Conditional Restricted Boltzmann Machines for Mono/Polyphonic Composer Identification

Loeckx, Johan

Publication date:
2015

Conditional Restricted Boltzmann Machines for Mono/Polyphonic Composer Identification

Johan Loeckx[a]

[a] *Artificial Intelligence Lab, Vrije Universiteit Brussel*
Pleinlaan 2, 1050 Brussel, Belgium

Abstract

In this paper, the effectiveness of Conditional Restricted Boltzmann Machines (CRBMs) as universal feature extractors for classifying symbolic music is investigated. An average monophonic classification accuracy of 72% was achieved when discriminating between string quartets movements of Mozart and Haydn. When the decisions of individual monophonic parts were combined using a basic voting scheme, a polyphonic classification accuracy of 96% was achieved, a substantial improvement compared to the best state-of-the-art performance to date of 80%. It was observed that the classification rate depended heavily on the exact composition of training and test set: the classifier performance deteriorated when the time of composition increased between training and test set. This supports the observation that the "style" of a composer is not a fixed given but varies over time.

1 Introduction

Using machine learning techniques to classify symbolic music fragments is not at all recent [14]. Existing techniques include, amongst others, event feature models like n-grams [7], global feature models [11], string (compression) methods [4] and similarity functions [1]. Although existing methods attain impressive accuracy scores of up to 80% when trying to discriminate between string quartets of Mozart and Haydn, the features are hand-crafted / specifically designed for the composer identification task [6].

Deep learning methods, on the other hand, are believed to model high-level abstractions automatically by learning multiple levels of representations by means of unsupervised learning algorithms to form a hierarchy of concepts. These kinds of techniques have already proven very successful in various settings, most notably in recognizing written digits and for modelling time series [8]. From this perspective, the interest in these methods for music purposes is not illogical [10] and the approach suggested in this paper is therefore inspired upon the work done by Taylor & Hinton on the modelling of motion style [15]. Furthermore, using Conditional Restricted Boltzmann Machines (CRBMs) for music has proved its effectiveness in the past for modelling temporal dependencies in music [3] and in automated tagging of music [9].

In this paper, we will investigate whether Conditional RBMs make a good candidate for composer identification. We will compare performance in the monophonic and polyphonic case to discriminate between compositions of Haydn and Mozart. Research by Dor and Reich concluded that the discrimination between string quartets of these two classical composers turned out to be the most challenging [5] as their composing style is very homologous [6]. First, the methodology is introduced and the chosen architecture is laid out. Next, experiments are discussed. Special attention will be given to the impact of test/training division on classification performance. It turns out that the more training and test set are intertwined with respect to composition time, the better the prediction performance, which suggests that composition style tends to change over a composer's life. Lastly, conclusions are stated.

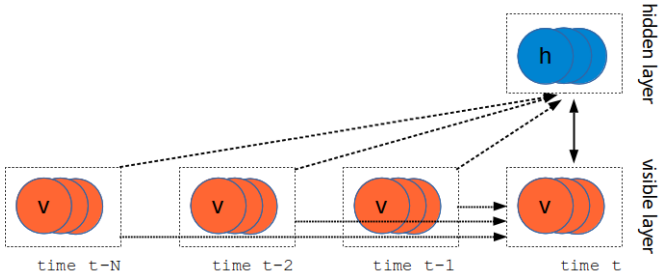

Figure 1: A Conditional Restricted Boltzmann Machine (CRBM) of order N. CRBMs attempt to capture temporal relations by introducing extra (unidirectional) dependencies between delayed inputs and the current input nodes, and from delayed inputs to the hidden layer.

2 Methodology

Deep learning is a class of machine learning training algorithms that attempt to learn higher order abstractions by training multiple layers of non-linear processing units. The difference with classical methods are the methods of learning (often a combination of unsupervised and supervised), the mechanisms of training and the structural limitations enforced (e.g. the number of hidden units, the nature of connections). In this section, we will investigate how a particular element, Conditional Restricted Boltzmann Machines (CRBMs) can be employed as a universal feature extractor that serves as input to a traditional classifier to perform symbolic composer identification between string quartets of Mozart and Haydn.

