A Neural Probabilistic Model for Predicting Melodic Sequences

Srikanth Cherla, Artur d'Avila Garcez, and Tillman Weyde

School of Informatics
City University London
Northampton Square
London EC1V 0HB
{abfb145,a.garcez,t.e.weyde}@city.ac.uk

Abstract. We present an approach for modelling melodic sequences using Restricted Boltzmann Machines, with an application to folk melody classification. 1Results show that this model's predictive performance is slightly better in our experiment than that of previously evaluated n-gram models [**?**]. The model has a simple structure and in our evaluation it scaled linearly in the number of free parameters with length of the modelled context. A set of these models is used to classify 7 different styles of folk melodies with an accuracy of 61.74%.

Keywords: restricted Boltzmann machine, n-gram model, musical pitch, prediction, sequence classification, monophonic melody

1 Introduction

We are interested in modelling melodic sequences with machine learning. Musical pitch serves as a starting point for building models for more comprehensive analysis of sequential structure in music that include other musical features and polyphonic structures. The challenge here lies in effectively generalising over the large number of variations that are possible in sequences of individual musical features and in their interactions, given their limited number of occurrences in data. Although Markov models have been widely used for modelling sequences in music [**?**, **?**], they are often faced with a problem related to data sparsity, known as *curse of dimensionality*. This refers to the exponential rise in the number of model parameters with the length of the modelled sequences. Neural Networks are increasingly being considered as scalable alternatives to Markov models for sequence learning in both music and language [**?**, **?**, **?**].

In this paper, we present an approach using the Restricted Boltzmann Machine (RBM) [**?**] for learning sequences of musical pitch. We demonstrate that the predictive performance of this model improves with context length up to 7 notes and the number of model parameters increases linearly in the process. The model in our experiment performs slightly better than previously evaluated n-gram models on a dataset of J.S. Bach chorale melodies. As an example application, we use a set of these models to classify 7 classes of folk melodies with an accuracy of 61.74%.

2 Restricted Boltzmann Machine

The Restricted Boltzmann Machine (RBM) is an undirected graphical model consisting of a set of visible units $\mathbf{v}$ and a set of hidden units $\mathbf{h}$, with connections only between units belonging to different sets. In its original form, the RBM has binary, logistic units in both layers. It is an energy-based model in which the joint probability over the observed and latent variables is given by

$$p(\mathbf{v}, \mathbf{h}) = \frac{e^{-Energy(\mathbf{v},\mathbf{h})}}{Z} \tag{1}$$

where Z is a normalization factor determined by the *partition function* [**?**]. The energy function of the RBM is given by $Energy(\mathbf{v}, \mathbf{h}) = -\mathbf{b}^\top \mathbf{v} - \mathbf{c}^\top \mathbf{h} - \mathbf{h}^\top \mathbf{W} \mathbf{v}$, where $\mathbf{v}$ and $\mathbf{h}$ are activation vectors, and $\mathbf{b}$ and $\mathbf{c}$ are bias vectors for the visible and hidden units, respectively. The matrix $\mathbf{W}$ represents the connection weights between the hidden and visible units. Learning a sequence of symbols in the visible layer of the RBM amounts to maximizing the log-likelihood of the joint distribution $p(\mathbf{v})$. While computing the exact gradient of the log-likelihood function for $p(\mathbf{v})$ is not tractable, an approximation of this gradient called Contrastive Divergence (CD) gradient has been found to work well in practice [**?**].

3 The Prediction Model

We employ the RBM trained generatively with CD, as demonstrated in [**?**] to model the distribution $p(s_t | s_{(t-n+1)\ldots(t-1)})$, where s_t is the t^{th} element in a sequence $s_{1\ldots t}$ of MIDI pitch values. The visible layer of the network contains $(n-1)$ sets of binary one-of-$(m+1)$ softmax units (the context) and another similar set of m softmax units (the prediction), where n is the length of the learned subsequence and m the size of the alphabet. The additional unit in each set of context units handles the absence of a context at the start of a melody. The model is illustrated in Figure 1. The model is trained generatively using the first instantiation of Contrastive Divergence learning (CD_1) [**?**]. Given an incomplete sequence in the first $(n-1) \times (m+1)$ visible units, it predicts a probability distribution over the pitch values in the remaining m visible units [**?**].

