A Comparative Study on Music Genre Classification Algorithms

Wojciech Stokowiec

Abstract Music Genre Classification is one of the fundamental tasks in the field of Music Information Retrieval (MIR). In this paper the performance of various music genre classification algorithms including Random Forests, Multi-class Support Vector Machines and Deep Belief Networks is being compared. The study is based on the "Million Song Dataset" a freely-available collection of audio features and metadata. The emphasis is put not only on classification accuracy but also on robustness and scalability of algorithms.

Keywords Music genre recognition · Million Song Dataset · Machine learning

1 Introduction

Musical genres are categorical descriptions that are used to characterize music. Although classification criteria may seem subjective and arbitrary, humans have shown remarkable skill at genre recognition. As the number of songs, which are available to listeners, grows exponentially, it seems implausible that they would be equipped with appropriate textual information such as music genre. Music genre classification is considered to be a great practical component of music retrieval and recommendation systems, thus techniques for automatic genre classification would be a valuable tool.

This work describes systems of automatic music genre recognition based exclusively on audio features. The paper is structured as follows: firstly, a brief review of related work is presented in Sect. 2. Secondly, the dataset, on which this study has been based, is detailed in Sect. 3. Next, the algorithms that have been compared are described in Sect. 4 and results are given in Sect. 5. Finally, conclusions and future work can be found in Sect. 6.

W. Stokowiec (✉)
National Information Processing Institute, Warsaw, Poland
e-mail: wojciech.stokowiec@opi.org.pl

D. Ryżko et al. (eds.), *Machine Intelligence and Big Data in Industry*,
Studies in Big Data 19, DOI 10.1007/978-3-319-30315-4_11

2 Related Work

Although Music Genre Classification is a vibrant research field, it is difficult to find papers based directly on Million Song Dataset Benchmarks (MSDB).

Nevertheless, several classification models have been trained on features provided by MSD. For instance, [4] derived beat-aligned timbre and chroma features from music audio data contained in MSD in an effort to train a convolutional Deep Belief Network on all the data, and then used the computed parameters to initialize a convolutional multilayer perceptron. It achieved 29.5 % accuracy on genre recognition task with 20 genres and concluded that the gains obtained with pretraining are rather modest and are not advantageous for genre classification. Another interesting study has been conducted by [7], where authors proposed a framework of model blending based on combing features from audio and lyrics, which lead to accuracy of 38.6 % (10 hand-picked genres).

Several articles that exploit Deep Belief Networks in the context of automatic music genre classification have been published. In [14], DBN have been trained on GTZAN [12] dataset using greedy, layer-wise, unsupervised learning algorithm with short and long-term features, additionally fine-tuned via back propagation algorithm. They were able to reach 78.7 % accuracy, whereas best-performing, widely used classifier (SVM) achieved 75.9 % on the same dataset. It is worth noticing, that in GZTAN dataset there are only 10 classes, which renders the classification task easier. For comparison, our highest scoring algorithm (Random Trees) trained on dataset with 13 classes reached 62 % accuracy. One can claim, that in the setting of 10 genres classification task the performance of Random Trees would appropriately increase.

To our knowledge, our study is the first to evaluate Deep Belief Networks on MSDB dataset in the context of music genre recognition.

Authors of MSDB conducted a preliminary analysis of the classification algorithms such as Naive Bayes, Support Vector Machines, k-nearest Neighbours ($k = 1$), J48 Decision Tree and Random Forest using WEKA Machine Learning Toolkit [10]. They received best results on Statistical Spectrum Descriptors (with 168 dimensions) feature dataset with MSD Allmusic Genre Dataset (21 classes) with 66 % training set split. Highest accuracy has been achieved with Support Vector Machines and k-NN classifiers, both of which scored more than 27 % on SSD feature set. This is in contrast to our results, which suggest that Random Forest and Decision Trees are top performers.

A comprehensive study of recent advances in Music Genre Recognition can be found in [11].