2.1 Neural Network Topology

A presumption often made when considering individual data points in a distribution of sequential data, is that nearby located points are also (statistically) closely related. One can take advantage of this property by describing the probability of an event occurring based on the events preceding it. It is this fact that has led to the conception of Conditional Restricted Boltzmann machines pictured in Fig. 1.

Conditional Restricted Boltzmann Machines are variants of the Restricted Boltzmann Machine (RBM) [13]. They introduce extra connections from past input samples to the current input neurons and hidden layer to model time-dependent effects, as shown in Fig. 1. They are part of the encoder/decoder paradigm, models that are able to encode their input data into features and decode those features back into the original representations. Above properties make that these architectures are often used to form (sparse) distributed representations of the input data that serve as feature extractors, employed in turn as input for more classical classifiers [2].

We will use a similar architecture, shown in Fig. 2. From raw MIDI data of the whole collection of Mozart and Haydn's string quartets, a piano roll representation is constructed and quantized at the level of 16th notes. Each quartet is played at 120 BPM, which means that a sixteenth note takes 125 ms to play. This representation is fed into a Conditional RBM pictured at the bottom of the figure. The sparse and distributed feature vector with 5000 binary elements, output of the CRBM, is used as input for a traditional pattern recognition neural network with 40 hidden nodes ($tanh$ activation) and two output nodes (softmax activation). The specific details of the model parameters are shown in Table 1. The dataset consisted of all Haydn and Mozart string quartets, obtained from the CCARH online database [12].

The goal here is not to test the efficiency of the MLP, but to assess the ability of the CRBM to automatically extract relevant features from a symbolic time-series in an unsupervised manner. In order to allow comparison with existing published methods, the same dataset has been used as in [7] and [6]. The fact that our method achieved an accuracy of 96% wrt. 80% for the best state-of-the-art method, obviates the need to explore other supervised classifiers.

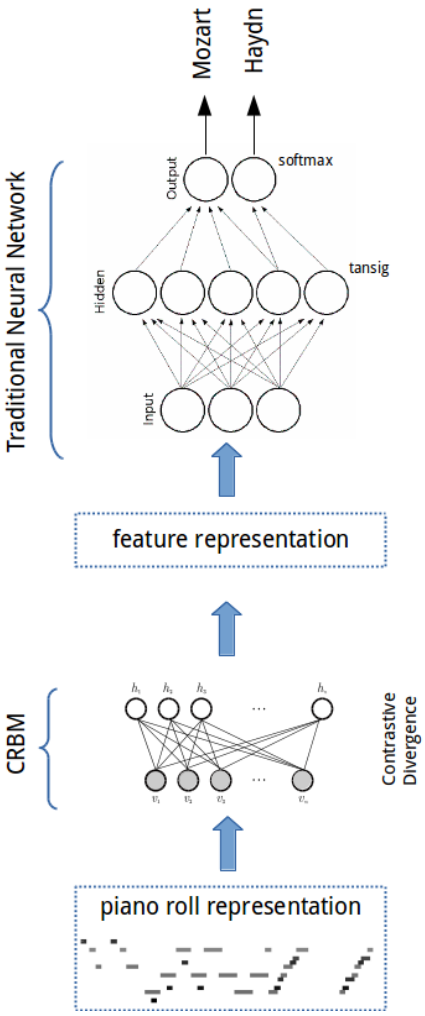

Figure 2: A Conditional RBM is trained in an unsupervised manner to learn a sparse and distributed binary feature representation, based on chunks of 10s time-variant piano roll input. The output of this network is fed into a multi-layered perceptron for pattern recognition that is trained in a traditional, supervised manner.