4 Evaluation

We carried out a two-fold evaluation. In the first, we compared the average cross-entropies of the RBM models of different subsequence lengths on a dataset of 185 chorale melodies with those of corresponding n-gram models previously evaluated on the same dataset. The present model was compared with the Long-term Model evaluated there. The same cross-validation folds were used as in [**?**]. Table 1 shows that the predictive performance of the RBM model is slightly better than that of n-gram models[1] for most context lengths and when choosing

[1] The n-gram cross-entropies in Table 1 were provided by Dr. Marcus Pearce.

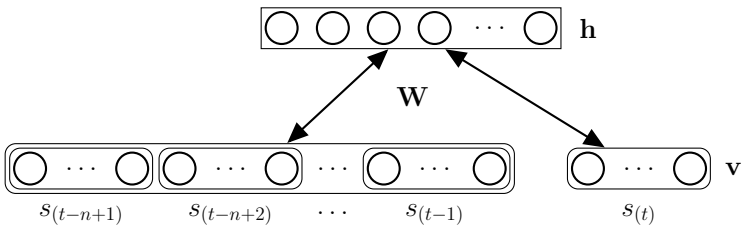

Fig. 1. The structure of the prediction model. Each set of nodes s_τ in the visible layer is the binary one-of-m representation of a pitch value. The sets grouped together to the left make up the context $s_{(t-n+1)\ldots(t-1)}$ of length $(n-1)$, and contain $(n-1)\times(m+1)$ nodes. The set of m nodes to the far right corresponds to the pitch $s_{(t)}$ to be predicted.

n	2	3	4	5	6	7	8	9	∞
n-gram	2.737	2.565	2.505	2.473	2.460	2.457	2.455	2.451	2.446
RBM	2.698	2.530	2.490	2.470	2.454	2.433	2.536	2.486	N/A
	(0.100)	(0.112)	(0.134)	(0.125)	(0.129)	(0.127)	(0.134)	(0.135)	N/A

Table 1. Comparison between n-gram and RBM models over a range of orders n, on the chorale melody dataset. The last column corresponds to unbounded order, which is not applicable to RBMs. The second and third rows are the means and corresponding standard deviations of cross-entropy across folds respectively for the RBM.

the best overall model. The performance progressively improves until a sequence length of 7. In a grid search over [100, 200, 400], 100 hidden units was found to be the best choice overall, leading to linear model growth in this evaluation.

In the second evaluation, a set of prediction models was employed as a one-vs-all classifier in the classification of 7 different folk-melody styles [**?**] from the Essen Folk Song Collection [**?**]. We refer the reader to [**?**] for details of the subset of this collection used here. With one model trained on pitch sequences from each class, a given test melody was assigned to the class whose model returned the lowest average cross-entropy value over that melody. Each of the models was trained on sequences of length 6. The extra visible unit for a missing context is not used here and the first 5 notes of a melody are ignored. The results are shown in Table 2. The overall classification accuracy is 61.74%. There is a relatively high degree of confusion between classes corresponding to geographically close regions of Europe (Alsace, Yugoslavia, Switzerland, Austria, Germany), suggesting the need to further optimize the models, or add more musical features.

5 Conclusions & Future Work

In this paper, we presented a neural probabilistic model for modelling fixed-length musical pitch sequences in monophonic melodies. It was demonstrated that the model performs slightly better than corresponding n-gram models in

	Nova-Scotia	Alsace	Yugoslavia	Switzerland	Austria	Germany	China	Total
Nova-Scotia	117	6	2	2	2	13	10	152
Alsace	8	33	11	7	15	15	2	91
Yugoslavia	15	14	54	9	17	7	3	119
Switzerland	6	9	10	33	22	11	2	93
Austria	5	16	10	14	41	14	4	104
Germany	14	23	10	15	14	132	5	213
China	11	3	2	2	5	1	213	237

Fig. 2. Confusion Matrix with results of the folk melody classification task.

most cases. Moreover, the results indicate linear growth with the sequence length n in the presented set-up. While this basic model has been successful at modelling musical pitch sequences, it is to be seen whether the introduction of other musical features like note durations, intervals, etc. would lead to improved predictions. The statistical significance of the improvements observed here is to be determined. Extension of the present model to polyphonic music, and an analysis of the features learned by the model in its hidden units [**?**], is also of interest.

Acknowledgements. The authors would like to thank Dr. Marcus Pearce for sharing his evaluation of the n-gram models and providing helpful feedback on this research, Dr. Darrell Conklin for his valuable advice and discovering an error in the evaluation, and Son Tran for many useful discussions on RBMs.