Music Genre Classification Using Data Mining and Machine Learning

Nimesh Ramesh Prabhu, James Andro-Vasko, Doina Bein, and Wolfgang Bein

Abstract

With accelerated advances in internet technologies users make listen to a staggering amount of multimedia data available worldwide. Musical genres are descriptions that are used to characterize music in music stores, radio stations and now on the Internet. Music choices vary from person to person, even within the same geographical culture. Presently Apple's iTunes and Napster classify the genre of each song with the help of the listener, thus manually. We propose to develop an automatic genre classification technique for jazz, metal, pop and classical using neural networks using supervised training which will have high accuracy, efficiency and reliability, and can be used in media production house, radio stations etc. for a bulk categorization of music content.

Keywords

Automatic classification · Data mining · Machine learning · Music genre

52.1 Introduction

With accelerated advances in internet technologies users make listen to a staggering amount of multimedia data available worldwide. Apple's website iTunes, MP3.com, Napster.com, all boast millions of songs and over 15 genres Musical genres are descriptions that are used to characterize music in music stores, radio stations and now on the Internet. Music comes in many different types and styles ranging from traditional rock music to world pop, jazz, easy listening and bluegrass.

Data mining is a process of analyzing data from different perspectives and summarizing it into useful information that can be used to classify music samples. Basically data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Machine learning is a branch of artificial intelligence which works with construction and study of systems that can learn from data. The core of machine learning deals with representation and generalization. Representation of data instances and functions evaluated on these instances are part of all machine learning systems. Generalization is the property that the system will perform well on unseen data instances. Neural networks techniques will be used in this paper for classification.

Music choices vary from person to person, even within the same geographical culture. Presently Apple's iTunes and Napster classify the genre of each song with the help of the listener, thus manually. But manual classification is time consuming and classification is difficult when the song is in a language unknown to the listener. Classifying songs automatically into proper genres using machine learning rather than manual process which will save time and manpower and there is little research on this topic due to the difficulty to achieve low error rates [1].

We propose to develop an automatic genre classification technique for jazz, metal, pop and classical using neural networks using supervised training which will have high accuracy (between 80% and 90%), efficiency and reliability.

The paper is organized as follows. In Sect. 52.2 we present basic concepts, followed by related work in Sect. 52.3. A detailed description of our hardware-software system and what it achieves is given in Sect. 52.4. Experimental results

N. R. Prabhu · D. Bein (✉)
Department of Computer Science, California State University, Fullerton, Fullerton, CA, USA
e-mail: nimesh5@csu.fullerton.edu; dbein@fullerton.edu

J. Andro-Vasko · W. Bein
Department of Computer Science, University of Nevada, Las Vegas, NV, USA
e-mail: androvas@unlv.nevada.edu; wolfgang.bein@unlv.edu

S. Latifi (ed.), *Information Technology – New Generations*, Advances in Intelligent Systems and Computing 738,
https://doi.org/10.1007/978-3-319-77028-4_52

are shown in Sect. 52.5. Concluding remarks and future work are presented in Sect. 52.6.

52.2 Basic Concepts

Machine learning is a subset of artificial intelligence where programs and systems are able to learn how to accomplish a task by learning through a training algorithm and a large amount of data. Supervised learning is a learning method where a program or model is trained with inputs that have target outputs. In other words, the input variables are mapped to output variables, allowing the system to learn in an assisted manner and be able to perform classification by adjusting for errors [2]. Regression and classification are the most common tasks for supervised learning, and it is also the most commonly used form of machine learning.

The robust capability of neural networks has made it a trending flavor of machine learning due to the complexity of modern classification and pattern matching problems, in addition to the rise in availability of large datasets [3].

