cs.waikato.ac.nz/ml/weka/).

The edit distance defines a pairwise distance metric, therefore the classification can be performed with an instance-based nearest neighbour classifier. Given a test instance, the prediction of its class label is solely based on the training instance which is closest with respect to the edit distance (instead of the usual Euclidean distance).

3.3 n-Gram Models

An n-gram model is a generative model for sequences which computes the probability of an entire sequence as the product of the probability of individual events within the sequence. Each event is conditioned on $n-1$ previous events and these conditional probabilities are estimated from a training corpus, with smoothing applied in order to avoid zero probabilities (Manning and Schutze 1999).

The n-gram model can be used for classification, by constructing a separate model for every class, and classifying a new sequence according to the model which generates the sequence with highest probability. To apply the model to music, every piece of the data set is transformed into an event feature sequence according to a feature of choice (e.g., duration ratio or melodic interval, see Sect. 3.1), and for each class the n-grams occurring in the class are compiled.

It is important to mention that the music representation is basically the same as for the edit distance approach, but the essential difference between these methods is that an n-gram model aims to model the transitions for a given class, whereas the edit distance computes a global pairwise similarity measure between pieces.

Table 1 The tenfold cross validation classification accuracies for our experiments on the three datasets

	Melodic int		Pitch	Duration ratio		Duration
	Alignment	Pentagram	Global	Alignment	Pentagram	Global
TuneFam-26	94.4 (3.9)	90.8 (3.7)	73.6 (8.9)	80.3 (5.9)	70.6 (5.9)	55.0 (7.5)
	92.0		*74.0*	*74.0*		*55.0*
Europa-6	49.5 (2.0)	64.1 (3.0)		47.5 (3.1)	55.1 (2.2)	
Dance-9	50.0 (2.6)	66.1 (2.2)		63.2 (1.4)	74.4 (2.0)	

Numbers in parentheses are the standard deviation over the ten folds. For comparison, the numbers italicized are the results by van Kranenburg et al. (2013)

4 Results and Discussion

In this section, the experimental results of the edit distance on the three datasets are reported for both the interval and the duration ratio representations. They are compared with a pentagram model (an n-gram model with $n = 5$), over the same representations. To assess the performance of both methods, we have set up a tenfold cross validation scheme to compute classification accuracies. The folds were taken in a stratified way, which is especially important for the results with TuneFam-26, to avoid that an entire tune family would be contained in the test fold, in which case a correct prediction would be impossible. We also ensured that the exact same folds were used for all experiments to do an impartial comparison.

The classification accuracies are reported in Table 1. The column on the left shows the melodic interval results, and the right column contains the duration ratio performances. The edit distance results are reported in the alignment columns, and for comparison we also include the results reported by van Kranenburg et al. (2013) (italicized numbers).

First of all, we observe higher accuracies on TuneFam-26 than on the other corpora. The edit distance approach classifies the tune family dataset with a high accuracy of 94.4 % on pitch intervals, which is very similar to the 92 % reported by van Kranenburg et al. (2013). This is remarkable since the edit distance is a simpler method than that used by van Kranenburg et al. (2013), which uses gap opening and extension weights in the computation. The edit distance slightly outperforms the pentagram model that still achieves an accuracy of 90.8 %, in other words there are only 13 more misclassified pieces.

With the duration ratios, the edit distance performs again very well on the tune family dataset with an accuracy of 80.3 %, outperforming both the pentagram model and the alignment method on duration ratio reported by van Kranenburg et al. (2013), though the high standard deviation of the accuracy estimate on both approaches should be noted (Table 1).

For the classification of geographic region or genre, the pentagram models clearly yield higher accuracies than the edit distance, with approximately 15 % difference for both datasets with the melodic interval representation. We remind the reader that 1 % on Europa-6 or Dance-9 corresponds to a larger amount of pieces due to the

difference in the sizes of the data sets, so this result shows that alignment methods are suitable for the specific task of tune family classification, but obtain much lower accuracies on more general types of classification tasks.

To summarize, all these results indicate that the tune family classification task is relatively easy. This finding specifically contradicts the statement of van Kranenburg et al. (2013) that the tune family classification task is more difficult than the region classification on Europa-6. They suggest that the heterogeneity of tunes between regions makes the task easier, but it appears in our results this is not the case. On the contrary, there is more heterogeneity within one region than there is in one tune family, which makes the region classification significantly harder.

We have also constructed two global feature models on TuneFam-26, based on global features derived from pitch on the one hand and duration on the other hand, similarly as in our previous work (Hillewaere et al. 2012). The accuracies obtained with an SVM classifier (with parameters tuned by a grid search) are reported in the respective columns of Table 1, and compared with the global feature results found by van Kranenburg et al. (2013). These accuracies confirm their statement that global feature approaches are of limited use for tune family classification.

5 Conclusions

In this paper we have investigated how a simple alignment method, called the edit distance, performs on three different folk music classification tasks: (a) classification of tune families, (b) classification of geographic region of origin, and (c) classification of dance types. Three folk tune datasets are used to assess the performance of the edit distance method in comparison with a pentagram model. Experimental results have shown the following:

- the edit distance approach performs well on the tune family dataset, yielding similar results to those reported by van Kranenburg et al. (2013);
- for edit distance, the tune family classification task is easier than classification of geographic region or dance type;
- for geographic region or dance type classification, an n-gram model is more appropriate.

We believe that these findings are due to the intrinsic concept of a tune family, since highly similar tunes are present within any tune family. Music retrieval methods using local sequential information, such as alignment methods and n-gram models, are capable of capturing this similarity and therefore lead to high performances. When pieces within classes are highly similar, alignment methods will achieve good classification results. On the other hand, when classes are more heterogenous the n-gram model is more appropriate.

References

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* (pp. 729–733). Kobe, Japan.

Hillewaere, R., Manderick, B., & Conklin, D. (2012). String methods for folk tune genre classification. In *Proceedings of the 13th International Society for Music Information Retrieval Conference* (pp. 217–222). Porto, Portugal.

Van Kranenburg, P., Garbers, J., Volk, A., Wiering, F., Grijp, L., & Veltkamp, R. (2007). Towards integration of MIR and folk song research. In *Proceedings of the 8th International Conference on Music Information Retrieval* (pp. 505–508). Vienna, Austria.

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.

Li, X., Ji, G., & Bilmes, J. (2006). A factored language model of quantized pitch and duration. In *International Computer Music Conference* (pp. 556–563). New Orleans, USA.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Müllensiefen, D., & Frieler, K. (2004). Optimizing measures of melodic similarity for the exploration of a large folk song database. In *Proceedings of the 5th International Conference on Music Information Retrieval*. Barcelona, Spain.

Uitdenbogerd, A. L., & Zobel, J. (1999). Matching techniques for large music databases. In *Proceedings of the ACM Multimedia Conference* (pp. 57–66). Orlando, Florida.

Urbano, J., Lloréns, J., Morato, J., & Sánchez-Cuadrado, S. (2011). Melodic similarity through shape similarity. In S. Ystad, M. Aramaki, R. Kronland-Martinet, K. Jensen (Eds.), *Exploring music contents* (pp. 338–355). Berlin: Springer.

Weihs, C., Ligges, U., Mörchen, F., & Müllensiefen, D. (2007). Classification in music research. *Advances in Data Analysis and Classification, 1*(3), 255–291.