Different Machine Learning Classifiers for Music Emotion Recognition

Rahul Suresh, Soumya A

Abstract*: *Music in an essential part of life and the emotion carried by it is key to its perception and usage. Music Emotion Recognition (MER) is the task of identifying the emotion in musical tracks and classifying them accordingly. The objective of this research paper is to check the effectiveness of popular machine learning classifiers like XGboost, Random Forest, Decision Trees, Support Vector Machine (SVM), K-Nearest-Neighbour (KNN) and Gaussian Naive Bayes on the task of MER. Using the MIREX-like dataset [17] to test these classifiers, the effects of oversampling algorithms like Synthetic Minority Oversampling Technique (SMOTE) [22] and Random Oversampling (ROS) were also verified. In all, the Gaussian Naive Bayes classifier gave the maximum accuracy of 40.33%. The other classifiers gave accuracies in between 20.44% and 38.67%. Thus, a limit on the classification accuracy has been reached using these classifiers and also using traditional musical or statistical metrics derived from the music as input features. In view of this, deep learning-based approaches using Convolutional Neural Networks (CNNs) [13] and spectrograms of the music clips for MER is a promising alternative.

Keywords*: *Decision Tree classifier, Gaussian Naive Bayes, K-Nearest-Neighbour (KNN), Music Classification, Music Emotion Recognition (MER), MIREX-like dataset, Random Forest classifier, Support Vector Machine (SVM), XGboost

I. INTRODUCTION

Music is a fundamental part of our lives and people listen to more music today than any other point of time in the past. Therefore, to better serve this need, identifying the emotion carried by musical tracks and categorizing them accordingly is essential. This task, formally known as Music Emotion Recognition (MER) has been a subject of intense research in recent times [4][10][13]. Competitions like the annual Music Information Retrieval Evaluation eXchange (MIREX) challenge are based on this exact task.

This paper aims to compare the accuracy of popular machine learning based classification algorithms on MER. Such algorithms include the XGboost, Random Forest, Decision Trees, SVM, KNN and Gaussian Naive Bayes. These classification algorithms have gained widespread use [23][24] in recent times due to their easy training and good performance. Normally musical or statistical features are extracted from the music clip [4][10] and used as input. However, in recent times, spectrograms of the music clip [13] are also used as input. The goal in both cases is to feed as much information about the music clip as possible to the classification model for it to make the best prediction.

The remainder of the paper is organized as follows. The latest advances in MER are covered in the Literature Survey. In the Theory section, the choice of classification labels and dataset is explained along with a note on the classification algorithms and oversampling techniques used in classifier models built as part of this paper. In the Experimental section, the composition of the MIREX-like dataset is explained along with the training and testing procedures that were used to build and evaluate the classification models. The process of hyper-parameter optimization and the resulting hyper-parameters are also explicated here. The Performance Analysis and discussion section follows, in which the accuracy, recall and F1 score achieved by each model is presented and the observations are noted. In the Conclusion section the important takeaways of this paper are presented followed by directions for future work.

II. LITERATURE SURVEY

The task of MER can be divided as static or dynamic emotion recognition. In static music recognition, a single emotion is identified for the entire music clip. The other is dynamic emotion recognition, in which the changing emotions over the duration of the music clip are identified. Additionally, two types of emotion-based classification can be performed. One is categorical in which music is classified according to emotion categories. Example happy, sad, etc. The other type is the dimensional approach according to which each music clip is allotted a point in a 2 or 3-dimensional emotion space. Russell [1] proposed a 2D emotion model, where the dimensions are arousal and valence thus creating an Arousal-Valence (A-V) plane. Arousal quantifies how exciting or calming the clip is. Valence measures positive or negative affectivity. It defines a continuous scale from pleasantness to unpleasantness [2]. Mehrabian [3] suggested a 3D model of Pleasure-Arousal-Dominance (PAD) in which Pleasure is the same as Valence, Arousal remains the same and Dominance reflects control or lack of control as represented by the clip.

In the effort by Jacek [4], a sliding window was used to perform dynamic music emotion recognition. The dataset used consisted of 324 clips each of 6 seconds belonging to different genres of music. Annotation of each of the clips was done by 5 experts who marked the valence and arousal of each clip.