Unlike other and older methods of classification, neural networks function as both a feature extractor and a classifier, providing both efficiency and capability in a range of machine learning tasks. A neural network is a system that is designed to model the way a human brain processes and performs a task, and it achieves this by employing a massive interconnection of simple computing cells that work as a parallel distributed processor [2]. These computing cells are referred to as “neurons” and are also regarded as “nodes” in the context of discussing the architecture of neural networks. Neural networks are visualized as consisting of multiple layers of nodes that are connected to each other. The basic structure of a simple neural network in modern applications consists of three layers: an input layer, hidden layer (or middle layer), and output layer. The input layers consist of the number of attributes or values, such as the 17 values of the five descriptors. The middle layer consists of one or more hidden layers, of which are responsible for the majority of the transformations on the input data into output signals, depending on their various synaptic weights and activation function [4]. The last layer, the output layer, combines all the signals or outputs from the last hidden layer and performs a classification or output transformation, such as the categorization of the song into the four genre. Most often, the output of the neural network does not match the actual (correct) result, so the error values acquired by comparing the output of the neural network against the actual target value for multiple such instances are then propagated backwards to each layer of the network to do adjustments to the weights. This process is called backpropagation and it is what gives the ability of neural networks to learn and improve from input data and solve problems beyond those that are only linearly separable [5]. Thus, backpropagation provides a method of splitting the total output error backwards into error values per node in every layer. The amount of which to adjust the weights based on the error values is handled by the method called gradient descent. Gradient descent utilizes the error function realized from the training process of the neural network and selects adjustments to the synaptic weights that causes a decrease in the slope of the error function until it reaches the minimum [1]. The change in synaptic weights via these adjustments from gradient descent can be very small, especially if it is applied on a per input basis, but over time it will cause the error value to converge to the minimum of the error function after many training samples [4].

There has been work done in the area of automated categorization [6]. This involves labeling texts to a set of predefined categories, this is otherwise known as text categorization. Text categorization is applied to document indexing, document filtering, metadata generation, word sense disambiguation, and in any scenario where document organization is required. In the past, text categorization was based on knowledge engineering, which classified documents under a set of given categories by manually defining a set of rules to the expert knowledge engine to perform the classification. This method has become less popular and this mechanism has been applied by using a machine learning paradigm where a general inductive process automatically builds a text classifier by learning from a set of pre-classified documents.

Neural networks also provide a sound knowledge representation for information retrieval systems. In an information representation using a neural network, each node can be a keyword or an author and a link used as an association in the network. Information is retrieved using a parallel relaxation method where nodes are activated in parallel and are traversed until the network reaches a stable state using a single-layered interconnected neurons and weighted links. The strategy is explained in [7].

Symbolic learning has also been applied for information retrieval systems. In [8], the ID3 and ID5R algorithms were introduced. The ID3 is a decision tree based algorithm that used divide and conquer strategy to classify mixed objects into their associated classes based on the attribute values of the objects. Each node from the tree contains either a class name (leaf node) or contains an attribute test (a non-leaf node). Every training instance is an attribute-value pair. The ID3 strategy picks an attribute and categorizes to a list of objects based on this attribute. Using the divide and conquer approach, the ID3 method minimizes the number of expected tests to classify an object.

There has been work done in the area of genetic algorithms involving information retrieval. The method in which a genetic algorithm solves a problem is that given a problem, we apply a function on the input (normally known as a fitness function) and obtain a result from the fitness function.

Typically, we have a set of various inputs and we apply the fitness function onto each of the inputs. Once we generate the outputs we place them into a pool in which they are used again with the fitness function. When new solutions are added into the pool, certain solutions get discarded if they do not show improvement from previous generations. Then the idea is that the fitness functions generates new solutions from the pool and then inserts new solutions and/or discards new or old solutions (which is a generation), and this process continues until we obtain the desired solution.

Selecting a solution in the pool can be determined by applying a cross over which attempts to find the next best solution in the pool for the next generation and then we mutate the item to create a new generation. A genetic algorithm can be applied on NP problems to attempt to generate a solution quickly, or a quicker method than the brute force approach. The fitness function for a genetic algorithm can use some heuristic to speed up the process and try to obtain a solution without having too many generations.

Feature extraction is part of data mining technique in which set of features will be created by decomposing the original data. A feature is a combination of attributes that is of special interest and captures important characteristics of the data. A feature becomes a new attribute. Feature extraction make us describe data with a far smaller number of attributes than the original set. Feature extraction is an attribute reduction process which results in a much smaller and richer set of attributes.