3 Dataset

Million Song Dataset (MSD) is a freely-available collection of audio features and metadata for a million contemporary popular music tracks. It contains approximately 280 GB of data with 1,000,000 song files and 44,745 unique artists [2]. Traditionally, due to licensing issues, research in music information retrieval on commercial-scale dataset was limited to the industry. However, this has been circumvented by providing precomputed features instead of raw audio. The MSD enables researchers to test algorithms on a large-scale collection in real-world-like environments [10]. Unfortunately, some authors [9] claim that the absence of accurate documentation of the extraction algorithm renders the audio features provided by MSD of limited use.

In this study we used Million Song Dataset Benchmarks (MSDB) where wide range of audio features have been extracted from audio samples downloaded from external content provider. Detailed description of extraction procedure and employed software can be found in [10]. An overview of features is given in Table 1.

There are three datasets mapping individual tracks to their appropriate, expert annotated genre. The first one, the MSD Allmusic Genre Dataset (MAGD) consist of 21 genres. The second one, MSD Allmusic Top Genre Dataset (top-MAGD), consist of 13 genres—the top 10 genres from MAGD including three additional ones (Vocal, Folk and New Age). The last one, the MSD Allmusic Style Dataset (MASD), has 25 classes. All classes from aforementioned datasets are listed in Tables 2 and 3.

For detailed description of data collection and dataset building please refer to [10].

In order to facilitate repeatability of experiments, several partitions of the dataset have been prepared in the MSDB. We have decided to use stratified partition with frequently used 2/3 training and 1/3 test split and partition with fixed number of training samples, equally sized for each class with 2,000 samples per genre.

Table 1 Overview of selected features from Million Song Dataset Benchmarks

#	Feature set	Dimensions
1	Rhythm patterns	1440
2	Rhythm histograms	60
3	Temporal rhythm histograms	420
4	MARSYAS timbral features	124
5	MFCC features	26

Table 2 MSD Allmusic Genre Dataset (MAGD)—upper part represents the MSD Allmusic Top Genre Dataset (Top-MAGD)

Genre name	Number of tracks
Pop/rock	238,786
Electronic	41,075
Rap	20,939
Jazz	17,836
Latin	17,590
R&B	14,335
International	14,242
Country	11,772
Religious	8,814
Reggae	6,946
Blues	6,836
Vocal	6,195
Folk	5,865
New age	4,010
Comedy/spoken	2,067
Stage	1,614
Easy listening	1,545
Avant-garde	1,014
Classical	556
Childrens	477
Holiday	200
Total	**422,714**

4 Algorithms

In this study several classification algorithms have been compared: Decision Tree (DT), Random Forest (RF), Multi-class Support Vector Machine (SVM), Multinomial Logistic Regression (LR) and Deep Belief Network (DBN). Short description of selected algorithms are given in the following subsections. As far as algorithm's implementations are concerned, Apache Spark's MLlib has been used for training Decision Trees, Random Forest, and binary classifiers such, as SVM and Logistic Regression. Due to the fact, that Multi-class classification is currently not supported for SVM and Logistic Regression, we have chosen to implement them in Scala, building upon MLlib binary classificators. Deep Belief Networks have been trained using DL4J, an open-source library written for Java and Scala.

Table 3 MSD Allmuisc Style Dataset (MASD)

Genre name	Number of tracks
Pop indie	18,138
Rock college	16,575
Rock contemporary	16,530
Hip hop rap	16,100
Dance	15,114
Metal alternative	14,009
Pop contemporary	13,624
Rock hard	13,276
Rock alternative	12,717
Experimental	12,139
Country traditional	11,164
Rock neo psychedelia	11,057
Electronica	10,987
Metal heavy	10,784
Jazz classic	10,024
Metal death	9,851
Folk international	9,849
Punk	9,610
Pop latin	7,699
Gospel	6,974
Blues contemporary	6,874
Grunge emo	6,256
RnB soul	6,238
Reggae	5,232
Big band	3,115
Total	**273,936**

4.1 Decision Tree

Decision tree is a greedy classification algorithm that performs a recursive binary split of the feature space [5]. In our study Gini impurity measure has been used as a splitting criterion. Our preliminary study has shown that it yields best results on Million Song Dataset. Moreover, based on empirical considerations (grid-search with 3-fold cross-validation), the maximum depth of a tree has been limited to 10 and the number of bins used when discretizing continuous features has been set to 64.

