AUTOMATIC CLASSIFICATION OF DRUM MACHINE SAMPLES USING ARTIFICIAL NEURAL NETWORKS

Roman B. Gebhardt, Jonas Margraf, Alexander Voß

Audio Communication Group
Technische Universität Berlin

ABSTRACT

In this paper, we describe the application of Artificial Neural Networks for drum machine classification. We extract a set of different audio features from the samples and input it to the model. We present different evaluations based on various databases as well as a comparison of the Artificial Neural Networks approach to other machine learning methods, namely Support Vector Machines, k-Nearest-Neighbour Classifiers and Decision Trees. Our results rom the Artificial Neural Networks model vary a lot with the size and kind of the used database. Especially for small ones it can compete with the other named models. Especially hi-hats seem to be well classifiable with Artificial Neural Networks.

***Index Terms*—** Machine Learning, Drum Classification, Artificial Neural Networks

1. INTRODUCTION

During the last decade, a growing interest in machine learning and pattern recognition has been brought up in the area of audio engineering. In line with the steadily expanding amount of data the internet offers, a way to distinguish between different kinds of audio data is getting more and more important. While automatic genre classification is an important issue for radio and streaming platforms which also represents a major part of audio focussed machine learning research, classification tasks can also be interesting e.g. for musicians who are looking for sounds in big sample libraries. This puts the focus on instrument recognition and classification which is based on a (mostly heavyset) collection of audio features and therefore represents a perfect example of use for machine learning. Several publications have especially thematised drum set classification [1, 2, 3, 4]. In these works, used machine learning models are Support Vector Machines, Decision Trees, Hidden Markov Models, k-Nearest-Neighbour and Bayes-Classifiers. However, to the best of our knowledge, there exists no work that approaches a percussion classification problem with an Artificial Neural Network model, which has gained more attention in the last years. On top of that, research has only been targeted on the recognition of sounds recorded from drum sets instead of such that were synthesised. Taking into consideration that, as mentioned above, a classifier would especially be of interest for producers of electronic music, who might want to sort or search in their sample collections, it should be obvious to use drum machine samples as database. These two points represent a strong motivation to attempt a drum machine classification by aid of Artificial Neual Networks.

2. METHOD

2.1. Dataset

While most studies build upon recorded drum sounds, which are for the most part drawn from various sample CDs, we focus on drum sound samples drawn from drum machines. We state that for a task like ours this brings the advantage that we do not have to deal with velocity issues and the fact that a drum hit will never sound the same even if performed by the same player with the same material. Apart from that, in an application like a music production software (which might be a possible environment for our system), the user will most likely deal with drum machine sounds in many cases. For this study, we used Chih-Wei Wu's *200 Drum Machine* database [5] which contains samples of percussion sounds taken from 200 different drum machines. The samples were arbitrarily labelled by instrument that was aimed to model, e.g. *kick* or *hi-hat*, however for every drum machine there were different numbers as well as kinds of instruments. In total, the numbers of samples per instrument class in our dataset were as follows: 746 kick drums, 862 snare drums, 159 claps, 596 toms, 305 closed hihats, 245 open hihats, 101 crash cymbals and 117 ride cymbals.

2.2. Descriptors

To gain descriptors for the machine learning task, we applied an audio feature extraction to the audio signals from the database. Since there exists high accordance of the audio features used as descriptors amongst the authors of related works (compare [1, 2]), we followed with the extraction of the most common features. These can be grouped into three categories, namely energy, spectral and time related. In terms of the energy related features, we computed the overall RMS (root means square) of the signal as a very basic loudness descriptor. We also computed the band-wise RMS to achieve information about the spectral distribution of the sounds' energy. To achieve this, we applied different band-pass filterings to the signal before computing the RMS values. These filterbands inherited frequencies typical for kick, snare and hi-hat (for comparison see figure 1).