Parameter	Value
Type	Binary CRBM
Training algorithm	CD-1
# visible nodes	25 x 80 (delayed inputs)
# hidden nodes	5000
$\epsilon_{b_i}, \epsilon_{b_j}, \epsilon_w, \epsilon_A, \epsilon_B$	5.10^{-3}
w_{decay}	10^{-3}
momentum	0.9

Parameter	Value
Type	Multi-layered Perceptron
Topology (nodes)	5000–40–2
Training algorithm	scaled-conj. grad. back-propagation
regularization	0.1

Table 1: Model parameters and training details of the Conditional RBM and Multi-Layered Perceptron.

Part	Accuracy [%]
Violin I	70.2
Violin II	71.8
Viola	82.6
Cello	46

Table 2: Comparison of monophonic prediction accuracies for each part, using 5-fold cross-validation. The viola scored best, while the violoncello part scored as bad as a random predictor.

2.2 Implementation

The CRBM is trained in an unsupervised manner using a Contrastive Divergence algorithm, in an auto-encoding manner so that difference between the original visible layer input (the time-series vector) and the reconstruction of the visible layer (shaded neurons) by means of the hidden layer is minimized. The neural network is trained using scaled conjugate gradient backpropagation. Training of the traditional neural network succeeded in less than five minutes time, though the unsupervised training took 2 days on a 4-core 2.5Ghz 64-bit i7 Intel.

Each movement was split into 10s "phrases" that were used as input for the Conditional RBM. To create more training data for the unsupervised stage, a sliding window was used to create a 10s-phrase per quantization step rather than using non-overlapping phrases. In the classification phase, the movement to classify was again split into 10s phrases and classified for each (monophonic) phrase separately. A basic voting scheme was then used to calculate the movement prediction.

3 Experiments

3.1 Monophonic classification

In a first experiment, monophonic classification accuracy was assessed for each voice separately, using 5-fold cross-validation. This scheme was chosen as it is the same validation strategy as used in the best state-of-the-art solution to date [6], in order to allow a fair comparison. The results are shown in Table 2. The viola part scored best, with a 82.6% accuracy – already better than the best method known state-of-the-art polyphonic counterpart that scores 80%. Remarkably, the predictions based on the cello part scored particularly bad, with 46% being more or less equal to a random prediction. A possible explanation is that the cello is the lowest part (in pitch), and serves a supporting role (bass) that is rather style-specific (Mozart and Haydn are both classical composers) and less composer-specific. Interestingly enough, our finding contradict the findings of Hillewaere & Conklin[7], who found 3-gram models based on the violoncello part scored best when validated with a "leave-one-out" training/test strategy.

3.2 Polyphonic classification

In a second phase, the individual monophonic predictions of the viola and two violin parts were combined using a uniform weighing scheme to yield the polyphonic classification. As mentioned before, each movement was split into phrases of 10s and for each phrase independently, the composer was identified. This process was repeated for each voice (except the cello) and the total number of predictions summed. The improvement was impressive. Using a "leave-one-out" validation strategy (using all-but-one training samples, and one test sample), perfect prediction was attained, compared to 75.4% in [7]. When validated with 5-fold cross-validation, a more realistic strategy, still an accuracy of 96% was achieved – compared to 80% for the best state-of-the-art hand-crafted, feature based method [6]. Because of the very good accuracy, no other classifiers or topologies were investigated.

Method	Validation strategy	Accuracy
monophonic n-gram models [7]	leave-one-out	75.4%
monophonic CRBM-MLP	leave-one-out	**100 %**
feature-based* [6]	5-fold cross-validation	80 %
polyphonic CRBM-MLP	5-fold cross-validation	**96%**

Table 3: Comparison of classification accuracy with state-of-the-art methods. *To the best of the author's knowledge, this is the best method available. Our method significantly improves the current state-of-the-art.

3.3 Impact of training/test set composition

A surprising fact was that the average performance accuracy over all parts for monophonic composer identification varied considerably for different folds in the cross-validation[1]. Often there was one fold that scored 15–20% worse than the others. It seemed that the exact composition of movements in training and test set had severe impact on the prediction accuracy.

