Improving Balance in Automatic Chord Recognition with Random Forests

Jeff Miller
Centre for Digital Music
Queen Mary University of London
UK

Ken O'Hanlon
Centre for Digital Music
Queen Mary University of London
UK

Mark B. Sandler
Centre for Digital Music
Queen Mary University of London
UK

***Abstract*—Automatic Chord Recognition (ACR) is a popular task to which Deep Learning (DL) has recently been successfully applied. ACR is considered as a classification problem wherein temporal frames of a piece of music are labelled according to some given chord vocabulary. When performing ACR using DL and larger chord vocabularies, an imbalance may be observed where popular chord classes are recognised more easily. In this paper, we propose random forest (RF) approaches in conjunction with existing DL strategies to mitigate such imbalance in ACR. We find improved balance in ACR is achieved when using DL-extracted chroma features alongside a RF balanced per-chord sampling strategy. Similar scores are achieved for accuracy and the balanced Average Chord Quality Accuracy (ACQA) metric on a Sevenths chord vocabulary, although the Sevenths accuracy is diminished relative to less balanced cases. Mapping the Sevenths estimations onto a MajMin vocabulary reveals little performance loss relative to initial MajMin estimations.**

I. Introduction

Automatic Chord Recognition (ACR) is a long-standing and well-studied topic [1]. Although other perspectives are possible [2], ACR was usually performed by classifying a feature, often chroma, as belonging to a certain chord label consisting of a root and chord quality [1]. The number of potential theoretical chord qualities is large, therefore chord vocabularies are employed to limit the potential complexity of the classification problem. ACR was first proposed in [3] where a large chord vocabulary was employed in experiments performed on synthetic, single-instrument data. However, ACR on real-world music was found to be difficult due to the presence of non-chord tones, percussion, and other artifacts in the signal. Hence, while some early work considered large vocabularies [4], smaller and simpler vocabularies were usually employed. The most well-known of these, referred to here as MajMin, contains only 25 chords: Major and Minor qualities in 12 transpositions and a 'no chord' class [5]. Such small chord vocabularies can appear to increase the accuracy of an ACR model but offer a limited perspective, particularly for certain styles of music.

Recently, Deep Learning (DL) has become the dominant methodology in ACR and has led to improved ACR performance. While small vocabularies are still often used [6] [7], several recent works have considered larger vocabularies in what seems to be a growing interest [8] [9] [10] [11] [12] [13]. There is a noted tendency of DL-ACR systems to perform better on popular chord qualities than on rarer qualities. This may be partially due to the uneven distribution of chords in available labelled datasets, used for training, that emphasise Western popular music. Furthermore, rarer complex chords are often extensions of more popular chords, from which they may be hard to distinguish. Recent methods have been proposed to tackle this problem, using compound classifiers with structured targets [10] [12], even chance training [9], and adaptive loss functions [13]. The classification imbalance leads to misleading results for larger chord vocabularies, with rarer chords often classified incorrectly. However, achieving balanced ACR remains a difficult task with accuracy falling when balance improves due to the previously skewed identification mode [12].

In this paper we propose an alternative approach to the imbalance problem in ACR. Rather than employ complex classifiers in DNNs, we simply employ DNN-extracted chroma features to which we apply a random forest chord classifier that utilizes a balanced per-chord sampling strategy.

The rationale for achieving chord balance with this approach is twofold. First, it is common for pitch-shift data augmentation to be applied in DL-ACR. This leads to balanced activity across all pitch class targets at training time, but this augmentation does not change the balance between different chord qualities. Second, the random forest consists of many decision trees, each of which can be trained in a less biased manner through the use of the balanced sampling strategy. We also provide a thorough ablation of this approach by comparing the performance of ACR systems utilising Constant-Q Transform (CQT) and chroma input features. For these features, we also propose a novel multi-median filtering approach to exploit temporal context as is performed in the convolutional networks that extract chroma features. The proposed approaches are described in Section 2. We then outline experiments comparing these approaches in Section 3, before concluding.

II. Methodology

Originally proposed by Ho [14], random forests are an ensemble learning classification method that employ several decision trees. While decision trees are effective classifiers

This research was supported by EPSRC grant EP/R512072/1 and the British Broadcasting Corporation through an Industrial CASE studentship in collaboration with the BBC Audio Research Partnership.

due to their execution speed, their potential complexity can be limited by a loss of generalisation accuracy when presented with previously unseen data. By applying the random subspace method to a large collection of decision trees, potential bias is distributed throughout the collection. A training dataset is uniformly sampled with replacement to generate an arbitrary number of smaller, incomplete datasets that are independent of each other. During training, a small subset of features is searched to find the optimal split at a given node in a decision tree, thereby reducing correlation between individual trees. The result is a large number of similar yet distinct decision tree models, thus reducing variability and overfitting. When a majority vote function is applied, the random forest acts as a more stable classifier than a single tree.

