Analysing Speech for Emotion Identification using Catboost Algorithm in Machine Learning

Saranya CP[1, 3], Amal Mohan N[2, 4], Lijo Tony A R[2, 5], Naveen Chanth R[2, 6] and Vishnu K [2, 7]

[1]*Assistant Professor, Department of CSE, Coimbatore Institute of Engineering and Technology, Coimbatore*

[2]*Student, Department of CSE, Coimbatore Institute of Engineering and Technology, Coimbatore*

[3] sarancp@gmail.com, [4] amalmohan414@gmail.com, [5] lijo.tony@gmail.com,

[6] naveenchanth5@gmail.com, [7] vishhnukr@gmail.com

Abstract:

Speech Emotion Recognition (SER) has become famous nowadays. This is because a better accurate SER strategy may contribute to further customization of existing applications like voice assistant, medical diagnosis etc.,. Evolution of artificial intelligence, especially machine learning algorithms have contributed much for the improvement of SER. Upon those, algorithms like neural networks, deep learning etc., provided better accuracy than the other algorithms. This paper provides an attempt to implement Catboost algorithm to perform SER. The implemented algorithm took MFCC as the base feature and trained using the RAVDESS dataset. The algorithm's accuracy was tested for the emotions like neutral, happy, sad, angry, fear and produced an accuracy of 74.07% which is better than other classifiers.

Keywords: **Speech Emotion Recognition (SER), Catboost Algorithm, Machine Learning, Artificial Intelligence, MFCC, RAVDESS**

1. Introduction

During classic days, text was the only form of data used in computer systems. In course of time, audio (speech) became as a valuable input format. Various algorithms, applications were developed to recognize speech and convert them into mere text. Advancement in technology paved the path to extract emotion from the speech file. Hence the Speech Emotion Recognition (SER) has become an interesting and challenging research area to work on. Let us discuss some of the research works.

In this work [1], the researchers have tried several deep neural network architectures for emotion classification. Along with those architectures, they have also tried with the combination of text format of the specific audio file. They concluded that the text along with Mel frequency cepstral coefficient (MFCC) in a convolutional neural network has produced an increased accuracy than the other methods.

In this work [2], the researchers have extracted about 3800 features from the audio file. Among those features, the significant features were extracted using the Waikato Environment for Knowledge Analysis (WEKA) software. They have tried various classifiers to detect the emotion recognition. Upon those, they have found that after normalization and discretization of the input parameters, support vector machine trained using the sequential minimal optimization has produced a better performance.

Here [3], initially the researchers took the features like pitch, MFCC, mel-band energies, etc., as the base features and found that pitch and energy are the most significant factors. Using those, they have tried various classifiers like support vector machine, linear discriminant analysis, hidden Markov model and quadratic discriminant analysis. They summarized that Gaussian Support Vector Machine has performed well for the 4-class speaking style classification.

Inception net v3 is used here [4] to build the emotion recognition model. Since Inception net is developed by Google for Image classification, they have used transfer learning to reduce the computation cost. The accuracy of the model was 35%.

This work [5] concentrates on application perspective of the emotion recognition. It explains various features of a speech. In this work, deep neural network along with the k-means algorithm was used to detect the emotions. They have also suggested an application to detect the scary emotion from speech of a patient and notify an alert via cloud.

The authors in this work [6] have segmented the audio data by resampling and windowing techniques. Such segments are decomposed using discrete wavelet transform and the effective characteristics of the signals are investigated. Such features are used to build a model using artificial neural network.

In our paper, we have chosen to use MFCC as the base feature to extract emotional characteristics of an audio file. This because, this paper [7] discusses deeply about the usage of MFCC to recognize speech from the audio file.

Emotion detection is applicable in various fields. One among them is in medical field. This work [8] shows that there is a relationship between the heart rate from ECG and the features extracted from MFCC are in close relation. Also they have suggested a medical application to detect heart rate changes from the emotions detected from the audio file and to suspect the heart related problems earlier.

Our paper concentrates on the basic knowledge required to perform the speech emotion recognition in section II. Our proposed idea is to use the Catboost Algorithm to perform SER which will be covered in section III. The proposed algorithm will be implemented on the RAVDESS dataset in section IV. The result and performance of the proposed work will be discussed in section V and conclusion is included as section VI.

2. Basic information

2.1. Mel Frequency Cepstral Coefficient (MFCC)

Every sound that is produced by the human is based on each individual's tongue, teeth and vocal tract shape. Normal time-power spectrum of such sounds can be covered with an envelope to represent the vocal tract. The coefficients that construct the Mel-frequency cepstrum (MFCC) perfectly represents such an envelope.