Revised Manuscript Received on November 15, 2019

***** Correspondence Author

Rahul Suresh, Master of Science in Artificial Intelligence, Boston University.
Email: sureshmotortech@gmail.com

Soumya A*, Department of Computer Science and Engineering, RV College of Engineering, Bangalore, and Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India.
Email: soumyaa@rvce.edu.in

The data collected was then averaged to draw up a point on the A-V plane. The independent variables (features) were extracted from the 6 second clips using the Essentia [5] and Marsyas [6] libraries. Using Essentia the spectral, tonal, rhythmic and time-domain features were obtained and then its mean, geometric mean, power mean, etc were also extracted to get a total of 530 features. Using Marsyas, the features extracted were Spectral Centroid, Zero Crossings, etc to get 31 features in all. For each of these features, 4 statistic features (the mean of the mean, the standard deviation of the standard deviation, the standard deviation of the mean and the mean of the standard deviation) were also calculated to get a total of 124 features. The regressors used for the model were Support Vector Regressor (SVR), Reduced Error Pruning Tree (REPT) and the M5P [7][8] all implemented in WEKA [9]. The best correlation metrics were 0.88 for the Arousal and 0.74 for Valence and was obtained using the SVR after feature selection on the Essentia features.

A categorical classification [10] was performed across 4 categories whose labels were fear, happy, relax and sad. Using the APM [11] database, 100 audio tracks under each label was obtained. Each individual track was of 35 seconds in length. The two audio channels signal are processed and standard audio features like the Root Mean Square (RMS) energy, Zero crossing rate, Voiceprob, etc were extracted using the OpenSMILE [11] library. Additionally, Electroencephalogram (EEG) audio features are also extracted. In all, 384 features are used. A Random Forest classifier with the hyperparameters of the number of decision trees as 800 and number of features as 200 was utilized. The overall accuracy using all the 2 channel and EEG features was 83.29%.

Another approach was based on the belief that by extracting musical or statistical features from music, as is the general technique, there is no way to ascertain which features are the important ones and contribute mainly to the emotion of the music. Therefore, a CNN [13] was directly used on the spectrogram of the audio file with the belief that the neural network will itself pick the important features from the spectrogram. The 2D A-V space was divided into 4 zones. The space from -45° to 45 ° was pleased or happy, 45° to 135° was aroused or alarmed, 135° to 225° was miserable or sad and 225° to 315° was soothing or tired. The 1000 song dataset [14] was used which was actually 744 songs after removing the duplicates. In the 1000 songs dataset each clip was 45 seconds long and annotated with an A-V value pair. 45 seconds was however too large to create a proper spectrogram; therefore, the clip was divided into multiple 5 second clips with the A-V value of the 45 second clip being replicated that many times. Then, the A-V value pair is mapped to one of the four zone of the A-V plane as described earlier. A grayscale Spectrogram was generated using Short-Time Fourier transform on audio signal. In total 30498 spectrograms were generated and used. The input shape of the spectrogram to the CNN was 128 * 128 * 1. The design of the CNN was c1-p1-c2-p2-c3-p3-f1-f2 where c is convolutional, p is pooling and f is the fully connected layer. A Rectified Linear Unit (RELU) nonlinearity was used. Before the final soft-max output layer, dropout was added to prevent overfitting. Using the chosen hyper-parameters, the model was trained and evaluated with 10-fold cross validation. This model gave an average accuracy of 72.4%. Other innovations like Moodnet [15] make use of multimodal data – both Mel spectrograms and the lyrics. This however would be hard if one seeks to make models which are language independent.

III. THEORY

Given the task of classifying music according to its emotion, various choices were possible. Giving independent emotion labels was an option or using the big 6 emotions [16] - happiness, sadness, anger, fear, surprise and disgust was also possible. However, a previously noted problem in the field of MER was the use of independent classification labels that often led to results that were not directly comparable. Hence the MIREX labels, born from the annual MIREX mood classification competition, which are popularly used [17][18] was chosen. The MIREX labels are composed of 5 clusters as shown in Table I.