A. *Input features*

ACR is chiefly concerned with converting a spectral representation of a piece of music into chord labels. Most commonly the spectral representation is a semitone Constant-Q-Transform (CQT) [15] which is a log frequency spectrogram in which each frequency bin corresponds to a pitch on the equal temperament scale. It is possible to map a CQT, P, directly to chord label space L using some chord model $\mathcal{M}$:

$$\mathcal{M}(P) \longrightarrow L. \tag{1}$$

A direct mapping from CQT to a chord label space was rarely performed before DL-ACR. An intermediate summary chroma feature, C, was usually extracted from the CQT using the pitch folding operation, $\mathcal{F}(P) \longrightarrow C$, which sums energy for each pitch class across all octaves of the CQT. A chroma-based chord model is then applied, such that ACR is expressed as

$$\mathcal{M}(\mathcal{F}(P)) \longrightarrow L. \tag{2}$$

The chroma feature is a 12D vector, with each coefficient representing the presence or absence of its associated pitch class set in the original signal. In this way, chroma is a succinct feature in which each dimension is semantically meaningful, which is most apt for analysis of symbolic domain sources. However, in the case of audio, transformation of spectral information into a chroma vector may result in information loss with obfuscation of important acoustic relationships within the original signal. On the other hand, use of the unfolded CQT directly as an input feature may maintain structural relationships between the constituent frequencies of a note and reduce obfuscation due to harmonic smearing, in which the harmonic overtones of one note either reinforce or conflict with the harmonic frequencies of another note. The effect of harmonic smearing is to overemphasise some pitch classes for certain types of chords, while introducing conflicting noise to others. Training a model on an unfolded full spectrum CQT retains the frequency relationships necessary to identify these situations, and may thus increase tolerance to clashes in higher harmonic bins, resulting in improved ACR.

Recent DL-ACR methods tend to employ CQT inputs, often employing a network $\mathcal{N}$ to map these directly to chord labels: $\mathcal{N}(P) \longrightarrow L$. However, other approaches [16] [6] learn a feature from the network: $\mathcal{N}'(P) \longrightarrow C$, similar to pitch folding and to which a separate model can be applied:

$$\mathcal{M}(\mathcal{N}'(P)) \longrightarrow L. \tag{3}$$

We compare three input features to the random forests corresponding to the cases (1)(2)(3). In each case, we employ the same base CQT: the semitone RA-CQT feature that is formed using spectral reassignment, as defined in [17]. This is used directly as an input feature to the random forests (1), as the source of a chroma feature under pitch folding (2) [17], and as the input to a chroma-extracting Deep Neural Network (DNN) (3). In this third case, we employ the convolutional FifthNet [18]. By comparing these features, we examine whether the pitch folding process may result in a loss of information that may be detrimental to ACR, and test whether the DNN is successful in extracting the most relevant information from the CQT. It is noted that the DNN employs temporal context in extracting a single chroma feature.

B. *Weighting vs. Sampling*

The problem with chord imbalance has been noted [1] [12]. To address imbalance in random forests and many machine learning methods, a weighting scheme is often introduced that places higher weights on rarer chord classes to boost their contribution to training the model. A common weighting scheme considers the number of samples of a given class in a dataset, and uses the inverse of this number as a relative weighting for that class. We employ this weighting approach as a baseline for comparison.

Alternatively, we drop the random sampling associated with random forests and propose a even distribution sampling whereby the same number of samples from each class are used to train each tree. Such samples are still drawn randomly with replacement for each class. In this way, a richer model of each rarer chord class is built at each tree than might be the case where only a very few samples might represent a class.

C. *Temporal Context*

One advantage of using deep learning for ACR appears to be the use of methods that exploit temporal context. Convolutional Neural Networks (CNNs) consider blocks of input time frames to classify one output frame [7] while recurrent networks [11] and transformers [19] consider longer temporal patterns. In the context of random forests, we consider that a multi-frame block, similar to those often employed with CNNs, is unsuitable due to the very large dimensions involved. Hence we propose an approach using multiple median filters that present features that are smoothed at different time scales. Specifically, we employ four median filters, each applied in the pitch, or pitch class, dimension only. The filter dimensions were set to $\{3, 7, 11, 15\}$ as 15 frames has been previously been considered good in DL-ACR [6] [20]. Each filter produces an extra feature of the same dimension as the original feature $X \in \mathbb{R}^{S \times T}$. The extra features generated through filtering are simply stacked on top of the original feature in the pitch dimension, resulting in the filtered feature $X_m \in \mathbb{R}^{(5\times S)\times T}$.

