Modeling Rhythmic Attributes in Music At Scale with Tree Ensembles and the Music Genome Project

Matthew Prockup+* MPROCKUP@DREXEL.EDU
Andrew J. Asman* AASMAN@PANDORA.COM
Andreas F. Ehmann* AEHMANN@PANDORA.COM
Fabien Gouyon* FGOUYON@PANDORA.COM
Erik M. Schmidt* ESCHMIDT@PANDORA.COM
Youngmoo E. Kim+ YKIM@DREXEL.EDU

+Drexel University, ECE Dept., 3141 Chestnut St., Philadelphia, PA 19104
*Pandora Media, Inc., 2101 Webster St, Oakland, CA 94612

Abstract

Musical meter and attributes of the rhythmic feel (e.g., swing) are crucial when defining musical style, but they are often difficult to capture in music audio signals. In this work, we employ a set of tree ensemble models trained using a rhythm-inspired acoustic feature set. We model a total of nine rhythmic attributes covering meter and feel using over one million examples labeled by experts from Pandora® Internet Radio's *Music Genome Project®*. While linear models are shown to be somewhat effective, the complexities of rhythm are better represented using more powerful, non-linear, tree ensemble methods.

1. Introduction

Rhythm is one of the fundamental building blocks of music, and perhaps the simplest aspect for humans to identify with. Previous work has studied the general recognition of rhythmic styles in music audio signals (Gouyon & Dixon, 2004) but few efforts have focused on the deconstruction and quantification of the foundational components of global rhythmic structures. Furthermore, it has been shown that these components have very important relationships to definitions of style and musical genre (Prockup et al., 2015a).

The fundamental components of rhythm are metrical structure, tempo, and timing (Gouyon & Dixon, 2005). There is a large body of prior work that attempts to estimate these components (Ellis, 2007; Gouyon et al., 2006; Peeters & Papadopoulos, 2011), but in extracting only beats, tempo, or meter, much of the rhythmic subtlety and feel is discarded. A mid-level representation known as the *accent signal* (Böck & Widmer, 2013), which measures the general presence of musical events, is better suited to represent this rhythmic subtlety.

In previous work, we have shown that linear models trained with compact features derived from the accent signal are quite effective when representing rhythmic structures (Prockup et al., 2015b). In this work, we focus on modeling rhythm-related attributes of meter and feel (e.g., "swing") using tree ensembles, which can more powerfully describe the non-linear complexities these attributes may contain.

2. Rhythmic Attributes and Acoustic Features

The targeted attributes are compositional constructs such as the meter or well-defined components of the rhythmic feel. Namely we focus on the 9 rhythmic attributes shown in Table 1 (top), which are labeled by musical experts from Pandora® Internet Radio's *Music Genome Project®*(MGP)[1]. In order to capture aspects of each rhythm attribute, a set of rhythm-inspired features was implemented (bottom Table 1). Each relies on global estimates of an accent signal (Böck & Widmer, 2013). The accent signal can also be separated into to multiple versions that are each constrained to specific frequency sub-bands, allowing for rhythms with different compositional functions (e.g., bass, lead) to be captured separately (see (Prockup et al., 2015b) for more details).

[1] "Pandora" and "Music Genome Project" are registered trademarks of Pandora Media, Inc. http://www.pandora.com/about/mgp