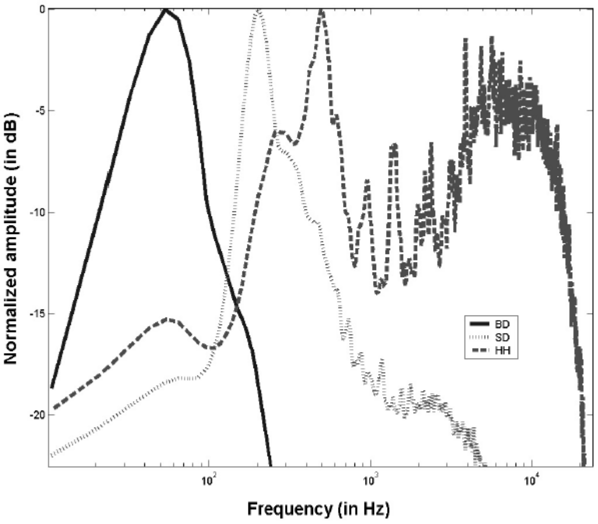


Fig. 1. *Amplitude spetra of different drum instruments (accumulated). Graphic taken from [2]*

We expected high RMS values for sounds that were filtered with a 'correct' filterband and vice versa low for 'wrong' filterbands. As descriptors, we used the differences between band-wise RMS and overall RMS as well as the differences between the band-wise RMS values amongst each other. For the temporal features, we extracted the *zero-crossing rate* as a raw pitch descriptor displaying low values for membrane sounds (with a discrete pitch). The *time spread* was also computed, which we expected to give information about the sustain of a sound, whereas the *temporal crest factor* should rate its impulsiveness. The *temporal centroid* was chosen to describe the "volume envelope", meaning if the main energy of a sound is distributed in the beginning or the end of the sound's duration. The spectral features contained the *spectral centroid* as both pitch and tonalness descriptor as well as *spectral spread, spectral skewness* and *spectral kurtosis*, all describing the characteristics of timbral shape. The *spectral crest factor* and *spectral flatness* were chosen to describe the noisiness of the sound and the *spectral rolloff* for the timbre's distribution of energy in high frequencies. The *spectral slope* depends on the spectral decay and therefore contains information about the spectral envelope over time. This is also the case for *Mel Frequency Cepstral Coefficients (MFCCs)* of which we calculated the first 13. For the feature extraction, we based our code mainly on the implementations offered online by Alexander Lerch [6]. His functions were either directly adopted or modified to our needs. To allow comparison between the obtained vectors x_f of sample-wise values for a certain feature f, we applied a transformation to standard score z_f, according to:

$$z_f = \frac{x_f - \mu}{\sigma} \tag{1}$$

with μ and σ representig the mean and the standard deviation of x_f respectively.

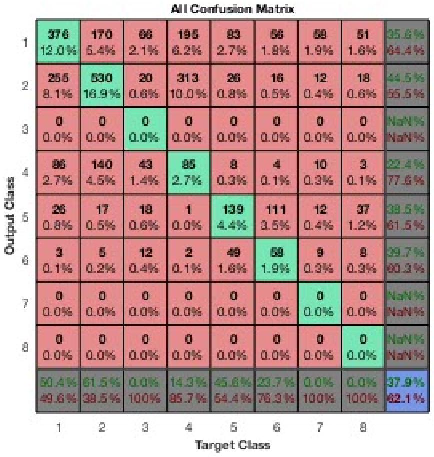
2.3. Instrument Classification

After the preprocessing stage in which we extracted the previously mentioned audio features, we used the MATLAB Statistics and Machine Learning and Neural Network toolboxes to classify our dataset and compare the results of the different machine learning algorithms. We focussed specifically on pattern recognition with the Neural Network toolbox (further referred to as ANN) and on Support Vector Machines (SVM), k-Nearest Neighbor (kNN) and Bagged Decision Trees as the algorithms to compare. For the neural network pattern recognition algorithm, we used 3 hidden layers and divided the dataset as follows: 70% were used for training, 15% for validation and 15% for testing. The following output classes were defined: 1 (kick drums), 2 (snare drums), 3 (claps), 4 (toms), 5 (hihat closed), 6 (hihat open), 7 (crash), 8 (ride). With the Statistics and Machine Learning toolbox we were able to test several variants of the SVM (linear, quadratic, cubic), kNN (fine, medium, cubic, cosine) and decision trees (complex, boosted, bagged) algorithms.

3. EVALUATION

3.1. Comparison of ANN, Bagged Trees, SVM and kNN

The overall most succesful algorithm for classification of the different drum machine samples was the Bagged Tree algorithm, which correctly identified 84.3% of all samples. Cubic SVM had the highest number of correct classifications of all SVM algorithms with 81.2%. The best performing kNN variant was Fine kNN with 71.2% accuracy. The neural network based pattern recognition algorithm was only able to correctly classify 37.9% of all samples. It is noteworthy that the neural network did not identify any of the samples belonging to the clap, ride cymbal or crash cymbal classes, as can be seen in Figure 2.