52.3 Related Work

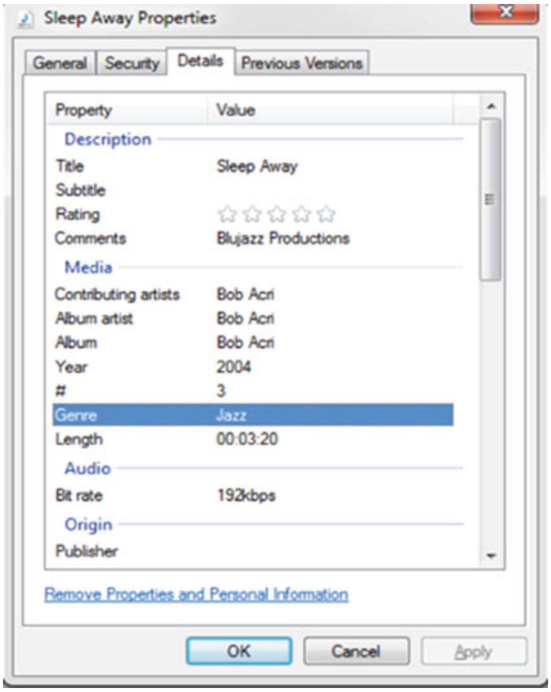
Genetic algorithms can be applied in information retrieval and document indexing, as in [9]. The keywords in a document are altered using genetic mutation and crossovers. The association of words with the documents are preserved in the chromosomes and each gene of the chromosome is a keyword associated to a document. After several generations and using a fitness function with the fitness score, the best population is generated which is a set of keywords that best describes the document. In [10] the authors extend the method to document clustering. Document clustering has been studied in [11, 12] where a genetic algorithm is applied on a weighted information retrieval system and a Boolean query was modified to improve recall and precision. In [13], a genetic algorithm approach is used for parallel information retrieval strategy.

Classifying songs automatically into proper genres using machine learning is a much needed but challenging task, due to the large rate of songs uploaded daily on Internet. Ujlambkar and Attar [14] analyzed various classification algorithms in order to learn, train and test the model for Indian music classification. Okuyucu et al. [15] performed a feature and classifier analysis for the recognition of similar Environmental sound categories using MFCC along with Zero Crossing Rate as audio feature. Vyas and Dutta [16] used three set of features, namely MFCC, peak difference and frame energy, followed by the K-means algorithm to classify Indian music. Baniya et al. [17] combined the extreme learning machine (ELM) with bagging classifier; a majority score decided the final classification. Baniya et al. [18] uses various audio features of different weights to decide on the final score for genre classification.

52.4 Research Approach and Methodology

In this section, we first present the dataset of song fragments, the features chosen, and the neural network. We used the music dataset from GTZAN Genre Collection. Marsyas (Music Analysis, Retrieval, and Synthesis for Audio Signals) is an open source framework from which audio tracks, each 30 s long. It contains 10 genres, each represented by 100 tracks. The tracks are all 22,050 Hz Mono 16-bit audio files in .wav format. For this project we have chosen only four genre out of 10 genres as related past work has indicated that accuracy decreases when classification categories increases. The chosen genre are jazz, classical, metal and pop. The genre of a song is available under song's properties (Fig. 52.1).

Feature extraction is an attribute reduction process which results in a much smaller and richer set of attributes. We have chosen six features (with 16 values in total) which will be



Property	Value
Description	
Title	Sleep Away
Subtitle	
Rating	☆☆☆☆☆
Comments	Blujazz Productions
Media	
Contributing artists	Bob Acri
Album artist	Bob Acri
Album	Bob Acri
Year	2004
#	3
Genre	Jazz
Length	00:03:20
Audio	
Bit rate	192kbps
Origin	
Publisher	