Table I: MIREX Emotion Clusters

Cluster No.	MIREX labels
Cluster 1	Passionate, rousing, confident, boisterous, rowdy
Cluster 2	Rollicking, cheerful, fun, sweet, amiable, good natured
Cluster 3	Literate, poignant, wistful, bittersweet, autumnal, brooding
Cluster 4	Humourous, silly, campy, quirky, whimsical, witty, wry
Cluster 5	Aggressive, fiery, tense/anxious, intense, volatile, visceral

A similar conundrum was faced while choosing the dataset to use. While using independent dataset composed of arbitrarily chosen songs was an option, the need to maintain a community standard shone through again. Different open source music archives such as APM dataset [11], Free Music Archive [19], Million songs dataset [20]. Most available datasets have music classified only based on genre. Another popular music dataset is the 1000 Songs Database [14] but the issue was that the label was dimensional and not categorical. Each music track in it had a value of arousal and a value of valence and thus translated to a point on the A-V plane instead of an emotion category. While attempts have been made to translate A-V dimensional values to categories [13], this again is subjective and not standardized. Not many big standard datasets exist because of copyright on musical records and in fact that is the reason why the original MIREX dataset itself is not publicly available. To keep with standards the MIREX like dataset [17] which follows MIREX labels was chosen and used.

Models applied for MER were Decision Tree classifier, KNN classifier, Gaussian Naive Bayes classifier, Random Forest classifier, SVM classifier and XGboost classifier. Oversampling methods such as SMOTE [22] and ROS were also tried out. For the classifiers all relevant hyperparameters were optimized using a combination of RandomizedSearchCV and GridSearchCV in Sklearn[21].

Dummy classifiers with different prediction strategies such as most frequent, stratified, etc were used to set the baseline metrics.

IV. EXPERIMENTAL RESULTS

The MIREX like dataset [17] was composed of 5 cluster labels and 903 music clips each of 30 seconds duration. The clips were distributed across the clusters as 18.8% for cluster 1, 18.2% for cluster 2, 23.8% for cluster 3, 21.2% for cluster 4 and 18.1% for cluster 5. This distribution of the dataset is shown in Table II. The Librosa [25] library was used for the feature extraction. Lyrics are not used to keep the model extendable for songs of all languages. For each clip, the features extracted were the Tempo, Beats, Chroma Short-Time Fourier Transform (Chroma_stft), Chroma Constant-Q chromagram (Chroma_cqt), Chroma Energy Normalized (Chroma_cens), Melspectrogram, Root Mean Square Energy (RMSE), Spectral Centroid, Spectral Bandwidth, Spectral Contrast, Spectral Rolloff, Polynomial features (Coefficients of fitting an nth-order polynomial to the columns of a spectrogram), Tonnetz, Zero Crossing Rate (ZCR), Mel-Frequency Cepstral Coefficients (MFCC) and Delta (local estimate of the derivative of the input data along the selected axis). The harmonic and percussive elements were also extracted to get 2 more features. Additionally, the onsets were detected by picking crests in an onset strength envelope to identify frames in the clip. After this, the time (in seconds) of each frame is extracted to get another feature called simply as the frame feature. Three statistical functions, namely the mean, standard deviation and variance are applied on each of the above features, except for the tempo and beats. Therefore 17 x 3 or 51 features are obtained. Together with the tempo, average beats and total beats a total of 54 features describing each clip of the dataset are obtained.

The dataset was initially divided into the training dataset 'A' and testing dataset 'B' after which the testing dataset 'B' was kept aside and not seen by the model until calculating the final metrics. An 80%-20% split was used for creating the training and testing datasets as shown in Table III for training dataset 'A' and Table IV for testing dataset 'B'. The split of the MIREX like dataset into 'A' and 'B' datasets was done using a random state value of 25. A combination of RandomizedSearchCV and GridSearchCV with 10-fold cross validation was used to determine the best hyper-parameters. Initially with a large range of parameters RandomizedSearchCV is used and then to fine tune the parameters further GridSearchCV is used. As the definition of 10-fold cross validation, 9 training splits and 1 testing split are created in each of the 10 runs. The chosen oversampling algorithm such as SMOTE or Random Oversampling was applied on the 9 training splits created for each run. In other words, the oversampling was done after creating the splits and on the 9 training splits and the model parameters was evaluated on the left out 1 testing split. Therefore, for each classifier 3 separate models were built using SMOTE, ROS or no oversampling for the oversampling step. The results of hyperparameter tuning for each model is as shown in Table V. Finally, after the best parameters were found out, the model was trained using these best parameters on the entire 'A' dataset and results were evaluated on the 'B' dataset. For all the classifiers to generate reproducible results a random state value of 25 was used.