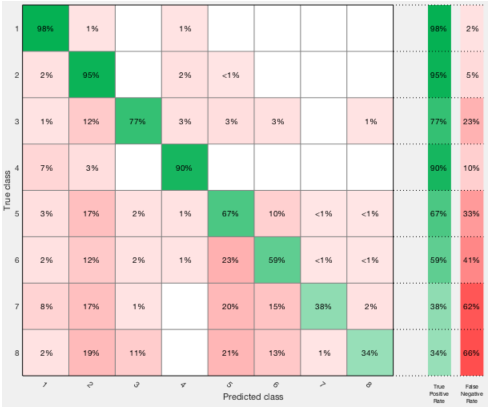


	1	2	3	4	5	6	7	8	
1	376 12.0%	170 5.4%	66 2.1%	195 6.2%	83 2.7%	56 1.8%	58 1.9%	51 1.6%	35.6% 64.4%
2	255 8.1%	530 16.9%	20 0.6%	313 10.0%	26 0.8%	16 0.5%	12 0.4%	18 0.6%	44.5% 55.5%
3	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
4	86 2.7%	140 4.5%	43 1.4%	85 2.7%	8 0.3%	4 0.1%	10 0.3%	3 0.1%	22.4% 77.6%
5	26 0.8%	17 0.5%	18 0.6%	1 0.0%	139 4.4%	111 3.5%	12 0.4%	37 1.2%	38.5% 61.5%
6	3 0.1%	5 0.2%	12 0.4%	2 0.1%	49 1.6%	58 1.9%	9 0.3%	8 0.3%	39.7% 60.3%
7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
8	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
	50.4% 49.6%	61.5% 38.5%	0.0% 100%	14.3% 85.7%	45.6% 54.4%	23.7% 76.3%	0.0% 100%	0.0% 100%	37.9% 62.1%

Fig. 2. *Confusion Matrix for the ANN classification with the 200 Drum Machines dataset. The output classes are labelled as follows: 1 (kick drums), 2 (snare drums), 3 (claps), 4 (toms), 5 (hihat closed), 6 (hihat open), 7 (crash), 8 (ride).*

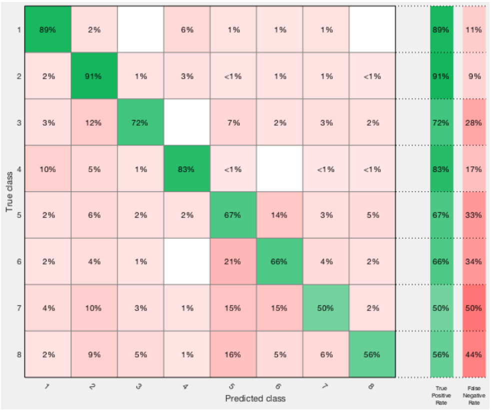
Best performances are shown for the classes snare (of which 44.5% were correctly classified), followed by open and closed hihats (39.7%, 39.5% respectively) and kick drums (35.6%). In comparison with the relatively good performance of the other mambrane sounds kick and snare, the toms are only classified correctly in 22.4% of all cases. Genereally, membrane instruments are almost exclusively confused with other membranes. Many confusions exist as well between the open and closed hihat classes. Surprisingly, all classes that have 0% true positive classifications are mostly classified as kick drums, which however does not represent the class with the highest general number of samples. A reason for this could be the (also perceptually noticable) higher diversity of kick sounds in comparison to snares, which represents the class with most samples.

A major difference in comparison with the three other models outperforming the ANN model is that three membrane classes are the ones holding the highest accuracy, while the ANN performs better for the hihat classes than for the toms. The confusion matrices of Bagged Tree, SVM and kNN are shown in Figures 3, 4 and 5 respectively. This is an interesting inspection, as membrane sounds should generally be more easy to describe, as they are amongst the most pitched percussion instruments. Therefore, they can be distinguished well by descriptors as e.g. the zero crossing rate, as it is not possible for more noisier sounds [4]. This aligns well with human experience, where one can say that it is easier to differentiate between tom and snare than between open and closed hi-hat.