Fig. 52.1 Genre of a song, stored as a file

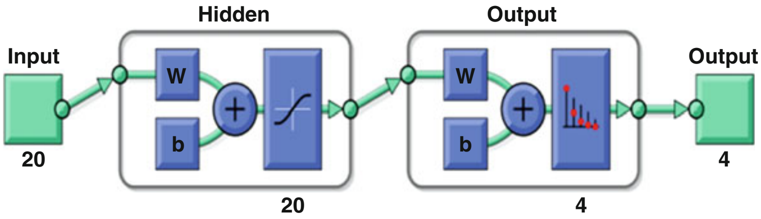


Fig. 52.2 Values of the 17-value features for the first 20 songs

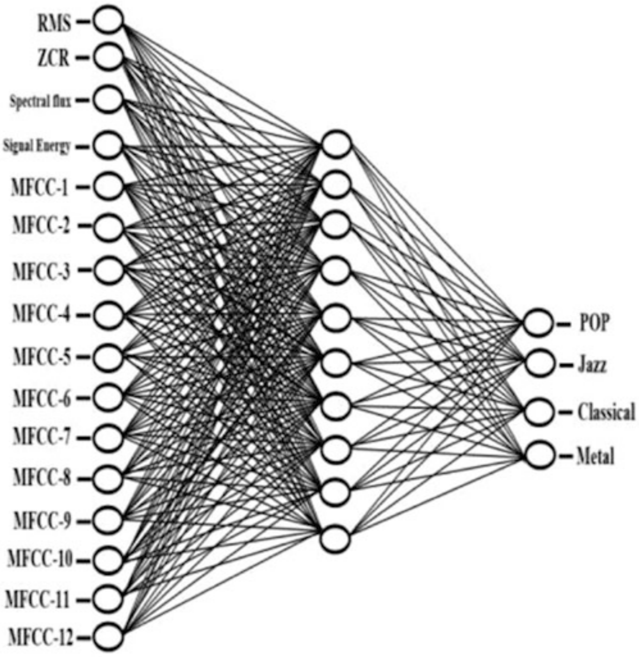


Fig. 52.3 Neural network used for classification of songs

extracted using the back propagation algorithm: Root Mean Square level, Zero Crossing Rate, Signal Energy, Spectral Flux, and Mel Frequency Cepstral Coefficients (12 in total). A snapshot of how the values are computed for the first 20 songs is shown in Fig. 52.2.

The neural network consists of 16 neurons in input layer, 4 neurons in output layer, and 10 neurons in hidden layer (see Fig. 52.3).

The number of neurons in hidden layer is not fixed but it is usually kept as an average of the neurons in input and output layer. We have chosen this network by trials and error, all the other networks gave worse performance in classification.

Since the neural network uses a supervised learning technique, out of 400 data samples, 300 data samples are used for training and validation, and remaining 100 are used for testing. The input to the neural network are the 16 values (from the five features) which are extracted during feature process (Fig. 52.4).

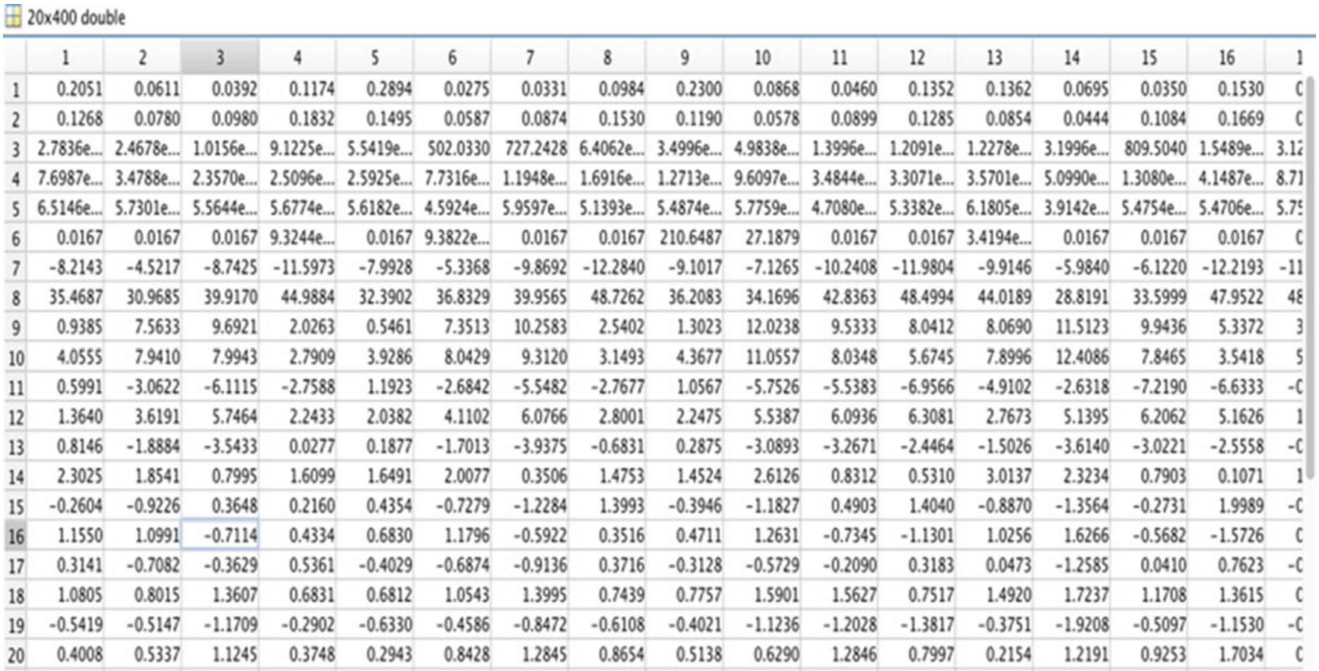
The network will give labels to the output neurons corresponding to a particular genre. The output for the first four songs is shown in Fig. 52.5.