Meter attributes denote musical meter distinct from simple duple, namely: cut-time, compound-duple, triple, odd.
Swing denotes longer durations on the beat followed by a shorter duration. It usually occurs on the 2^{nd} and 4^{th} beats.
Shuffle is similar to swing, but the motif is equally perceived on all beats.
Syncopation is confusion created by early anticipation of the beat or obscuring meter with emphasis against strong beats.
Back-Beat Strength is the emphasis placed on the 2^{nd} and 4^{th} beat or grouping in a measure or set of measures.
Danceability is the utility of a song for dancing. This relates to consistent rhythmic groupings with emphasis on the beats.
Beat Profile features are statistics of a quantized version of the accent signal between consecutive beat estimates.
Tempogram Ratio features are tempo-invariant fractional relationships of musical event timings. An estimate of tempo removes tempo-scaling in a tempogram.
Mellin Scale Transform is a naturally tempo-invariant transform of a signal (Holzapfel & Stylianou, 2011).
Mellin Periodicity emphasizes Mellin transform periodicities using the discrete cosine transform, median removal (subtracting the local median) and half-wave rectification.
Mel-Frequency Cepstral Coefficients (MFCC) capture instrument timbre. A block-based approach is used (Seyerlehner et al., 2011).

Table 1. Explanations musical attributes (top), and acoustic features (bottom).

3. Experiments: Tree Ensembles

In this work we employ both Random Forests (RF) (Breiman, 2001) and Gradient Boosted Trees (GBT) (Ye et al., 2009) formulated for both classification of binary attributes (e.g., meter) and regression of continuous attributes (e.g, danceability). Additionally, similar to work by (He et al., 2014), we can treat each tree ensemble as a feature transformation and use the output of each leaf as input features to a simpler classification or regression model (RF-H, GBT-H). The output decisions of each tree in the ensemble can be used as a new feature set that exploits the relationships of ensemble predictions. In this work, we use the leaf outputs of each tree as inputs into linear classifiers (logistic regression) and regressors (linear regression) trained using stochastic gradient decent (SGD).

For each method, a 70%:30% (train:test) split of the data was used, and no artists were shared between the training and testing sets. For comparison, we will evaluate these methods against the linear models presented in previous work (Prockup et al., 2015b). When training each model, the following tree parameters were tuned across generally accepted ranges: tree depth (3-8), number of estimators (50-250), percentage of features used per estimator (12.5%-50%). Only the best models will be evaluated.

4. Results

The results for each of the experiments are shown in Table 2. Each model was trained using only rhythm features, only timbre features, or their combination. It is seen across the board, and similar to previous work (Prockup et al., 2015b), that the rhythm features perform better than timbre features when modeling rhythmic attributes. The combination of rhythm and timbre performs only slightly better than when using rhythm features alone. Each of the tree models outperform each of the linear models, suggesting that the relationship of rhythm features to rhythm attributes are more complex than those captured by purely linear models. When considering the tree ensemble models, the GBTs and GBT-Hs generally outperform RFs and RF-Hs respectively. For GBTs. the hybrid approach (GBT-H) in this context is not very helpful. However, for RFs, the RF-H approaches are helpful, especially for regression of the continuous attributes (danceability, back-beat), with model performance approaching that of the GBT and GBT-H models.

		AUC		Comp.					**R^2**	Back-
Features	Model	Cut	Triple	Duple	Odd	Swing	Shuf.	Sync.	Dance	Beat
	Linear	0.797	0.794	0.663	0.745	0.781	0.719	0.705	0.400	0.309
	RF	0.842	0.811	0.702	0.791	0.830	0.758	0.738	0.515	0.340
Timbre	GB	**0.877**	0.808	0.689	0.769	0.828	0.761	0.737	**0.570**	**0.401**
	RF-H	0.853	0.817	0.707	**0.793**	0.835	0.762	0.750	0.560	0.373
	GBT-H	0.868	**0.820**	**0.713**	**0.793**	**0.839**	**0.764**	**0.759**	0.553	0.372
	Linear	0.901	0.924	0.946	0.859	0.903	0.919	0.768	0.505	0.316
	RF	0.926	0.938	0.960	0.870	0.916	0.926	0.779	0.554	0.364
Rhythm	GB	**0.944**	**0.956**	**0.962**	0.862	**0.932**	**0.928**	0.779	**0.615**	**0.463**
	RF-H	0.930	0.943	0.960	0.875	0.922	0.927	**0.787**	0.602	0.444
	GBT-H	0.939	0.951	0.961	**0.886**	0.925	**0.928**	0.786	0.603	0.449
	Linear	0.905	0.920	0.943	0.865	0.903	0.919	0.777	0.486	0.441
Timbre	RF	0.930	0.936	0.960	0.881	0.921	0.930	0.791	0.594	0.417
+	GB	**0.949**	**0.956**	0.960	0.877	**0.935**	0.930	0.793	**0.645**	**0.505**
Rhythm	RF-H	0.934	0.943	0.960	0.883	0.926	0.931	0.802	0.634	0.477
	GBT-H	0.946	0.953	**0.962**	**0.899**	0.931	**0.932**	**0.805**	0.631	0.487