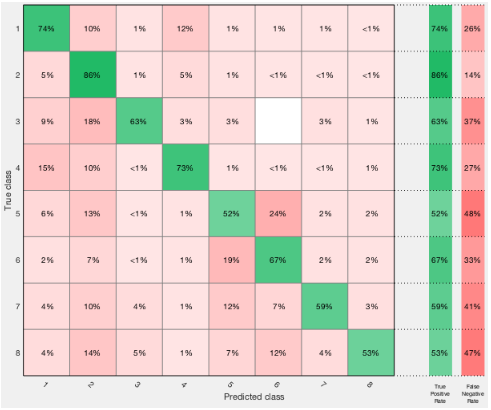
	1	2	3	4	5	6	7	8	True Positive Rate	False Negative Rate
1	98%	1%		1%					98%	2%
2	2%	95%		2%	<1%				95%	5%
3	1%	12%	77%	3%	3%	3%		1%	77%	23%
4	7%	3%		90%					90%	10%
5	3%	17%	2%	1%	67%	10%	<1%	<1%	67%	33%
6	2%	12%	2%	1%	23%	59%	<1%	<1%	59%	41%
7	8%	17%	1%		20%	15%	38%	2%	38%	62%
8	2%	19%	11%		21%	13%	1%	34%	34%	66%

Fig. 3. *Confusion Matrix for The bagged Tree classification with the 200 Drum Machines dataset. Class labels as in Fig. 2*



	1	2	3	4	5	6	7	8	True Positive Rate	False Negative Rate
1	89%	2%		6%	1%	1%	1%		89%	11%
2	2%	91%	1%	3%	<1%	1%	1%	<1%	91%	9%
3	3%	12%	72%		7%	2%	3%	2%	72%	28%
4	10%	5%	1%	83%	<1%		<1%	<1%	83%	17%
5	2%	6%	2%	2%	67%	14%	3%	5%	67%	33%
6	2%	4%	1%		21%	66%	4%	2%	66%	34%
7	4%	10%	3%	1%	15%	15%	50%	2%	50%	50%
8	2%	9%	5%	1%	16%	5%	6%	56%	56%	44%

Fig. 4. *Confusion Matrix for the SVM classification with the 200 Drum Machines dataset. Class labels as in Fig. 2*



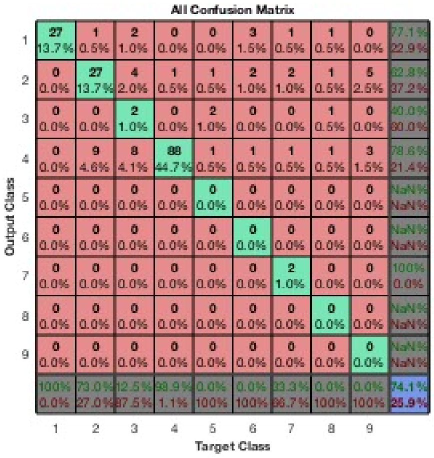
	1	2	3	4	5	6	7	8	True Positive Rate	False Negative Rate
1	74%	10%	1%	12%	1%	1%	1%	<1%	74%	26%
2	5%	86%	1%	5%	1%	<1%	<1%	<1%	86%	14%
3	9%	18%	63%	3%	3%		3%	1%	63%	37%
4	15%	10%	<1%	73%	1%	<1%	<1%	1%	73%	27%
5	6%	13%	<1%	1%	52%	24%	2%	2%	52%	48%
6	2%	7%	<1%	1%	19%	67%	2%	2%	67%	33%
7	4%	10%	4%	1%	12%	7%	59%	3%	59%	41%
8	4%	14%	5%	1%	7%	12%	4%	53%	53%	47%

Fig. 5. *Confusion Matrix for The kNN classification with the 200 Drum Machines dataset. Class labels as in Fig. 2*

Similarly to Tanghe et al. [2], we did not improve their results by dimensional reduction (which they stated to be a consequence of proper choice of features, which we also found to be the case throughout our paper reviews). In consequence, all results named above are derived from all features in equal weighting.