For this reason, we dug deeper into the relationship between training and test set performance based on different kinds of partitions. Our hypothesis was that the style of a composer changed during his life and that quartets written in the beginning of his life, were not representative for the style of later quartets. The dataset of quartets was thus sorted by composition date and two strategies for partitioning were investigated, with increasing levels of blending/mixing as shown in Fig. 3:

- **Scenario I (mixed movements)**
  Both training and test set consist of movements from the complete lifespan of the composers, but not all movements are considered.

- **Scenario II (partitioned movements)**
  Training occurs on movements that have been composed strictly before the set of test movements.

Scenario I and II can roughly be considered as the performance on future and contemporary movements respectively, where scenario I attempts to model the impact of incomplete sampling (similar to "album filtering" in audio-based tasks). Scenario II imposes a strict time-separation between training and test set to investigate whether early quartets are good predictors for later quartets. *"Spread"* refers to the number of movements between each training and test movement in scenario I. *"Distance"*, on the other hand, refers to the number of movements between the training and test partition that belong to neither test or training set in scenario II.

In the *first scenario (I)*, depicted in Fig. 4, prediction accuracy was regarded with respect to the "spread". Again, it shows that prediction accuracy was very high in case training and test set were "close". In case every training movement was preceded by a test movement (spread = 1), an accuracy of 94.7% was reached. However, when test performance was assessed on later compositions, accuracy dropped as the information obtained from earlier quartets did not apply to later ones.

Figure 5 depicts accuracy when there is a *strict partition between training and test data* (scenario II). In this case, all movements in the test dataset have been composed after the movements in the training dataset. When the distance is relatively close (less than 5 movements), prediction accuracy stays high (about 100% for movements). From then on, accuracy drops, with exception of distance 8 or 9. This is probably due to the fact that there is a quartet that is similar to an earlier one — the effect is very non-linear. For example, training classification for the second movement of Mozart's quintet KV 157 was misclassified as Haydn with 96.4% confidence, while the third movement was classified correctly as Mozart when quartet KV 80 and KV 155 were in the training set. No significant impact of the part on performance was observed, except for the violoncello part, which was considerably worse than all other parts. This may indicate that style is less clearly observed in the "supporting" parts.

[1] Please note that one fold existed of *a whole movement of a string quartet*. Training and test data thus never belonged to the same movement.

Figure 3: The prediction accuracy depends heavily on the amount of mixing between test and training set. The less they are intermingled, the less accurate predictions become. *Spread* refers to the number of movements between each training and test movement in scenario I whereas *distance*, refers to the number of movements between the training and test partition that belong to neither test or training set (scenario II.)

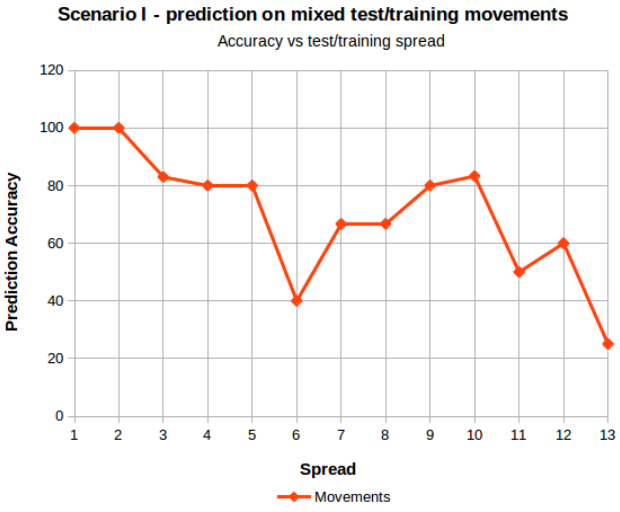

Figure 4: Prediction accuracy drops as the distance between test and training movements (”spread”), increases. This result seems to indicate that information from one movement cannot be generalized other, unless they are sufficiently ”close” (ie. composed at more or less the same date.)

