EMOTION CLASSIFICATION USING MULTI-SVM

Project Report

KRITIK MATHUR, 140905586
NIMISH AGRAWAL, 140905596

PROBLEM STATEMENT

With the advent of virtual assistants in computers, speech recognition has become more and more important. Speech recognition has been around for a lot of years but it still is a field that observes continuous improvements like accent recognition, language identification, real time translation, etc.

The goal of this project is to predict the mood of the speaker based on the MFCC values using the Support Vector Machine (SVM) as the classifier.

INTRODUCTION

"Emotional prosody refers to the melodic and rhythmic components of speech that listeners use to gain insight into a speaker's emotive disposition."

Like, for example, studies have shown that sad emotions are produced with a higher pitch, less intensity but more vocal energy (2000 Hz), longer duration with more pauses, and a lower first formant. Such emotional prosody based studies will be the basis for our mood prediction algorithms.

The feature that we have used in this project for classifying different emotional states is Mel-frequency cepstral coefficients (MFCC).

DATASET DESCRIPTION

The dataset has been taken from the Linguistic Data Consortium - https://catalog.ldc.upenn.edu/LDC2002S28

This dataset consists of certain phrases recorded in different emotions by a theatre actor and a text file consisting of intervals of phrases and the emotion they are in.

The format of these clippings is *Wave* (.wav).

SURVEY

During present scenario, for human emotion recognition an extensive research is made by using different speech information and signal. Many researchers used different classifiers for human emotion recognition from speech such as Hidden Markov Model (HMM), Neural Network (NN), Maximum likelihood Bayes classifier (MLBC), Gaussian Mixture Model (GMM), Kernel deterioration and K-nearest Neighbours approach (KNN), support vector machine (SVM), Naive Bayes classifier.

Of all these, we have used SVM as a classifier to classify different emotional states.

MFCC

Mel Frequency Cepstral Coefficents (MFCCs) are a feature widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since. Prior to the introduction of MFCCs, Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) (click here for a tutorial on cepstrum and LPCCs) and were the main feature type for automatic speech recognition (ASR), especially with HMM classifiers.

The steps for computing the MFCCs for a given audio signal are as follows:

1. Frame the signal into short frames.
2. For each frame calculate the periodogram estimate of the power spectrum.
3. Apply the mel filterbank to the power spectra, sum the energy in each filter.
4. Take the logarithm of all filterbank energies.
5. Take the DCT of the log filterbank energies.
6. Keep DCT coefficients 2-13, discard the rest.

The Mel scale relates perceived frequency, or pitch, of a pure tone to its actual measured frequency. Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Incorporating this scale makes our features match more closely what humans hear.

The formula for converting from frequency to Mel scale is:

$$M(f) = 1125\ln(1 + f/700) \qquad (1)$$

To go from Mels back to frequency:

$$M^{-1}(m) = 700(\exp(m/1125) - 1) \qquad (2)$$

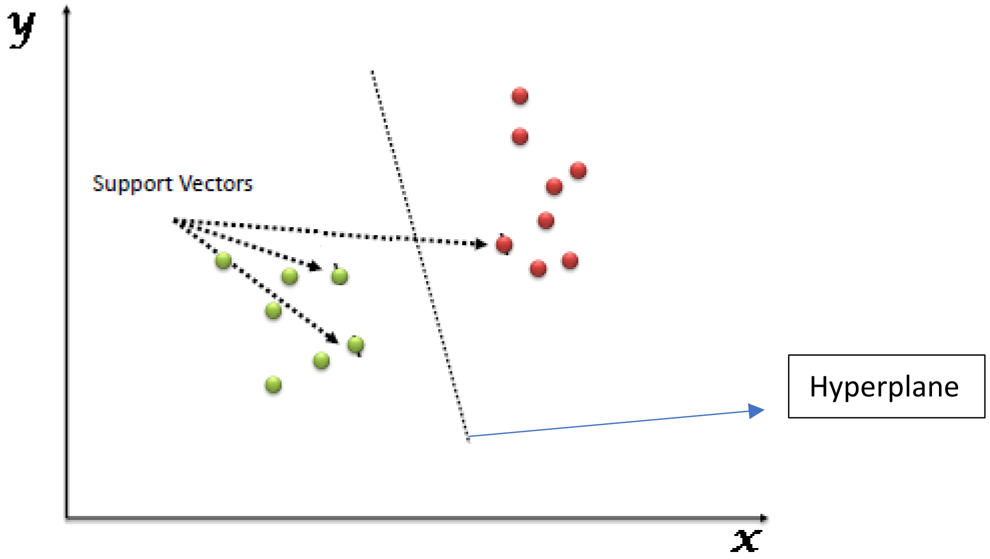
SUPPORT VECTOR MACHINES

The SVM is a high dimensional vector supervised learning method that is based on emotion assumptions. It predicts that the presence (or absence) of a specified feature of a class is not related to the presence (or absence) of all other features. It is very simple to program and execute it, its parameters are simple to assume, even on very large databases learning or training is very fast and effective and its accuracy is comparatively better in comparison to the other techniques.