52.5 Experimental Results

All experimental results were gathered in the MATLAB environment using the Signal Processing Toolbox to extract features and Neural Network Toolbox: used for training & classification. The performance of the neural network is shown in the confusion matrix. The confusion matrices produced by MATLAB show two green squares which represent correct classifications and two red squares representing incorrect classifications. Correct classifications on the confusion matrix are represented as true positive and true negative, where true positive refers to correct classifications of class membership and true negative refers to correct classifications of class non-membership. Conversely, the incorrect classifications are represented as false positive and false negative rates. Intuitively, false positives represent incorrect class membership classification and false negatives represent incorrect class non-membership.

The performance percentages are calculated by dividing the total number of correct classifications by the total number of classifications.

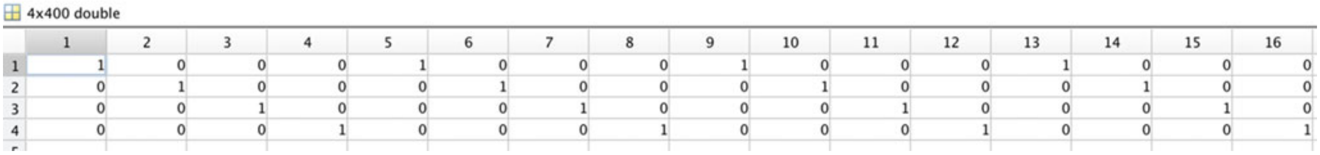
MATLAB also displays multiple instances of confusion matrices of each phase of the neural network: training, validation, and testing. These individual confusion



20x400 double

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	1
1	0.2051	0.0611	0.0392	0.1174	0.2894	0.0275	0.0331	0.0984	0.2300	0.0868	0.0460	0.1352	0.1362	0.0695	0.0350	0.1530	C
2	0.1268	0.0780	0.0980	0.1832	0.1495	0.0587	0.0874	0.1530	0.1190	0.0578	0.0899	0.1285	0.0854	0.0444	0.1084	0.1669	C
3	2.7836e...	2.4678e...	1.0156e...	9.1225e...	5.5419e...	502.0330	727.2428	6.4062e...	3.4996e...	4.9838e...	1.3996e...	1.2091e...	1.2278e...	3.1996e...	809.5040	1.5489e...	3.1
4	7.6987e...	3.4788e...	2.3570e...	2.5096e...	2.5925e...	7.7316e...	1.1948e...	1.6916e...	1.2713e...	9.6097e...	3.4844e...	3.3071e...	3.5701e...	5.0990e...	1.3080e...	4.1487e...	8.71
5	6.5146e...	5.7301e...	5.5644e...	5.6774e...	5.6182e...	4.5924e...	5.9597e...	5.1393e...	5.4874e...	5.7759e...	4.7080e...	5.3382e...	6.1805e...	3.9142e...	5.4754e...	5.4706e...	5.7
6	0.0167	0.0167	0.0167	9.3244e...	0.0167	9.3822e...	0.0167	0.0167	210.6487	27.1879	0.0167	0.0167	3.4194e...	0.0167	0.0167	0.0167	C
7	-8.2143	-4.5217	-8.7425	-11.5973	-7.9928	-5.3368	-9.8692	-12.2840	-9.1017	-7.1265	-10.2408	-11.9804	-9.9146	-5.9840	-6.1220	-12.2193	-11
8	35.4687	30.9685	39.9170	44.9884	32.3902	36.8329	39.9565	48.7262	36.2083	34.1696	42.8363	48.4994	44.0189	28.8191	33.5999	47.9522	48
9	0.9385	7.5633	9.6921	2.0263	0.5461	7.3513	10.2583	2.5402	1.3023	12.0238	9.5333	8.0412	8.0690	11.5123	9.9436	5.3372	3
10	4.0555	7.9410	7.9943	2.7909	3.9286	8.0429	9.3120	3.1493	4.3677	11.0557	8.0348	5.6745	7.8996	12.4086	7.8465	3.5418	5
11	0.5991	-3.0622	-6.1115	-2.7588	1.1923	-2.6842	-5.5482	-2.7677	1.0567	-5.7526	-5.5383	-6.9566	-4.9102	-2.6318	-7.2190	-6.6333	-C
12	1.3640	3.6191	5.7464	2.2433	2.0382	4.1102	6.0766	2.8001	2.2475	5.5387	6.0936	6.3081	2.7673	5.1395	6.2062	5.1626	1
13	0.8146	-1.8884	-3.5433	0.0277	0.1877	-1.7013	-3.9375	-0.6831	0.2875	-3.0893	-3.2671	-2.4464	-1.5026	-3.6140	-3.0221	-2.5558	-C
14	2.3025	1.8541	0.7995	1.6099	1.6491	2.0077	0.3506	1.4753	1.4524	2.6126	0.8312	0.5310	3.0137	2.3234	0.7903	0.1071	1
15	-0.2604	-0.9226	0.3648	0.2160	0.4354	-0.7279	-1.2284	1.3993	-0.3946	-1.1827	0.4903	1.4040	-0.8870	-1.3564	-0.2731	1.9989	-C
16	1.1550	1.0991	-0.7114	0.4334	0.6830	1.1796	-0.5922	0.3516	0.4711	1.2631	-0.7345	-1.1301	1.0256	1.6266	-0.5682	-1.5726	C
17	0.3141	-0.7082	-0.3629	0.5361	-0.4029	-0.6874	-0.9136	0.3716	-0.3128	-0.5729	-0.2090	0.3183	0.0473	-1.2585	0.0410	0.7623	-C
18	1.0805	0.8015	1.3607	0.6831	0.6812	1.0543	1.3995	0.7439	0.7757	1.5901	1.5627	0.7517	1.4920	1.7237	1.1708	1.3615	C
19	-0.5419	-0.5147	-1.1709	-0.2902	-0.6330	-0.4586	-0.8472	-0.6108	-0.4021	-1.1236	-1.2028	-1.3817	-0.3751	-1.9208	-0.5097	-1.1530	-C
20	0.4008	0.5337	1.1245	0.3748	0.2943	0.8428	1.2845	0.8654	0.5138	0.6290	1.2846	0.7997	0.2154	1.2191	0.9253	1.7034	C