While DL-ACR methods generally exploit temporal context, it is still common to employ a sequential classifier such as a Hidden Markov Model (HMM) [16] or Conditional Random Field [7] [11] as a post-processing step to smooth noisy labels. Here, a HMM is employed that uses a simple transition matrix with homogeneous off-diagonal entries equal to α, a user tuned parameter [21] [17].

III. Experiments

We employed the dataset from [20] [17], which comprises well-known datasets including the Beatles [22], Queen and Zweieck subsets of Isophonics [23], RWC-POP [24], US-POP [25], and Robbie Williams [26]. This compound dataset consists of 598 songs of approximately 40 hours duration. For training, the data was split six ways, using the same splits as in [18]. Cross-validation was then employed, with 4 splits used for training, one for validation and one for test in each run.

Two chord vocabularies were used for classification. The first, MajMin [27], uses Major and Minor qualities at 12 root positions and one No-Chord (NC) class, leading to a total of 25 classes. Other chord types (including seventh chords) are mapped into one of these classes as is commonly performed [27]. The second vocabulary, Sevenths (also referred to as 7s) extends the MajMin vocabulary with the addition of dominant7, major7, and minor7 chord qualities. Again, chord types with further extensions (in this case, 9ths, 13ths, etc.) are mapped onto these qualities [27]. The distribution of chord qualities in the Sevenths vocabulary for the dataset employed is given in Table 1.

Chord classification was performed on a framewise basis using a frame size of 100ms. Beat synchronisation was not applied. Given the chord notations and durations, each frame was assigned a label according to the chord active for the majority of the frame duration. The simplest metric is Accuracy, which relates the percentage of correctly estimated frames:

$$Acc = \frac{Number\ of\ correctly\ estimated\ frames}{Total\ number\ of\ frames} \times 100\%$$

To measure the balance in the chord qualities the Average Chord Quality Accuracy (ACQA) is employed. First, accuracy is calculated separately for each quality: $Acc(q)$, and the mean is taken to define ACQA :

$$ACQA = \frac{\sum_{q\in Q} Acc(q)}{|Q|}$$

We consider Q to consist of six qualities; the five chord qualities in the Sevenths vocabulary and the NC class. In addition, to provide an auxiliary metric for comparison to classifications made using the MajMin vocabulary, we mapped the Sevenths chord estimations onto the categories of the MajMin vocabulary. Major7 & Dominant7 chord types were mapped to the 'Major' class, and the Minor7 chord type was mapped to the 'Minor' class. To distinguish this mapping from the MinMaj vocabulary, we refer to this mapping as MM.

Quality	Maj	Min	Dom7	Maj7	Min7	NC
%	61.4	16.8	6.9	3.3	7.4	4.3

TABLE I
Distribution of chord qualities from Sevenths vocabulary in the dataset employed.

Experiments were run comparing the CQT and both standard and DNN-extracted chroma features on the Sevenths vocabulary. Each feature was tested using framewise and multi-median filter inputs. In all these cases we compared classwise weighting of random samples with the balanced chord class distribution sampling. After initial experiments, we employed tree depths of 12 for unfiltered chroma, and 36 for all other features, as we found no improvement beyond these sizes. Each forest consisted of 1000 trees, each of which was trained using 500 samples per chord class. We also trained using the MajMin vocabulary with similar hyperparameters. Results were recorded for both direct framewise classification, and classification with HMM post-processing.

A. Results

The results are given in Table 2 where it is seen that performance using CQT is most strongly related to use of the balanced datasets. Without such balance the forest does not perform well in terms of ACQA, regardless of how the temporal aspect is considered. With balanced data, the CQT feature performs better when the data is filtered rather than unfiltered; however, when the HMM is also applied the pattern is reversed, with small improvements for the filtered data and big improvements for the unfiltered balanced data.

The chroma feature performs worse than CQT over all metrics. Similar to CQT, balanced sampling results in better performance in terms of ACQA and MM, while the HMM post-processing affects the results for filtered data by relatively small amounts and poor performance is seen for filtered, unbalanced data. However, unlike the CQT, when the HMM is applied to labels extracted from unfiltered, unbalanced chroma, performance metrics approach those found using the balanced options.

Finally, the DNN-extracted chroma performs better than both other features, with improvements of over 6% in terms of the ACQA metric, regardless of use of the HMM. A similar pattern to the chroma feature is seen; performance is much better for unbalanced data when it is also unfiltered, while the best performance is seen using the balanced data. Filtering is seen to have a small negative effect on metrics apart from Accuracy, while it is noted that the DNN chroma feature is extracted using a convolutional network that considers temporal context.