3.2. Further evaluation of the ANN model

Due to the bad performance of the ANN method, we first tested with equal distributions of 100 samples per class to exclude the possibility of the class sizes to distort our results. However, this did not improve the results, on the contrary the overall accuracy even decreased to only 16.4%. We then tested the ANN with the samples of only one despite of the 200 drum machines in order to minimise the variety of sounds. For this, we chose 124 samples from the Roland TR-808 labelled equally to the 8 categories as mentioned above. With this, an accuracy of 96% was achieved, which now even indicated overfitting. To also eliminate this possible item of review, we tested the trained machine with samples form a similar drum machine (Roland TR-909). This way, we could achieve an accuracy of 74.1%, which now lies in the range of the other models (see 3.1). However, it must be mentioned that this comparison must be interpreted in careful manner, as the numbers derive from the analysis of different databases. The corresponding confusion matrix is shown in 6.


All Confusion Matrix
Output Class
Target Class

	1	2	3	4	5	6	7	8	9	
1	27 13.7%	1 0.5%	2 1.0%	0 0.0%	0 0.0%	3 1.5%	1 0.5%	1 0.5%	0 0.0%	77.1% 22.9%
2	0 0.0%	27 13.7%	4 2.0%	1 0.5%	1 0.5%	2 1.0%	2 1.0%	1 0.5%	5 2.5%	62.8% 37.2%
3	0 0.0%	0 0.0%	2 1.0%	0 0.0%	2 1.0%	0 0.0%	0 0.0%	1 0.5%	0 0.0%	40.0% 60.0%
4	0 0.0%	9 4.6%	8 4.1%	88 44.7%	1 0.5%	1 0.5%	1 0.5%	1 0.5%	3 1.5%	78.6% 21.4%
5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 1.0%	0 0.0%	0 0.0%	100% 0.0%
8	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
9	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
	100% 0.0%	73.0% 27.0%	12.5% 87.5%	98.9% 1.1%	0.0% 100%	0.0% 100%	33.3% 66.7%	0.0% 100%	0.0% 100%	74.1% 25.9%

Fig. 6. *Confusion Matrix for the ANN classification. Training set from Roland TR-808 samples, test set from TR-909 samples.*

The results now strongly resemble the results of Bagged Trees, SVM and kNN from 3.1. Also here, the membrane classes obtain the best results, where especially the toms now show a very good performance in opposite to before. Albeit, a limiting fact that has to be mentioned is, that the performance for the cymbal classes (which are represented by less samples) is now extremely weak which might be an effect of their small sizes.

4. CONCLUSION

In this paper, we described a comparison of various machine learning methods. We evaluated the performance of SVM, kNN, Decision Tree and (with a special focus) ANN models for the classification of the samples of 200 different drum machines. While the first three perform equally well, the ANN classifier did not show good results when trained and tested with the used database. Further testing showed that for smaller databases and lower variance of the sound characteristics, the accuracy of the ANN could be drastically improved. ANN showed a relatively good accuracy for the classification of hi-hats which is a promising fact for further research, as especially unpitched percussion sounds (to which hi-hats can be counted among) are relatively difficult to classify due to their noisy timbral characteristics. In terms of future work, we intend to test with more accurate databases to allow proper comparison with other machine learning models. We also plan to work on better descriptors for unpitched percussion sounds, as the features used so far in our opinion are too much low-level to allow an accurate classification of cymbals. In combination with our inspections of the good performance of the ANN for these, we state that a better classification results can be achieved.

5. REFERENCES

[1] P. Herrera, A. Yeterian, and F. Gouyon, “Automatic classification of drum sounds: A comparison of feature selection methods and classification techniques,” in *ICMAI ’02 Proceedings of the Second International Conference on Music and Artificial Intelligence*, 2002, pp. 69–80.

[2] K. Tanghe, S. Degroeve, and B De Baets, “An algorithm for detecting and labeling drum events in polyphonic music,” in *Proceedings of the 1st Annual Music Information Retrieval Evaluation Exchange*, 2011, pp. 11–15.

[3] M. Blass, “Drum pattern classification using hidden markov models,” in *Int. Workshop on Machine Learning and Music*, Prague, Czech Republic, 2013, pp. 11–14.

[4] V.M.A. Souza, G.E.A.P.A. Batista, and N.E. Soza-Filho, “Automatic classification of drum sounds with indefinite pitch,” in *International Joint Conference on Neural Network, 2015, Killarney*.

[5] “200 Drum Machines Dataset web presence,” http://www.hexawe.net/mess/200.Drum.Machines/, Accessed: 2016-09-27.

[6] “Audio Content Analysis web presence,” http://audiocontentanalysis.org/, Accessed: 2016-09-27.