To be more precise, MFCC [7] features project the unique features of a speech because they use to predict the actual vocal tract of the human being. We have used those coefficients in our paper to predict the distinct speech/emotion characteristics.

2.2. RAVDESS Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song which is known as the RAVDESS is the data set used in our work. The given dataset was contributed by 12 male and 12 female characters. Each character had uttered the allocated texts in 8 different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised) in varying intensities. It includes speech, song files in both audio and video format.

Though there are 7365 files in the dataset, we only used speech files with five emotions (neutral, happy, sad, angry, fear) to train and test our machine learning model. All the audio speech files were preprocessed to a mono wav format and a frequency of 16,000 Hz.

3. Proposed Idea

3.1. Catboost Algorithm

Ensemble algorithms in machine learning are famous for their increased proficiency and accuracy. Such ensemble algorithms can be classified into two major types. Bagging is the type in which parallel execution of weak learners occurs. Boosting technique is a type of ensemble learning in which weak learners are trained and combined sequentially to produce a strong learner. Boosting algorithms have impressed a lot in the recent years of machine

learning. Adaboost, Gradient boosting, XGBoost, LightGBM are some of the famous boosting algorithms. In that list, Catboost was released as an open source machine learning library by Yandex in 2017.

Catboost [9] stands for Categorical Boosting. The name reason for this algorithm is it can handle the various categories of data in the form as they appear using the process of one-hot encoding. Since our dataset is not in text or table format and falls under other category as an audio type, Catboost can handle such category with ease than any other existing algorithms.

Catboost algorithm also depends on the ordered boosting technique. Ordered boosting technique is a modified form of gradient boosting algorithm. Gradient boosting algorithm was supposed to have the problem of prediction shifting. Prediction shifting happens due to special kind of target leakage. Ordered boosting was implemented in a way such that it diminishes the problem of prediction shifting. The below figure 1 describes the working of ordered boosting in catboost [9] algorithm.

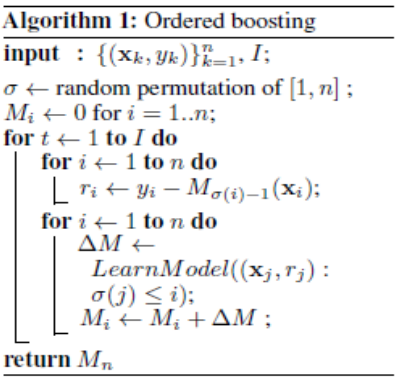

Figure 1. Ordered Boosting Algorithm

Hence Catboost was proposed on the basis of the two primary components – processing categorical features and the ordered boosting technique. Catboost makes use of the gradient boosted decision trees. Numerous decision trees will be built during the training. Decision trees formed in the successive iteration will be having a minimized loss than the trees in previous iterations. In that way, the iteration will be continued until there is no significant reduction in the loss function. The working order of the catboost algorithm is as in the figure 2.