Fig. 52.4 The five features with 16 coefficients for genre classification



4x400 double

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
2	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0
3	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0
4	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1

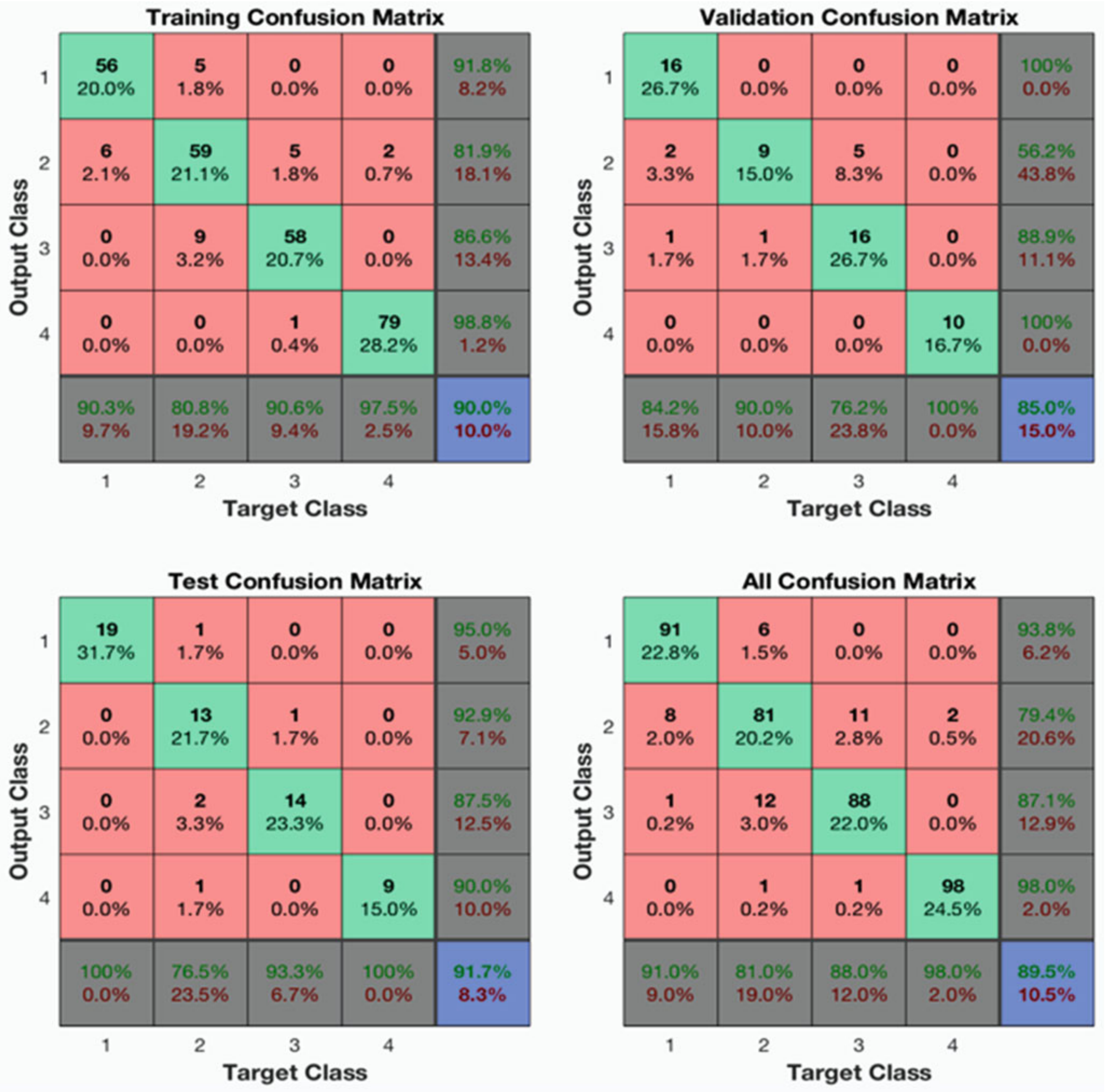
Fig. 52.5 Output of the neural network for the first four songs

