Proceedings of 2018 Eleventh International Conference on Contemporary Computing (IC3), 2-4 August, 2018, Noida, India

A Hybrid Model For Music Genre Classification Using LSTM And SVM

Prasenjeet Fulzele[1], Rajat Singh[2], Naman Kaushik[3]
Dept. of Computer Science,
Jaypee Institute of Information Technology,
Uttar Pradesh, India
{[1]prasenjeet.fulzele7, [2]rajatsinghcms,
[3]naman.k1804}@gmail.com

Kavita Pandey
Dept. of Computer Science,
Jaypee Institute of Information Technology,
Uttar Pradesh, India
kavita.pandey@jiit.ac.in

Abstract— **With today's cutting edge technology and intractable access to voluminous data files via the internet, it is important to meet the computational needs of every user. Machine learning is one such growing branch of artificial intelligence that has made such demands of the users viable. Machine learning models are paving the way for classification techniques such as in music genre classification, and have shown to be efficient in predicting classes to a great extent. To exploit the time dependent nature of the dataset Long Short-Term Memory (LSTM) Neural Network is used for music genre classification and combined with Support Vector Machine (SVM) classifier to enhance its performance. The hybrid model of these two classifiers resulted into an increase in the accuracy of prediction of the individual models. This hybrid model is imposed on GTZAN music dataset and is compared with the results of standalone models of LSTM and SVM. The proposed model exceeded the independent accuracies of the LSTM and SVM classifiers with an accuracy of 89%, reaffirming the efficient utilization of each classifier.**

Keywords—Music Genre Classification; Machine Learning; Long Short-Term Memory Neural Network; Support Vector Machine

I. Introduction

In the wake of an era of digitization where overflowing amounts of information be it textual, graphical, numerical all of which are being promised to be processed in a blink of an eye, automated music genre classification is one such area of research that is yet to deliver . With CDs, DVDs, and cassettes no longer being the only source of music, the advent of internet has a major control over the flow of information, that is, from the web to every other person in the world connected to this network. From broadcasts to podcasts, from sound packages to sound samples everything can be downloaded within a few clicks. The result is a massive collection of songs, audio files, piled up indiscriminately in various folders, making it a gruesome task for a user to keep track of the genre of each and every song and arrange them likewise. Segregating audio tracks by tagging them to appropriate genre happens to be the most logical approach in order to manage audio files present in such huge numbers.

Automated genre-tagging using machine learning based models have opened up new possibilities in this active area of research with promising results. Feature extraction proves to be a crucial method being widely adopted in predicting the actual labels of any audio file [1]. Numerous models such as the deep neural networks (DNNs) have been employed along with combination of other models such decision trees, to predict the genre with greater accuracies [2]. DNNs have also proved to be efficient in handling and training vast amounts of data remarkably well [3-7]. Likewise, other neural networks for example CNN has come very close in terms of correctly predicting the class of a given audio input file by employing the fact of characteristic spectrogram images [8]. It is understood that there exist n-number of genres and sub-genres in music theory but it is difficult to precisely map out the boundaries between each of them [9].Therefore, alternate methods to resolve a huge collection of audio files into their respective genres should be taken up further.

This paper has been precisely divided into five sections. The models used in this study are briefly described in Section II along with their usage in the proposed model. Section III comprises of the implementation of the proposed model including the training method used. In Section IV the performance of the hybrid model is shown using the confusion matrix along with the independent accuracies achieved for each genre. Finally the paper is concluded in Section V.

II. Methodology

The proposed model is implemented in two parts. The training of two individual classifiers- SVM and LSTM Neural Network is accomplished in the first part. Features from the audio files of dataset were extracted and arranged to form the input to be embedded into the two models. The second part involves fusing of the two models by using the sum rule. The separate posterior probabilities of the two models are added up and these combined results are used to make the final prediction.

A. *Training Classifiers*

The dataset to be used for training of the music genre classifier included audio files. To utilize the time dependent property of the data, a network with feedback and storage capabilities was to be used. The Long-Short Term Memory Neural Network is one such network. To improve further over the classification performance of the LSTM network, a strong classifier Support Vector Machine (SVM) was trained.

1) Long Short –Term Memory Neural Network

Neural Networks are information processing systems which ensure parallel processing with the help of

978-1-5386-6835-1/18/$31.00 ©2018 IEEE

interconnected neurons [10]. They have a layered structure and the type of connection between these layers defines the structure of the Neural Network. To handle sequential data the most appropriate type of neural network are the Recurrent Neural Networks (RNN). In order to work with long term time dependencies, another efficient architecture of RNN is Long Short-Term Memory (LSTM) Neural Network [11]. The hidden layers of LSTM have similar structure as that of RNN but they have four interconnected layers in the repeating module instead of one.