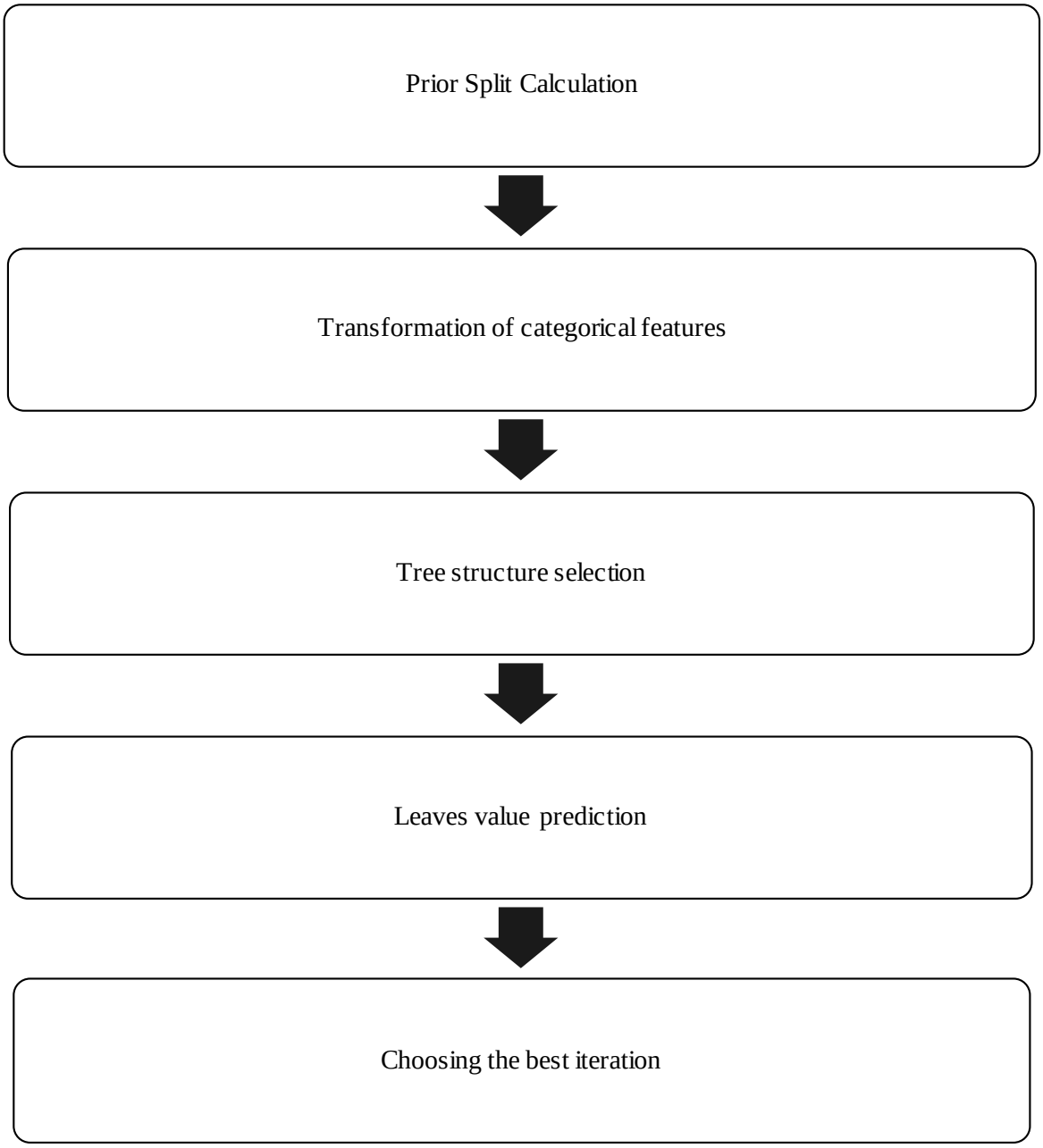

Figure 2. Catboost Algorithm Work Flow

In the prior split calculation step, quantization will be performed to form the buckets from data. Transformation of categorical features will be performed using permutation on the buckets randomly and picking up the acceptable integer values for the labels. Tree structure selection acts in a greedy way such that the features which are capable of splitting the tree are given higher priority. Then the decision trees will be formed along with the leave values formation. Such decision trees will be modified in further iterations in order to reduce the loss value.

4. Implementation

4.1. Training the model

Speech files of the RAVDESS dataset were used to train the model. Such files were regularized to mono stream and a frequency of 16,000 Hz. Neutral, happy, sad, angry and fearful were the five emotions used to train the model. 'Librosa', 'Catboost' and 'Sklearn' libraries in python were used to train and test our machine learning model. MFCC coefficients were extracted from the audio files. Since our problem statement falls under multiple classification, multiclass parameter was chosen for our classifier. An optimum learning rate was found to be 0.475 for the classifier. In order to reduce the over fitting problem, early stopping of the iterations was set to 300 iterations. 75% of the speech files was used to train the model. The remaining 25% of the speech files was used for the testing purpose. The complete process is shown in the figure 3.