4.2 Random Forest

Random forests are ensembles of Decision Trees [5]. They are especially appealing in the context of big-data because of the fact, that individual trees can be built independently, thus learning them is inherently parallel. Since trees should be built on subset of data, it straightforward to parallelize the whole process. As it was the case with Decision Trees, the hyperparameters have been optimised using grid-search in conjunction with 3-fold cross-validation.

4.3 Multi-class SVM

The Support Vector Machine (SVM) is a state-of-the-art linear binary classification algorithm. In this study we have decided to take advantage of the fact, that (in opposition to SVMs employing kernel-trick) linear SVM scale well with the number of examples [3, 5]. Moreover, the naive way of explicitly computing non-linear features does not scale well with the number of input features and in the case of Million Song Dataset occurred to be computationally prohibitive.

We have decided to employ One-Vs-The-Rest strategy to construct Multi-class SVM. It is worth noticing, that in the case of a multi-label classification task, choosing the category based on maximal posterior probability over all classes is the Bayes optimal decision for the equal loss case [8]. Unfortunately, Standard SVM do not directly provide posterior probability estimates, therefore those parameters have to be manually calibrated. Based on the work [8] we have decided to use a parametric model to fit the posterior probabilities directly. Following parametrization has been used:

$$P(y = 1 \mid x) \approx P_{A,B}(f) = \frac{1}{1 + \exp(Af + B)}, \quad \text{where } f = f(x). \tag{1}$$

The parameters A and B from Eq. 1 are fit by solving following maximum likelihood problem from a training set (f_i, y_i):

$$\arg \min_{A,B \in \mathbb{R}} F(z) = -\sum_{i=1}^{n} \Big(y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \Big), \tag{2}$$

where $p_i = P_{A,B}(f_i)$ and n is the number of training examples.

According to [8] for linear SVMs the bias introduced by using the same data set for training the model and estimating parameters from Eq. 1 is negligible for large datasets. This is due to the fact, that as in almost all cases, a maximum of $N + 1$ support vectors will lie on the margin (were N is the dimensionality of the input vector). In our case of linear SVM no additional preprocessing is needed.

4.4 Deep Belief Network

For classification, a DBN with ℓ layers models the joint distribution between target y, observed variables x_j and i hidden layers $\mathbf{h}^k$ made of all binary units h_i^k, as follows [6]:

$$P(\mathbf{x}, \mathbf{h}^1, \ldots, \mathbf{h}^\ell, y) = \left(\prod_{k=1}^{\ell-2} P(\mathbf{h}^k | \mathbf{h}^{k+1}) \right) P(y, \mathbf{h}^{\ell-1}, \mathbf{h}^{\ell-1}), \tag{3}$$

where $\mathbf{x} = \mathbf{h}^0$, $P(\mathbf{h}^k | \mathbf{h}^{k+1})$ is a conditional distribution for the visible units conditioned on the hidden units of the Restricted Boltzmann Machine (RBM) at level k and $P(y, \mathbf{h}^{\ell-1}, \mathbf{h}^{\ell-1})$ is the visible-hidden joint distribution in the top-level RBM. A RBM has the following form:

$$P(\mathbf{x}, \mathbf{h}) \propto \exp(\mathbf{h}' W \mathbf{x} + b' \mathbf{x} + c' \mathbf{h}) \tag{4}$$

with parameters $\theta = (W, b, c)$. During network training, contrastive divergence has been run only once, which has been shown to work surprisingly well. In an effort to adapt RBM to accept continuous input we have employed a Gaussian transformation on the visible layer and a rectified-linear-unit transformation on the hidden layer. Initial weights has been sampled from uniform distribution. We have decided to use different architectures depending on feature dataset. Number of hidden layers and their size has been chosen after initial empirical investigation.