2) Support Vector Machine

A strong supervised classifier used for classification of two or more classes is Support Vector Machine which distinguishes between two classes by creating a separation boundary between the two based on the data belonging to each class [12]. It separates the data of two classes by creating a hyper-plane which most accurately classifies the unknown future data. It works towards minimizing the expected error's maximum value.

B. Combining Predictions

The probability with which each sample is predicted as one of the genres is represented in the form of posterior probabilities. The two models are fused together by combining these posterior probabilities from both the models using sum rule. The advantageous behavior of both models is combined and the final predictions are based on the fused probabilities of the proposed hybrid model. The performance of the model is assessed over unknown files from the test data set.

III. IMPLEMENTATION

GTZAN music database is used to train the models and evaluate their performances. This dataset is available on the official website of MARSYAS software framework publically [13]. In the proposed model, SVM and LSTM Neural Network are trained separately. The performance of the model is evaluated by calculating the accuracy of the predictions made on the test set and by plotting the confusion matrix of the combined model. All models were built and coded in MATLAB 2017b on a 64-bit Windows operating system with Intel Core i5 processor and 4GB RAM.

A. Dataset and Features

The GTZAN dataset consists of 1000 music files in .au format for ten genres, with 100 files belonging to each genre. Every audio file is 30 seconds long falling under one of the ten genres: Hip-Hop, Rock, Reggae, Classical, Jazz, Blues, Pop, Disco, Country and Metal.

22050Hz was taken as the sampling frequency, F_s of the audio files for feature extraction. A total of 9 features were extracted from the audio files in order to make the audio files readable by the machine learning models used for training.

For training of the SVM multi-class classifier, all these features were computed to finally form a feature vector of dimensionality 1*333. The input of the LSTM Neural Network is an array of sequences. So all features being time dependent, sequences were combined to form an array of dimensionality 26*13. The final set of features taken is shown in Fig. 1.

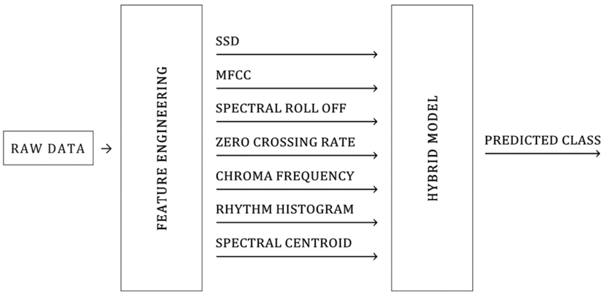

Fig. 1 Features used in proposed model

B. Models

Normalized data was organized and sent as input to the two models. First the models are trained individually and then they are combined to give the final prediction. Random Grid Search was used to tune the hyper-parameters of all the models.

The posterior probabilities of all the files in the test set from both the models were combined using sum rule. The final prediction was made on these resulting probabilities. The structure of the final hybrid model is shown in Fig. 2.

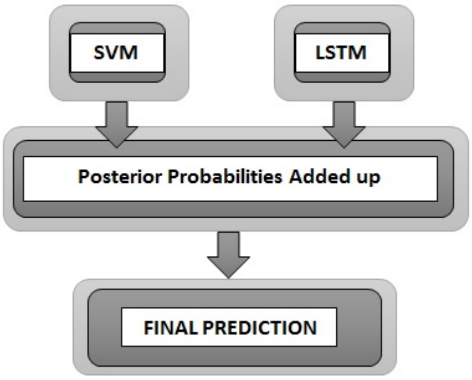

Fig. 2. Proposed Hybrid Model

IV. RESULTS

The dataset contains 1000 audio files from which 900 files were selected for training and rest of the 100 files for testing. Random Grid Search was used for hyper-parameter tuning to optimize the values of hyper-parameters of the models.

Both the individual models were outperformed by the combined hybrid model which gave the maximum accuracy among all the three models. On the same dataset, the accuracy of all models is evaluated and the proposed hybrid model gives the highest accuracy of 89%, as shown in Table I.

TABLE I. ACCURACY COMPARISON OF MODELS

Model	Accuracy (%)
SVM	84
LSTM	69
SVM+LSTM(Proposed)	89

Combining the two models resulted into a major improvement in classification. It enhanced the performance by combining the correct predictions from both models. The individual accuracy of SVM classifier was 84% and that of the LSTM Neural Network was 69%. The performance of the LSTM model was majorly improved by combining it with the SVM classifier. The combined hybrid model resulted in an accuracy of 89%. The performance of SVM model was also improved.

The classification accuracies of individual classes by the three models are depicted in Fig. 3. It can be stated evidently that for music genre classification, the proposed hybrid model is an improvement over the separate individual models for this dataset. To show the performance of the proposed hybrid model confusion matrix is plotted in Fig. 4.