The results for the forests trained on the MajMin vocabulary alone are also given in Table 2. These are most often comparable to the MM metric for the data trained with a Sevenths vocabulary. This is in contrast to Accuracy for the Sevenths vocabulary, which drops when the results are more balanced. This is interesting as it shows that the taper off in

			Framewise				HMM			
			Sevenths			MajMin	Sevenths			MajMin
Feature	Filtered	Balanced	*Acc*	ACQA	MM	Acc	*Acc*	ACQA	MM	Acc
CQT	×	×	56.8	30.8	66.0	66.0	63.5	33.9	74.1	74.6
	×	✓	51.8	45.6	65.2	64.9	*68.2*	*58.6*	*79.6*	*79.4*
	✓	×	*66.1*	40.6	77.0	*77.2*	66.7	40.6	77.8	78.3
	✓	✓	64.9	*55.3*	*77.5*	76.0	67.8	57.4	79.2	78.7
Chroma	×	×	40.3	35.2	58.3	61.0	62.8	55.5	76.9	*78.0*
	×	✓	38.7	35.7	58.6	59.1	60.2	*58.2*	*77.7*	77.9
	✓	×	*64.3*	39.6	*75.7*	*75.8*	*65.7*	39.8	77.2	77.4
	✓	✓	58.9	*51.2*	74.8	74.5	63.5	54.7	77.1	77.0
DNN Chroma	×	×	64.5	59.3	80.0	80.5	68.2	61.8	**81.4**	**81.7**
	×	✓	61.2	**61.4**	79.9	79.9	65.0	**65.0**	81.3	81.1
	✓	×	**68.6**	46.2	**80.4**	**80.6**	**69.7**	45.6	**81.4**	**81.7**
	✓	✓	61.0	60.4	79.3	79.6	65.5	64.2	81.0	81.0

TABLE II
Table showing ACR results, using CQT, chroma and DNN chroma features with options for filtering and balancing, both with and without HMM post-processing. Several metrics are given for Sevenths vocabulary. Accuracy values for MajMin vocabulary are included for comparison. Bold and italic fonts denote the best result for a given metric over all features, and for a given feature, respectively.

performance is actually slight, and it is suspected that the differences between the balanced and unbalanced Sevenths estimations lie chiefly between classes belonging to the same superclass e.g. when Minor7 and Minor are mistaken for each other. The best results for the MajMin vocabulary for the unbalanced DNN chroma feature are 81.7%, which was the same result for the same features when a Gaussian Mixture Model (GMM) classifier was applied in [18].

In general, we see that using balanced datasets tends to lead to better ACQA and MM metrics, although in some cases accuracy is sacrificed for the sake of balance. Filtering seems to have a positive effect on all results without the HMM being applied, although generally it is found that balanced data with HMM post-processing performs better when filtering is not employed.

It is notable that the MM metric is not affected much for either chroma feature, although a relatively large drop in MM was seen using the CQT feature without filtering or balancing. This drop is also seen for the data trained with MajMin labels. Apart from this unfiltered, unbalanced case the CQT does improve on the standard chroma feature. While one might assume that the extra information in the CQT relative to the chroma feature leads to CQT being superior for the task, this result suggests that it is more difficult to extract this information in a meaningful way.

On a similar note, given the improved performance seen for the DNN Chroma in general, we can assume that the neural network employed was capable of extracting relevant information from the CQT to produce a probabilistic chroma vector that is more apt for the task than the standard energy-based chroma feature. Equally, given the relatively strong performance of the DNN chroma feature without the HMM or filtering applied, this could be considered a result of temporal context processing in the CNN. However, it seems that there might be some room for improvement.

It is interesting that applying the HMM to the DNN chroma feature results in only a small increase in MM, while improving Acc and ACQA by larger amounts. This suggests that the decision boundaries between the different qualities in each of the Major / Minor superclasses in the Sevenths vocabulary could be sharper.

IV. Conclusions

We considered the use of random forests to address the chord imbalance problem in ACR, which we tested on a Sevenths vocabulary containing chords of varying cardinality. Particularly, we proposed an even chord class distribution training scheme for the forests, which was found to be successful for balancing the per-chord accuracy, while a DNN-extracted chroma feature was seen to be superior to other features. While the accuracy was seen to diminish using the Sevenths vocabulary, we note that the superclass accuracy, i.e. MajMin vocabulary, was not affected. Hence we consider that the approach is relatively effective in separating members of the same superclass, e.g. distinguishing between Minor7 and Minor. However, given the effect of HMM smoothing on the Acc and ACQA metrics we propose that further improvements might be made. While balancing chords seems a worthwhile endeavour, it is perhaps the case that the most balanced ACR system is not the optimal solution for every application. Even so, we consider that random forests might easily be tuned towards other considerations, simply by considering the training data balance. Future work will consider further chord qualities.