matrices offer a better glimpse into the performance of the network and insights onto possible improvements. The confusion matrix of the training sequence usually yields the highest performance rate and is normally regarded as the weakest indicator of true classification performance.

Validation and testing confusion matrices are the best indicators of true classification performance with validation performance usually being regarded as the indicator to be maximized when searching for the optimal number of hidden nodes in a network. The confusion matrix is shown in Fig. 52.6 the green squares represent correct classifications, the red squares represent incorrect classifications, and the blue square at the bottom right edge represents the total performance of the model's accuracy. The peak performance of the 10-hidden node neural network is for pop music at 91.7%, followed by metal at 90% (see Fig. 52.6).

52.6 Conclusions and Future Work

Music genre classification was achieved with 90% accuracy. Classification accuracy for pop (91.7%) and metal (90%) was higher while jazz (85%) and classical (89.5%) was lesser due to similarity in features. The adaptability and versatility of neural networks, along with the strong performance of classifying genre based on short music fragments, show a clear potential for the application of neural networks in automatic genre classification of songs. Addition of spectral features may further improve accuracy.



Training Confusion Matrix

	1	2	3	4	
1	56 20.0%	5 1.8%	0 0.0%	0 0.0%	91.8% 8.2%
2	6 2.1%	59 21.1%	5 1.8%	2 0.7%	81.9% 18.1%
3	0 0.0%	9 3.2%	58 20.7%	0 0.0%	86.6% 13.4%
4	0 0.0%	0 0.0%	1 0.4%	79 28.2%	98.8% 1.2%
	90.3% 9.7%	80.8% 19.2%	90.6% 9.4%	97.5% 2.5%	90.0% 10.0%

Validation Confusion Matrix

	1	2	3	4	
1	16 26.7%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
2	2 3.3%	9 15.0%	5 8.3%	0 0.0%	56.2% 43.8%
3	1 1.7%	1 1.7%	16 26.7%	0 0.0%	88.9% 11.1%
4	0 0.0%	0 0.0%	0 0.0%	10 16.7%	100% 0.0%
	84.2% 15.8%	90.0% 10.0%	76.2% 23.8%	100% 0.0%	85.0% 15.0%

Test Confusion Matrix

	1	2	3	4	
1	19 31.7%	1 1.7%	0 0.0%	0 0.0%	95.0% 5.0%
2	0 0.0%	13 21.7%	1 1.7%	0 0.0%	92.9% 7.1%
3	0 0.0%	2 3.3%	14 23.3%	0 0.0%	87.5% 12.5%
4	0 0.0%	1 1.7%	0 0.0%	9 15.0%	90.0% 10.0%
	100% 0.0%	76.5% 23.5%	93.3% 6.7%	100% 0.0%	91.7% 8.3%