Table 2. The rhythmic attribute learning is evaluated with area under the ROC curve for classification and R^2 for regression.

5. Conclusion

We found that tree ensembles are better than linear methods when modeling the complexities of rhythmic attributes. In most cases, Gradient Boosted Trees (GBT) perform best. For GBT, the addition of the hybrid approach (GBT-H) provides little gain. However, the hybrid approach for Random Forests (RF-H), was helpful when modeling continuous attributes, which is an intriguing result. Previous efforts in these hybrid approaches usually target binary classification (He et al., 2014), through which their interpretation is more intuitive. In future work, we will further explore why these methods also work well for regression, and try to better formulate its interpreted meaning.

References

Böck, Sebastian and Widmer, Gerhard. Maximum filter vibrato suppression for onset detection. In *Proc. of the 16th International Conference on Digital Audio Effects (DAFx-13)*, 2013.

Breiman, L. Random forests. *Machine learning*, 2001.

Ellis, Daniel PW. Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1):51–60, 2007.

Gouyon, F, Klapuri, A, Dixon, S, Alonso, M, Tzanetakis, G, Uhle, C, and Cano, P. An experimental comparison of audio tempo induction algorithms. *Audio, Speech, and Language Processing, IEEE Trans.*, 14(5):1832–1844, September 2006.

Gouyon, Fabien and Dixon, Simon. Dance music classification: A tempo-based approach. In *Proc. of the International Society for Music Information Retrieval Conference*, 2004.

Gouyon, Fabien and Dixon, Simon. A review of automatic rhythm description systems. *Computer music journal*, 2005.

He, X, Pan, J, Jin, O, Xu, T, and Liu, B. Practical Lessons from Predicting Clicks on Ads at Facebook. *ACM SIGKDD*, 2014.

Holzapfel, André and Stylianou, Yannis. Scale transform in rhythmic similarity of music. *IEEE Trans. on Audio, Speech and Language Processing*, 19(1):176–185, 2011.

Peeters, Geoffroy and Papadopoulos, Helene. Simultaneous Beat and Downbeat-Tracking Using a Probabilistic Framework: Theory and Large-Scale Evaluation. *Audio, Speech, and Language Processing, IEEE Trans.*, 19(6):1–17, August 2011.

Prockup, Matthew, Ehmann, Andreas F., Gouyon, Fabien, Schmidt, Erik M., Celma, Oscar, and Kim, Youngmoo E. Modeling genre with the music genome project: Comparing human-labeled attributes and audio features. In *Proc. of the International Society for Music Information Retrieval Conference*, 2015a.

Prockup, Matthew, Ehmann, Andreas F., Gouyon, Fabien, Schmidt, Erik M., and Kim, Youngmoo E. Modeling musical rhythm at scale using the music genome project. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2015b.

Seyerlehner, Klaus, Schedl, Markus, Knees, Peter, and Sonnleitner, Reinhard. A refined block-level feature set for classification, similarity and tag prediction. *Extended Abstract to MIREX*, 2011.

Ye, J, Chow, JH, Chen, J, and Zheng, Z. Stochastic gradient boosted distributed decision trees. *ACM Information and knowledge management*, 2009.