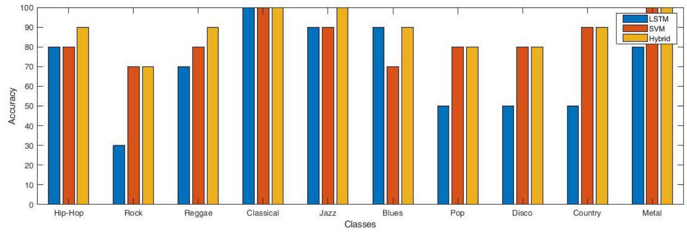

Fig. 3. Accuracy comparison for every genre

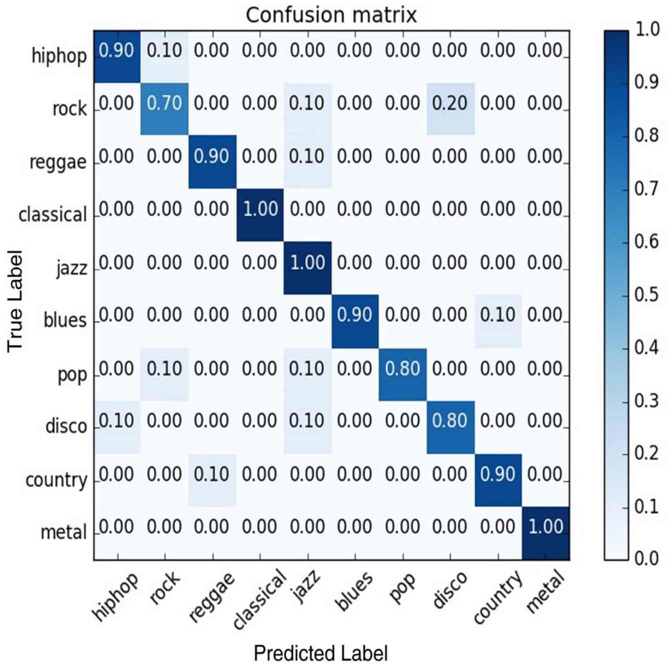

Fig. 4. Confusion Matrix for Proposed Hybrid Model

V. CONCLUSION

The paper presents a hybrid model of two classifiers namely SVM and LSTM and combines the separate results of these classifiers to improve their performances. Consequently, combination of LSTM with SVM also improved their accuracies by fair amount. At last the three different models were evaluated and deployed over the same training dataset GTZAN to avoid biased evaluation. Ten different genres were classified with an accuracy of 89% by the hybrid model of SVM-LSTM which in fact is the model proposed in this study. The proposed model employed the advantages from the two classifiers and combined them in order to classify all the 10 genres with maximum accuracy.

REFERENCES

[1] J. Irvin, E. Chartock, and N. Hollander, "Recurrent Neural Networks with Attention for Genre Classification".

[2] A. Kar, C. Ahuj, A. Mukherjee, "Music Classification using DNN's".

[3] Y. Zhao, D. P. Tao, S. Y. Zhang, and L. W. Jin, "Similar handwritten chinese character recognition based on deep neural networks with big data", Journal on Communications, vol. 35, no. 9, pp. 184-189, 2014.

[4] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 30–42, 2012

[5] Mcloughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust sound event classification using deep neural networks", IEEE/ACM Transactions on Audio Speech and Language Processing, vol. 23, no. 3, pp. 540 – 552, 2015.

[6] I. H. Chung, T. N. Sainath, B. Ramabhadran, M. Picheny, J. Gunnels, V. Austel, U. Chauhari, and B. Kingsbury, "Parallel deep neural network training for big data on bluegene/q," in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 745–753, 2015.

[7] A. R. Rajanna, K. Aryafar, A. Shokoufandeh, R. Ptucha, "Deep Neural Networks: A Case Study for Music Genre Classification", IEEE 14th International Conference on Machine Learning and Applications (ICMLA), pp. 655-660, 2015.

[8] C. Senac, T. Pellegrini, F. Mouret, J. Pinquier , "Music Feature Maps with Convolutional Neural Networks for Music Genre Classification", in Proceedings of CBMI, Florence, Italy, June 19-21, 2017.

[9] C. Xu, N. C. Maddage, X. Shao, F. Cao and Q. Tian , "Musical Genre Classification Using Support Vector Machines", in IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03, pp. 429-432, 2003.

[10] P. Kumar, P. Sharma, "Artificial Neural Networks-A Study", International Journal of Emerging Engineering Research and Technology, Vol. 2, no. 2, pp. 143-148, 2014.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] M. I. Mandel, G. E. Poliner and D. P. W. Ellis, "Support vector machine active learning for music retrieval", Multimedia Systems, vol. 12, no. 1, pp. 3-13, 2006.

[13] Marsyas (Music Analysis, Retrieval and Synthesis of Audio Signals) WebPage.Available:http://marsyasweb.appspot.com/download/data_sets/, accessed 3 October 2017.