Table II: MIREX like Dataset Distribution

Cluster	Number of samples
1	170
2	164
3	215
4	191
5	163
Total	903

Table III: Training Dataset 'A' Distribution

Cluster	Number of samples
1	133
2	130
3	171
4	159
5	129
Total	722

Table IV: Testing Dataset 'B' Distribution

Cluster	Number of samples
1	37
2	34
3	44
4	32
5	34
Total	181

Table V: Results of Hyper-parameter Tuning

Classifier	Hyperparameters	Oversampling Technique
Decision Tree classifier	max_depth: 6 max_features : 21 min_samples_split: 0.2	ROS
KNN classifier	leaf_size : 10 n_neighbours : 8 p :1 weights : distance	SMOTE
Gaussian Naive Bayes classifier	var_smoothing : 1e-21	None

Random Forest classifier	max_depth : 3 max_features : 18 min_samples_leaf : 0.1 min_samples_split : 0.1 n_estimators : 150	None
SVM classifier	c : 1e8 gamma : 1e-15	SMOTE
XGboost classifier	gamma : 0.03 learning_rate : 0.2 max_depth : 13 min_child_weight: 3 n_estimators: 90 reg_lambda : 0.5	None

V. Performance Analysis and Discussion

The baseline metrics set according to different metrics such as most frequent, stratified and uniform are as shown in Table VI. The classifiers were trained on the 'A' dataset and tested on the 'B' dataset.

Table VI: Baseline Classifier Results

Classifier	Accuracy (in %)	Recall (micro avg)	F1 (micro avg)
Most Frequent	24.3	0.24	0.24
Stratified	23.2	0.23	0.23
Uniform	19.88	0.20	0.20

The Gaussian Naive Bayes classifier gave the highest accuracy around 40.33 % and also the highest F1 score of 0.40. The XGboost classifier came second with an accuracy of 38.67 % and an F1 score of 0.39. The complete results are as shown in Table VII.

Table VII: Experimental Results using Different Classifiers

Classifier	Accuracy (in %)	Recall (micro avg)	F1 (micro avg)
Decision Tree Classifier	36.46	0.36	0.36
KNN Classifier	20.44	0.20	0.20
Gaussian Naive Bayes Classifier	**40.33**	**0.40**	**0.40**
Random Forest Classifier	35.35	0.35	0.35
SVM Classifier	32.59	0.33	0.33
XGboost Classifier	38.67	0.39	0.39

The results showcased the tough nature of the MER task and to an extent the inability of only musical or statistical features to provide enough grounds for differentiation among the clusters. A SVM with RBF kernel achieved only 32.59% accuracy. The XGboost classifier too maxed out at 38.67% accuracy and is further evidence to this. The results of oversampling too were varied. While the synthetically generated data points of SMOTE were useful for some classification algorithms such as KNN or SVM, the decision tree classifier performed best with randomly oversampled data points of ROS. The algorithms such as Gaussian Naive Bayes, Random Forest and the XGboost performed best with no oversampling. This was expected for the Gaussian Naive Bayes algorithm as it works by using the underlying probability distributions across the classes to make predictions.

VI. Conclusion

The aim of comparing different classification algorithms for the task of MER was realized. Results showed that the maximum possible accuracy for this task is 40.33% using Gaussian Naive Bayes. The results were a bit surprising as one expected better performance of Random Forest or XGboost algorithms. Even the performance of SVM with RBF kernel at only 32.59% accuracy is a bit unexpected. The effects of oversampling too were non-uniform with algorithms preferring SMOTE or ROS or no oversampling at all. Using only musical information a peak has been reached. Therefore, a different approach such as using spectrograms and CNNs trained on them can be performed in the future.