All Confusion Matrix

	1	2	3	4	
1	91 22.8%	6 1.5%	0 0.0%	0 0.0%	93.8% 6.2%
2	8 2.0%	81 20.2%	11 2.8%	2 0.5%	79.4% 20.6%
3	1 0.2%	12 3.0%	88 22.0%	0 0.0%	87.1% 12.9%
4	0 0.0%	1 0.2%	1 0.2%	98 24.5%	98.0% 2.0%
	91.0% 9.0%	81.0% 19.0%	88.0% 12.0%	98.0% 2.0%	89.5% 10.5%

Fig. 52.6 Confusion matrix of the 400 songs (top left is metal, top right is jazz, bottom left is pop, and bottom right is classical)

It will be useful to test our algorithm for another available database provided by the Technical University Dortmund [19].

Acknowledgements Doina Bein acknowledges the support by Air Force Office of Scientific Research under award number FA9550–16–1-0257.

References

1. M.M. Panchwagh, V.D. Katkar, Music genre classification using data mining algorithm, in *2016 Conference on Advances in Signal Processing (CASP)*, Pune, India, 2016
2. S. Haykin, *Neural Networks and Learning Machines* (Pearson Education, Inc., Upper Saddle River, NJ, 2009)

3. M. Copeland, What's the difference between artificial intelligence, machine learning, and deep learning?, https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/. Accessed 22 Nov 2017
4. T. Rashid, *Make Your Own Neural Network: A Gentle Journey Through the Mathematics of Neural Networks, and Making Your Own Using the Python Computer Language* (CreateSpace Independent Publishing, San Bernardino, CA, 2016)
5. C.M. Bishop, *Neural Networks for Pattern Recognition* (Clarendon Press, Oxford, 1995)
6. F. Sebastiani, Machine learning in automated text categorization. ACM Comput. Surv. **34**(1), 1–47 (2002)
7. J.J. Hopfield, Neural network and physical systems with collective computational abilities, in *Proceedings of the National Academy of Science*, 1982
8. H. Chen, L. She, Inductive query by examples (IQBE): A machine learning approach, in *27th Annual Hawaii International Conference on System Sciences (HICSS-27)*, Los Alamitos, 1994
9. M. Gordon, Probabilistic and genetic algorithms for document retrieval. Commun. ACM **31**(10), 1208–1218 (1988)
10. M.D. Gordon, User-based document clustering by redescribing subject descriptions with a genetic algorithm. J. Assoc. Inf. Sci. Technol. **42**(5), 311–322 (1991)
11. V.V. Raghavan, B. Agarwal, Optimal determination of user-oriented clusters: An application for the reproductive plan, in *Proceedings of the Second International Conference on Genetic Algorithms on Genetic Algorithms and Their Application*, Cambridge, Massachusetts, USA, 1987
12. F.E. Petry, B.P. Buckles, D. Prabhu, D.H. Kraft, Fuzzy information retrieval using genetic algorithms and relevance feedback, in *Proceedings of the ASIS Annual Meeting*, Medford, NJ, 1993
13. O. Frieder, H.T. Siegelmann, On the allocation of documents in multiprocessor information retrieval systems, in *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, NY, NY, 1991
14. A.M. Ujlambkar, V.Z. Attar, Automatic mood classification model for Indian popular music, in *Sixth Asia Modeling Symposium*, 2012
15. C. Okuyucu, M. Sert, A. Yazici, Audio feature and classifier analysis for efficient recognition of environmental sounds, in *International Symposium on Multimedia*, 2013
16. G. Vyas, M.K.K. Dutta, Automatic mood detection of Indian music using MFCCs and K-means Algorithm, in *Seventh International Conference on Contemporary Computing (IC3)*, 2014
17. B.K. Baniya, D. Ghimire, J. Lee, A novel approach of automatic music genre classification based on timbral texture and rhythmic content features, in *16th International Conference on Advanced Communication Technology (ICACT)*, 2014
18. B.K. Baniya, J. Lee, Z.-N. Li, Audio feature reduction and analysis for automatic music genre classification, in *International Conference on Systems Man and Cybernetics (SMC)*, 2014