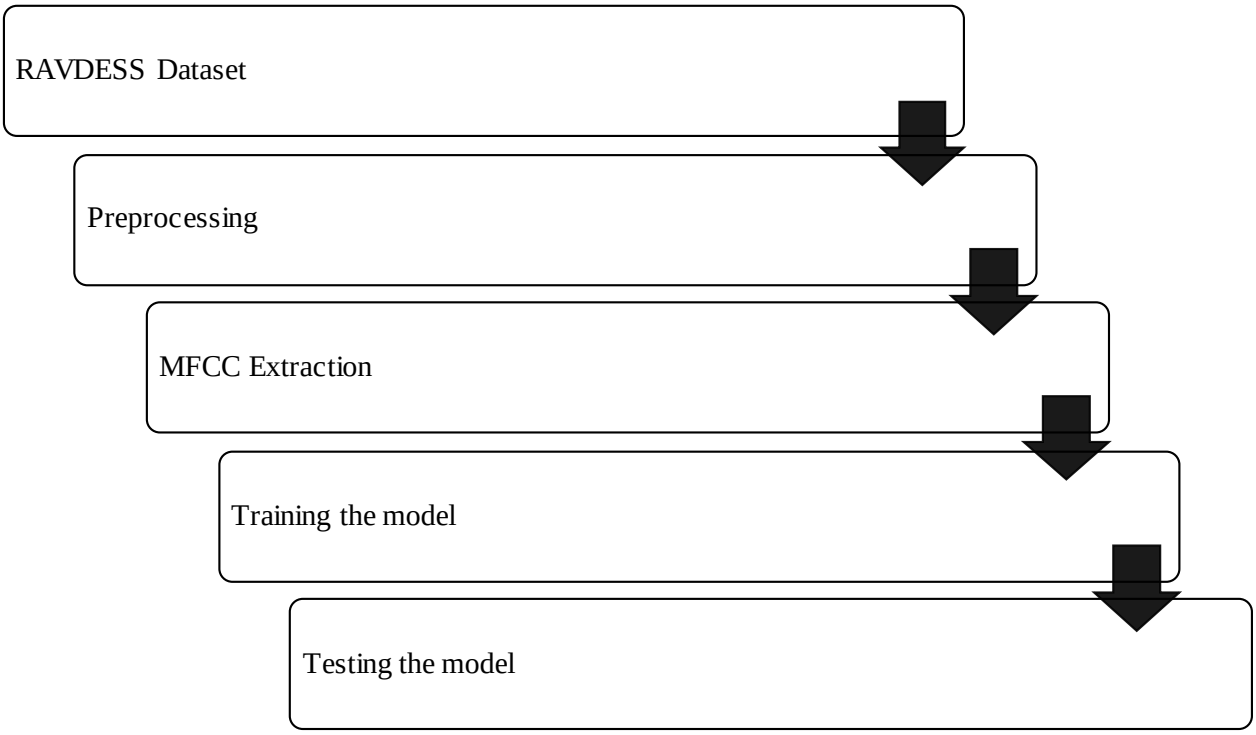

Figure 3. Implementation flow diagram

4.2. Testing the model against custom input

Our model was also tested against the custom inputs. For that the input audio file is regularized to mono wave format and a frequency of 16,000 Hz. Such a preprocessed audio file was given as the input to our model.

5. Result and Analysis

Sklearn's accuracy predictor helped us to evaluate the model's performance. Our model produced an accuracy of 74.07%. We also tried to predict the emotions on the same preprocessed dataset using different classifiers and their respective accuracies are shown in the table 1 and figure 4.When compared to other primitive classifiers like Naive Bayes, KNeighbours etc., our model produced an improved accuracy. This accuracy is more or less equal to the accuracy scores offered by the classifiers of neural network, deep learning, XGBoost. The main advantage of our model is it can handle the categorical data pretty much good than other classifiers.

Table 1. Various Classifiers and their Accuracies

Classifier	Accuracy
Neural Network	73.81%
Gaussian Naive Bayes Classifier	31.65%
KNeighbors Classifier	42.50%
Support Vector Machine Classifier	53.06%
Stochastic Gradient Descent Classifier	35.00%
Gradient Boosting Classifier	52.78%
Light Gradient Boosting Classifier	63.89%
XGB Classifier	63.33%
Catboost Classifier	74.07%