References

[1] J. Pauwels, K. O'Hanlon, E. Gomez, and M.B. Sandler, "20 years of automatic chord recognition from audio," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[2] K. Yoshii and M. Goto, "A vocabulary-free infinity-gram model for nonparametric Bayesian chord progression analysis," 2011, pp. 645–650.

[3] T. Fujishima, "Realtime chord recognition of musical sound: a system using Common Lisp Music," in *Proceedings of the International Computer Music Conference (ICMC)*, 1999, pp. 464–467.

[4] M. Mauch and S. Dixon, "Simultaneous estimation of chords and musical context from audio," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1280–1289, August 2010.

[5] C. Harte and M. B. Sandler, "Automatic chord identification using a quantised chromagram," Barcelona, Spain, May 28–31 2005.

[6] F. Korzeniowski and G. Widmer, "Feature learning for chord recognition: The deep chroma extractor," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2016.

[7] F Korzeniowski and G. Widmer, "A fully convolutional deep auditory model for musical chord recognition," in *Proceedings of the IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2016, pp. 1–6.

[8] J. Deng and Y.-K. Kwok, "A hybrid gaussian-hmm-deep learning approach for automatic chord estimation with very large vocabulary.," 2016, pp. 812–818.

[9] J. Deng and Y.-K. Kwok, "Large vocabulary automatic chord estimation using bidirectional long short-term memory recurrent neural network with even chance training," *Journal of New Music Research*, vol. 47, no. 1, pp. 53–67, 2018.

[10] B. McFee and J. P. Bello, "Structured training for large-vocabulary chord recognition," in *Proceedings of the International Society for Music Information Retrieval*, 2017.

[11] Y. Wu and W. Li, "Automatic audio chord recognition with midi-trained deep feature and blstm-crf sequence decoding model," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 27, no. 2, pp. 355–366, Feb. 2019.

[12] J. Jiang, K. Chen, W. Li, and G. Xia, "Large-vocabulary chord transcription via chord structure decomposition," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[13] M. Bortolozzo, R. Schramm, and C. R. Jung, "Improving the classification of rare chords with unlabeled data," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3390–3394.

[14] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995, vol. 1, pp. 278–282.

[15] J. C. Brown, "Calculation of a constant-q spectral transform," *Journal of Acoustic Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[16] E. J. Humphrey, T. Cho, and J. P. Bello, "Learning a robust Tonnetz-space transform for automatic chord recognition," in *Proceedings on the International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2012, pp. 453–456.

[17] K. O'Hanlon and M. B. Sandler, "Comparing cqt and reassignment based chroma features for template-based automatic chord recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, 2019.

[18] K. O'Hanlon and M. B. Sandler, "Fifthnet: Structured compact neural networks for automatic chord recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 2671–2682, 2021.

[19] T.-P. Chen and L. Su, "Harmony transformer: Incorporating chord segmentation into harmony recognition," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[20] K. O' Hanlon and M. B. Sandler, "The fifthnet chroma extractor," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[21] T. Cho and J. P. Bello, "On the relative importance of individual components of chord recognition systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 477–492, February 2014.

[22] C. Harte and M. Sandler, "Automatic chord identification using a quantised chromagram," in *Proceedings of the Audio Engineering Society Convention*, 2005.

[23] M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Kolozali, D. Tidhar, and M. Sandler, "Omras2 metadata project 2009," in *Late Breaking Demo, International Conference on Music Information Retrieval (ISMIR)*, 2009.

[24] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Popular, classical, and jazz music databases," in *The 3rd International Conference on Music Information Retrieval (ISMIR)*, 2002, pp. 287–288.

[25] T. Cho, *Improved techniques for automatic chord recognition from music audio signals*, Ph.D. thesis, New York University, 2013.

[26] B. di Giorgi, M. Zanoni, A. Sarti, and S. Tubaro, "Automatic chord recognition based on the probabilistic modeling of diatonic modal harmony," in *8th international workshop on multidimensional systems (nDS13)*, 2013.

[27] J. Pauwels and G. Peeters, "Evaluating automatically estimated chord sequences," in *Proceedings of the International Conference on Audio, Speech and Language Processing (ICASSP)*, 2013, pp. 749–753.