5 Results

Our experimental results show that Random Forests and Decision Trees outperform Multi-class SVM, Multinomial Logistic Regression and naively trained DBN independently of the chosen features, genre dataset or benchmarking partitions. Highest scoring algorithm (Random Forest) trained on dataset with 13 classes reached accuracy of 62 %.

Random Forests and Decision Trees are also easily parallelizable, which effectively makes them the algorithm of choice in setting with time-limited computing resources. In the case of DBN hyperparameter tuning, using grid-search has occurred to be computationally intractable and time constraints become an important issue. All of this led to unsatisfactory results.

In all cases Multi-class SVM had at least 10 % of accuracy less than Random Forest. One possible remedy for SVM's poor performance may be changing the way multiple classes are handled. It is worth investigating whether employing One-Vs-One strategy, instead of One-Vs-The-Rest, can bring benefits. We have conducted preliminary study in which 78, i.e. all possible genre pairs in top-MAGD dataset, One-Vs-One Random Forests have been trained. Results are promising: mean F-1 measure was equal to 81 and 90 % of scores were above 68 %. Highest and lowest

Table 4 5 highest One-Vs-One classfier F-1 scores for top-MAGD genres on Rhythm Histograms dataset

Genre	F-1 score
New age vs Pop rock	0.9834
Pop rock vs folk	0.9761
Vocal vs Pop rock	0.9746
Blues vs Pop rock	0.9721
Pop rock vs Reggae	0.9719

Table 5 5 lowest One-Vs-One classfier F-1 scores for top-MAGD genres on Rhythm Histograms dataset

Genre	F-1 score
Blues vs Folk	0.6351
International vs RnB	0.6313
International vs Latin	0.6055
RnB vs Latin	0.5914
Vocal vs Folk	0.5818

scoring pairs are shown in Tables 4 and 5 respectively. It is also worth noticing, that SVM's performance decreased with the increase in the size of the input vector: SVM reached highest F-1 scores on MFCC features dataset (26 features) and lowest on Temporal Rhythm Histograms feature dataset (420 features).

Interestingly, audio features perform quite poorly on guitar based styles from MSD Allmusic Style Dataset (MASD). Quite unexpectedly, classifiers have problems distinguishing between *Rock College* and *Metal Heavy*. Similar observation have been made by [7].

Because of the space considerations we have decided to restrict result presentation to Temporal Rhythm Histograms, Rhythm Histograms, MFCC features and MARSYAS timbral features datasets with 66 % training, stratified set split. Additionally, we have included random classifier choosing each class with equal probability, denoted R, as our point of reference. Results are shown in Tables 6, 7, 8 and 9.

Table 6 Classification F-1 score for genre datasets on Temporal Rhythm Histograms feature dataset

Genre	DT	RF	SVM	LR	DBN	R
MAGD	0.59	0.60	0.12	0.11	0.05	0.048
top-MAGD	0.61	0.62	0.20	0.25	0.09	0.077
MASD	0.15	0.17	0.06	0.07	0.05	0.040

Table 7 Classification F-1 score for genre datasets on Rhythm Histograms dataset

Genre	DT	RF	SVM	LR	DBN	R
MAGD	0.58	0.59	0.37	0.35	0.09	0.048
top-MAGD	0.60	0.61	0.38	0.36	0.11	0.077
MASD	0.16	0.18	0.10	0.09	0.05	0.040

Table 8 Classification F-1 score for genre datasets on MFCC features dataset

Genre	DT	RF	SVM	LR	DBN	R
MAGD	0.58	0.59	0.49	0.48	0.05	0.048
top-MAGD	0.61	0.62	0.51	0.50	0.08	0.077
MASD	0.18	0.21	0.16	0.15	0.05	0.040

Table 9 Classification F-1 score for genre datasets on MARSYAS timbral features dataset

Genre	DT	RF	SVM	LR	DBN	R
MAGD	0.61	0.62	0.52	0.53	0.11	0.048
top-MAGD	0.64	0.65	0.55	0.54	0.10	0.077
MASD	0.21	0.22	0.17	0.18	0.05	0.040