REFERENCES

1. J. Russell, "A circumplex model of affect", *Journal of personality and social psychology*, 1980, Volume 39, pp. 1161-1178
2. Bradley MM, Lang PJ, "Affective reactions to acoustic stimuli", *Psychophysiology*, 2000, Volume 37, pp. 204–215
3. Albert Mehrabian, "Some referents and measures of nonverbal behavior", *Behavior Research Methods & Instrumentation*, 1968, Volume 1, Issue 6, pp 203–207
4. Jacek Grekow, "Music Emotion Maps in Arousal-Valence Space", 15th IFIP *International Conference on Computer Information Systems and Industrial Management (CISIM)*, Vilnius, Lithuania, 2016, pp. 697-706
5. Bogdanov D, Wack N Gomez, E Gulati S, Herrera P, Mayor O, Roma G, Salamon J, Zapata J, Serra X, "ESSENTIA: an audio analysis library for music information retrieval", *Proceedings of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013, pp. 493–498
6. Tzanetakis G, Cook P, "Marsyas: a framework for audio analysis", *Organised Sound*, 1999, Volume 4, Issue 3, pp. 169 - 175
7. Ross Quinlan, "Learning with Continuous Classes", *Australian Joint Conference on Artificial Intelligence*, Hobart, Australia, 1992, pp. 343–348
8. Wang Y, Witten I.H, "Induction of model trees for predicting continuous classes", *Proceedings of the poster papers of the European Conference on Machine Learning*, Prague, Czechia, 1997
9. I Witten, E Frank, M Hall, "Data Mining: Practical Machine Learning Tools and Techniques", *Morgan Kaufmann Series in Data Management Systems*, 2011
10. Zhang F, Meng H, Li M, "Emotion Extraction and Recognition from Music", *12th International Conference on Fuzzy Systems and Knowledge Discovery*, Zhangjiajie, China, 2016, pp. 1728–1733
11. APM Music, http://www.apmmusic.com/about.
12. F. Eyben, M. Woellmer, and B. Schuller,"OpenSMILE, the Munich open Speech and Music Interpretation by Large Space Extraction toolkit", *ACM SIGMultimedia Records*, Volume 6, Issue 4, 2014, pp. 4-13
13. Liu Tong, Han Li, Ma Liangkai, Guo Dongwei, "Audio-based Deep Music Emotion Recognition", *AIP Conference Proceedings*, Volume 1967, 2018, pp. 040021
14. Mohammad Soleymani, Michael N. Caro, Erik M. Schmidt, Cheng-Ya Sha, Yi-Hsuan Yang, "1000 Songs for Emotional Analysis of Music", *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, Barcelona, Spain, 2013, pp. 1-6
15. Aniruddha Bhattacharya, Kadambari K.V, "A multimodal approach towards emotion recognition of music using audio and lyrical content", arXiv:1811.05760, 2018
16. Ekman P, Sorenson E.R, Friesen W.V, "Pan-cultural elements in facial displays of emotions", *Science*, Volume 164, Issue 3875, 1969, pp. 86–88

17. Panda Renato, Malheiro Ricardo, Rocha Bruno, Oliveira António, Paiva Rui Pedro, "Multi-Modal Emotion Music Recognition (MER): A New Dataset, Methodology and Comparative Analysis", *10th International Symposium on Computer Music Multidisciplinary Research, Marseille, France, 2013*, pp. 570-582
18. Napiorkowski S, "Music mood recognition: State of the Art Review", *MUS-15*, Volume 10, 2015
19. Free Music Archive, http://freemusicarchive.org/about
20. Thierry Bertin-Mahieux and Daniel P. W. Ellis and Brian Whitman and Paul Lamere, "The Million Song Dataset", *ISMIR*, 2011
21. Pedregosa F, et al, "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, Volume 12, 2011, pp. 2825-2830
22. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer, "SMOTE: synthetic minority over-sampling technique", *Journal of Artificial Intelligence* Research archive, Volume 16, Issue 1, 2002, pp. 321-357
23. B. Pan, "Application of XGBoost algorithm in hourly PM2.5 concentration prediction", *Conference Series: Earth and Environmental Science*, Volume 113, 2018, p. 012127
24. Primartha R, Tama BA, "Anomaly detection using random forest: A performance revisited", 2017 *International Conference on Data and Software Engineering (ICoDSE)*, 2017, pp. 1-6
25. McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, Oriol Nieto, "librosa: Audio and music signal analysis in python", *In Proceedings of the 14th python in science conference*, Austin, United States of America, 2015, pp. 18-25

AUTHORS PROFILE

Mr. Rahul Suresh is a graduate student at Boston University pursuing his Master of Science in Artificial Intelligence. He completed his Bachelor of Engineering in Computer Science and Engineering, RV College of Engineering, Bengaluru, securing 9th rank. He is interested in innovative applications of Machine Learning. His works are showcased at *github.com/R-Suresh.*

Dr. Soumya A, Associate Professor, Department of Computer Science and Engineering, RV College of Engineering, Bangalore. She has been awarded Doctoral degree, University of Mysore, Mysore. She has obtained M.S degree in Computer Cognition Technology, University of Mysore, Mysore and B.E in Computer Science and Engineering, Bangalore University, Bangalore. Her areas of research are Artificial Intelligence, Soft Computing, Pattern Recognition and Image Processing. She has to her credits 25 publications in International / National Journals and Conferences. She is currently guiding two research scholars towards PhD, several under-graduate projects and post-graduate projects. She is a member of several professional bodies namely IEEE, ISTE, CSI and IUPRAI