It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. Simply put, it does some extremely complex data transformations, then figures out how to separate your data based on the labels or outputs you've defined.

In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot).

Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes.



In the above figure, the green dots and the red dots represent two different classes which are separated by the hyperplane.

SCOPE

The voice based emotion recognition has many applications, especially in the field of Artificial Intelligence. The recognition of mood can help the AI behave in a certain way. For example, if the person's mood is found to be angry while the person is driving, the automobile AI can take some pre-defined actions, so as to make sure an accident does not take place. Similarly, if the mood is found to be sad, then the mobile phone based AI may change the theme of the phone or take some actions in an attempt to cheer the user up.

Another use of emotion recognition can be in theatrics, or in stage setups where the stage light colour may change based on the mood of the performer.

GAPS

The major shortcoming in our project is due to the lack of a reliable dataset. Recording voice samples require high-end equipment and so, recording our own voice samples was not a possibility. Instead, we had to rely on dataset we found on the web, which has sufficient data for only two emotion sates, namely anger and sadness.

OBJECTIVE

With this project, we aim to develop a software which takes speech signal as an input, processes it, and in return outputs the mood of the speaker in terms of probabilities.

METHODOLOGY

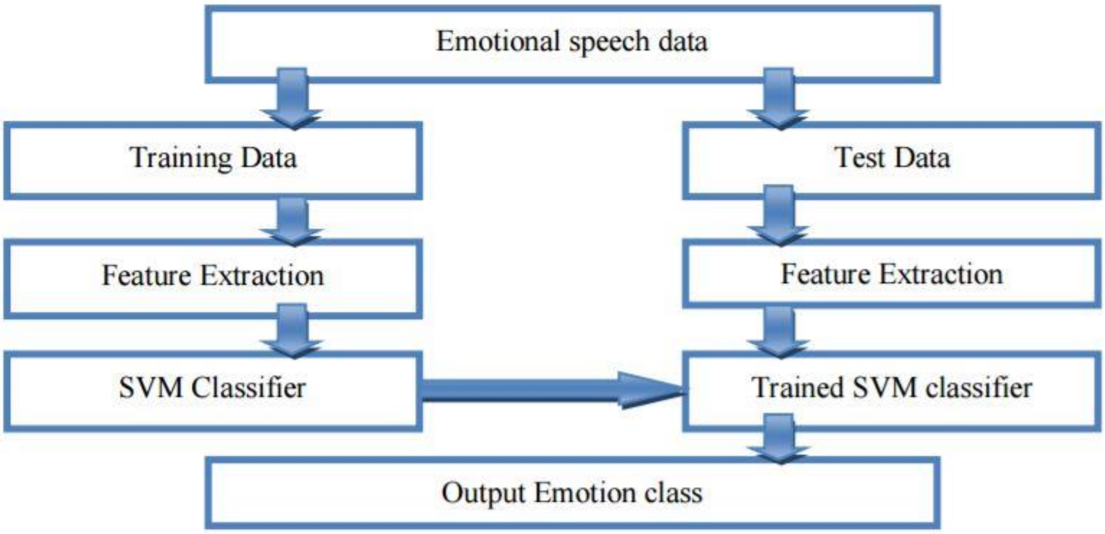
We use Support Vector Machines (SVMs) as a classifier to classify emotions such as Anger, Happiness, Sadness, etc.

There are a variety of temporal and spectral features that can be extracted from human speech. We use statistics relating to the pitch, Mel Frequency Cepstral Coefficients (MFCCs) and formants of speech as inputs to classification algorithms.

Broadly, different sub-stages for emotion recognition will be:

i) Feature Extraction – from stage-1 output
ii) Feature Labelling – to train the SVM
iii) SVM Training – to generate a model for mood prediction
iv) Feature Extraction – for test data
v) SVM classification – for testing using the generated model

This process along with training and testing phases can be represented by the below flowchart:



FLOWCHART FOR EMOTION RECOGNITION PROCESS

RESULTS

We were successfully able to classify four different mood states – anger, sadness, elation, and happiness with this project.

Both SVM and KNN successfully classified 19 out of the 21 given test samples.

However, the confidence value for SVM true positives was very high (over 0.6) everytime and fairly low for false positives (0.4-0.45) hence proving SVM better, on a larger test data.

CONCLUSION

The results were as expected, since SVM performed better than the KNN classifier.

The reason behind the accuracy not being 100 per cent was the lack of adequate amount of clean data for training the classifier.

REFERENCES

- “Pattern Classification” - Richard O. Duda, David G. Stork, Peter E.Hart
- practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/
- “Speech Emotion Recognition Based on SVM Using MATLAB” – Ritu D. Shah, Dr. Anil C. Suthar
- www.analyticsvidhya.com
- www.catalog.ldc.upenn.edu – for emotion samples and transcript
- www.yaksis.com
- MATLAB documentation