6 Conclusions and Future Work

Although Random Forest and Decision Trees yield satisfactory results, several things can be done to improve the classifier performance. To be more explicit, the following three aspects are worth studying:

1. Inclusion of text features;
2. Exploration of different pairwise coupling methods in an effort to obtain better probability estimates for Multi-class classification;
3. DBN adjustment and optimization;

Ad (1) At the intuitive level, the inclusion of lyrical features can boost classifier accuracy, as lyrics cover semantic information about song's contents not available in the audio features. Additionally, in many music sub-genres the dividing line is often subtle and runs through the topics discussed rather than artistic means of expression. Encouraging results have been obtained in [7], where authors achieved 40 % accuracy based solely on bag-of-words lyric features. One can suggest, that, to a certain extent, lyrics and audio features are orthogonal and combining them in single model can yield better accuracy. It is also worth investigating whether using Markov Models or TF-IDF can bring additional benefits.

Ad (2) Platt [8] uses a Levenberg–Marquardt algorithm to solve Eq. 2. Instead other methods for solving unconstrained optimization can be used in an effort to boost classifier performance. Furthermore, as show in [13] pairwise coupling in con-

junction with Platt scaling can yield satisfactory results. It is worth studying whether employing different probability estimation methods can boost classifier accuracy.

Ad (3) In future work, we would like to refine couple of aspect concerning the architecture of the network, such as the number and size of the hidden layers. Our experience with deep architectures shows that hyper-parameter optimization in large and multilayer models is not by any means an easy task. Recently, [1] showed that two sequential model-based optimization algorithms could outperform domain experts in the tuning Deep Belief Networks. It seems that it would be beneficial to incorporate algorithms proposed by [1] in subsequent works involving DBN.

Moreover, one can experiment with alternative ways of examining music genre classification system performance, for example, using multi-label classification may provide additional insight.

References

1. Bergstra, J., Bardenet R., Bengio Y., Kegl, B.: Algorithms for hyper-parameter optimization. In: Proceedings of the 24th Neural Information Processing Systems (NIPS 2011) (2011)
2. Bertin-Mahieux, T., Ellis, D., Whitman B., Lamere P.: The million song dataset. In: Proceedings of the 12th International Conference on Music Information Retrieval (2011)
3. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2007)
4. Dieleman, S., Brakel, P., Schrauwen, B.: Audio-based music classification with a pretrained convolutional network. In: Proceedings of the 12th International Society for Music Information Retrieval Conference (2011)
5. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. Springer, New York (2001)
6. Hinton, G.E., Osindero, S., Teh, Y.: A fast learning algorithm for deep belief nets. Neural Comput. **18**, 1527–1554 (2006)
7. Liang, D., Gu, H., O'Connor, B.: Music genre classication with the million song dataset. Machine Learning Department, CMU (2011). http://www.ee.columbia.edu/~dliang/files/FINAL.pdf
8. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers, pp. 61–74. MIT Press, Cambridge (1999)
9. Schindler, A., Rauber, A.: Capturing the temporal domain in Echonest Features for improved classification effectiveness. In: Proceedings of the 10th International Workshop on Adaptive Multimedia Retrieval (2012)
10. Schindler, A., Mayer, R., Rauber, A.: Facilitating comprehensive benchmarking experiments on the million song sataset. In: Proceedings of the 13th International Society for Music Information Retrieval Conference (2012)
11. Strum, B.L.: A survey of evaluation in music genre recognition. Adaptive multimedia retrieval: semantics, context, and adaptation. Lect. Notes Comput. Sci. **8382**, 29–66 (2014)
12. Tzanetakis, G., Cook, P.: Musical genre classification of audiosignals. IEEE Trans. Audio Speech Process. **10**(5), 293–302 (2002)
13. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. JMLR **5**, 975–100 (2004)
14. Yang, X., Chen, Q., Zhou, S., Wang, X.: Deep belief networks for automatic music genre classification, In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (2011)