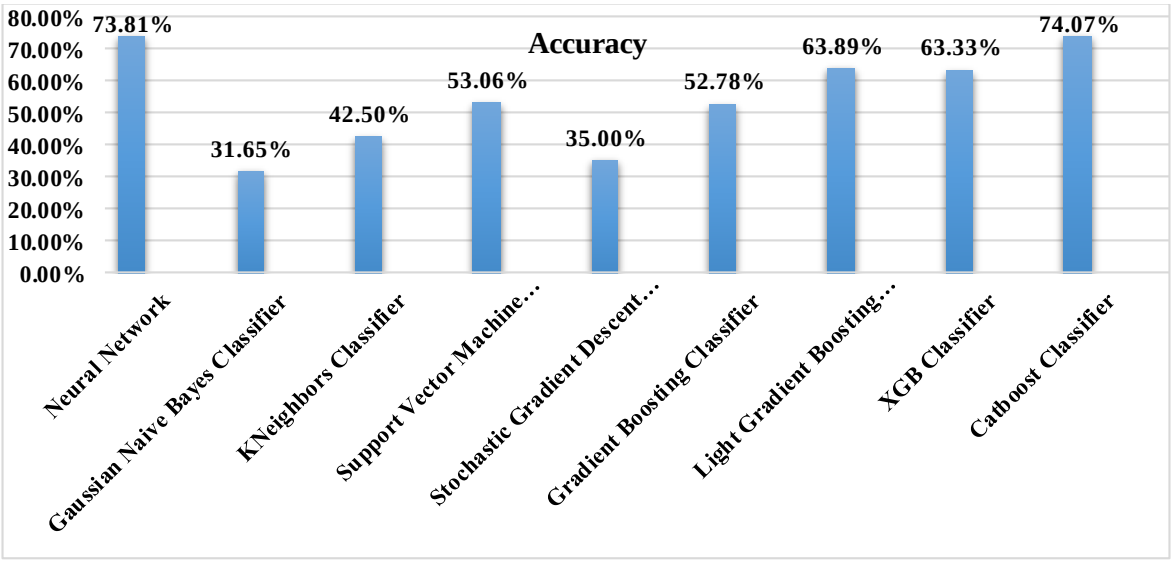

Figure 4. Comparison of Various Classifiers

6. Conclusion

Emotion Recognition from speech has numerous applications in medical field, personal assistance, smart home etc., Hence research works are being done in that field. Numerous machine learning models were implemented to detect emotions from speech. Our work makes an inclusion of catboost classifier in that list. The accuracy offered by this classifier on a reputed dataset shows that catboost is also a valid algorithm for emotion recognition from speech. In future, this algorithm can be tested against the various datasets for testing its efficiency in the emotion recognition. As well it can also be coupled with other powerful algorithms in order to produce a robust model to detect emotions more accurately from the speech.

Since emotions are also dependent on the accent of the language, accent based classifiers might also be developed for the proficient analysis of emotions. To be more precise and accurate, voice assistants can go through a particular person's speeches and create a machine learning model for that individual which may result in an optimized assistance.

References

[1] Suraj Tripathi, Abhay Kumar, Abhiram Ramesh, Chirag Singh, Promod Yenigalla, "Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions" , CICLing June 11 ***2019***.

[2] S. Casale, A. Russo, G. Scebba, S. Serranov ,"Speech Emotion Classification using Machine Learning Algorithms", International Journal of Computer Applications, IJCA Journal, Volume 118 - Number 13, ***2015***.

[3] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, Te-Won Lee, "Emotion recognition by speech signals", 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, ***2003***.

[4] Nithya Roopa S., Prabhakaran M, Betty.P, " Speech Emotion Recognition using Deep Learning ", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-7Issue-4S, November ***2018***.

[5] K.Tarunika, R.B Pradeeba, P.Aruna ,"Applying Machine Learning Techniques for Speech Emotion Recognition ", 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), ***2018***.

[6] Alejandro Uribe, Alejandro G´omez, Manuela Bastidas, O. Lucia Quintero, Damian Campo, "A Novel emotion recognition technique from voiced-speech ", IEEE 3RD COLOMBIAN CONFERENCE ON AUTOMATIC CONTROL (CCAC), ***2017***.

[7] Beth Logan,"Mel Frequency Cepstral Coefficients for Music Modeling", International Symposium on Music Information Retrieval, ***2000***.

[8] Aibek Ryskaliyev, Sanzhar Askaruly, Alex Pappachen James, "Speech Signal Analysis for the Estimation of Heart Rates Under Different Emotional States ", International Conference on Advances in Computing, Communications and Informatics (ICACCI), September 21-24, ***2016***.

*[9] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev,Anna Veronika Dorogush, Andrey Gulin, " CatBoost: unbiased boosting with categorical features ", v5, 20 January **2019**.*

*[10] Chernykh, V., Sterling, G.: Emotion Recognition From Speech With Recurrent Neural Net-works. In: arXiv:1701.08071v1, **2017**.*

*[11] T. S. Polzin, A. Waibel, "Emotion-sensitive humancomputer interfaces" ISCA Workshop, Speech and Emotion, **2000**.*

*[12] "Sequential k-nearest neighbor pattern recognition for usable speech classification", Jashmin K Shan, Brett Smolenki , Robert E Yantorno , Ananth N Iyer, **2004** 12th European Signal Processing Conference, IEEE Proceedings.*

*[13] N. A. Meseguer, "Speech analysis for automatic speech recognition," Norwegian University of Science and Technology, Masters Thesis, vol.109, **2009**.*

*[14] C.-W. Huang and S. S. Narayanan, "Characterizing Types of Convolution in Deep Convolutional Recurrent Neural Networks for Robust Speech Emotion Recognition," pp. 1–19, **2017**.*

*[15] S. Lugovic, I. Dunder, and M. Horvat, "Techniques and applications of emotion recognition in speech," 2016 39th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2016 - Proc., no. November 2017, pp. 1278–1283, **2016**.*