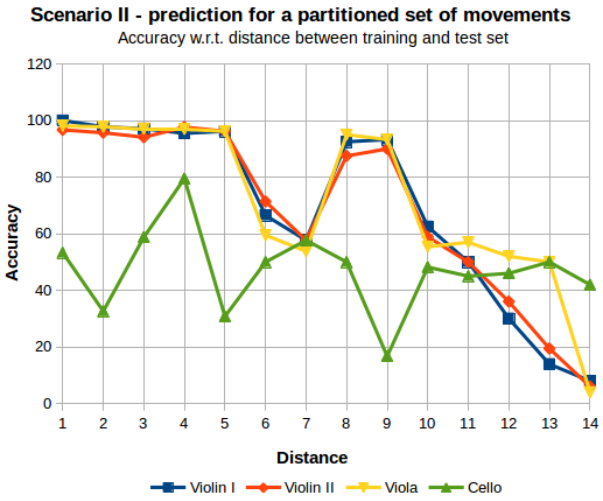

Figure 5: Classification accuracy vs. distance between a strictly partitioned set of training and test movements. As the distance increases, accuracy drops significantly. Accuracies for different monophonic parts is shown.

4 Conclusions and future work

This paper summarized the use of Conditional Restricted Boltzmann Machines (CRBM) as universal feature selectors for time-varying series, in the domain of symbolic composer identification. The proposed architecture of a CRBM followed by a multi-layered perceptron was very successful in classifying string quartets of Mozart vs. Haydn. The best monophonic classification accuracy achieved was 82% for the viola line. Combining monophonic classifications into a polyphonic predication improved the accuracy considerably upto 96%. Our method thus significantly improves the current state-of-the-art polyphonic performance of 80%, using universal feature extractors that need no manual intervention or crafting of features.

Furthermore, it was observed that the exact composition of training and test set had an important influence on the prediction accuracy. When the quartet movements used for training and test were well mixed (in time), a near-perfect movement classification was obtained, even in the monophonic case. However, when training was performed on earlier compositions to make predictions about later works (and vice-versa), the accuracy dropped significantly with increasing ”distance” in composing time.

5 Acknowledgments

This research has been supported by the EU FP7 PRAISE project #318770.

References

[1] Yoko Anan, Kohei Hatano, Hideo Bannai, Masayuki Takeda, and Ken Satoh. Polyphonic music classification on symbolic data using dissimilarity functions. In *ISMIR*, pages 229–234, 2012.

[2] Yoshua Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[3] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. *ICML 2012*, 2012.

[4] Rudi Cilibrasi, Paul Vitányi, and Ronald De Wolf. Algorithmic clustering of music based on string compression. *Computer Music Journal*, 28(4):49–67, 2004.

[5] Ofer Dor and Yoram Reich. An evaluation of musical score characteristics for automatic classification of composers. *Computer Music Journal*, 35(3):86–97, 2011.

[6] William Herlands, Ricky Der, Yoel Greenberg, and Simon Levin. A machine learning approach to musically meaningful homogeneous style classification. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[7] Ruben Hillewaere, Bernard Manderick, and Darrell Conklin. String quartet classification with monophonic models. In *ISMIR*, pages 537–542, 2010.

[8] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[9] Michael Mandel, Razvan Pascanu, Hugo Larochelle, and Yoshua Bengio. Autotagging music with conditional restricted boltzmann machines. *arXiv preprint arXiv:1103.2832*, 2011.

[10] Stephen McGregor, Geraint Wiggins, and Matthew Purver. Computational creativity: A philosophical approach, and an approach to philosophy. In *Proceedings of the Fifth International Conference on Computational Creativity ICCC-2014*, June 2014.

[11] Cory McKay and Ichiro Fujinaga. Automatic genre classification using large high-level musical feature sets. In *ISMIR*, volume 2004, pages 525–530, 2004.

[12] Craig Stuart Sapp. Online database of scores in the humdrum file format. In *ISMIR*, pages 664–665, 2005.

[13] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. 1986.

[14] Bob L Sturm. A survey of evaluation in music genre recognition. In *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation*, pages 29–66. Springer, 2014.

[15] Graham W Taylor and Geoffrey E Hinton. Factored conditional restricted boltzmann machines for modeling motion style. In *Proceedings of the 26th annual international conference on machine learning*, pages 1025–1032. ACM, 2009.