DS-864 Machine Learning

Project Report

Video Analysis

Submitted by

Roll No	Name of Students
MT2015004	Aditi Raghuvanshi
MT2015047	Grishma Ajmera
MT2015084	Priyanak Tiwari

Under the guidance of

Dr. Vaibhav Rajan

Prof. G Srinivasaraghavan

International Institute of Information Technology,
Bangalore

Autumn - 2016

Acknowledgement

We would like to thank our supervisor Dr. Vaibhav Rajan and Prof. G Srinivasaraghavan for all their help and guidance.
He was always available whenever we needed some advice or ran into any problem. We would also take the opportunity to thank the God and our parents for their blessings and our beloved friends for their encouragement and motivation during this period.

Aditi Raghuvanshi
Grishma Ajmera
Priyanak Tiwari

Contents

1 Introduction 1

2 Assumption 2

3 Analysis on Dataset 2

4 Feature Extraction 4

5 Algorithmic approach 6

6 Results and Observations 7

7 Future scope 8

References 9

1 Introduction

In the last few decades, the educational system has seen a major drift towards online video lectures. Websites like NPTEL, Coursera etc provides opportunities for the learners to enroll into the courses of highly qualified professors from MIT, Stanford, IITs etc which was not possible earlier. With time, amount of videos have increased tremendously. Videos can be studied for analysing content, discussion time, examples used for explanation, blackboard time and many more.

Our project aims to build a machine learning based system that analyzes videos and fetch statistics from the video. Statistics mainly comprise of analysing the overall explanation time,number of slides displayed, number of diagrams/figures, number of lines in a slide presented in a video etc.

Along with the mentioned factors, metadata of video is also considered where various other information like title, number of views, number of likes and dislikes, comments etc. These statistics can help in predicting popularity of the video.

Various approaches are used for calculating statistics related to a video which is described in detail in feature extraction heading. In this project we are not doing analysis on the video content. We are doing processing on video key frames,it's youtube metadata and audio processing for feature extraction.

Machine learning is one such field in computer science which can build a robust model classifying videos into various features and will keep learning itself with time. Rating of a video in youtube is a hidden metadata. By using feature extraction technique, we are taking rating value as our label which will be in the range 0-5. The model built help us to predict rating for video lecture based on the statistic calculated for it. Thus by using this model we can predict video rating based on its extracted features.

2 Assumption

- We are using videos of short duration since feature extraction takes a lot of time for longer lectures.
- Lecture should contain slides. It should not be backboard class.
- Lecture is available in youtube from where we can fetch youtube metadata details.
- Video should contain only lecture slide and a professor. No audience should be visible in video.
- Video should be available offline. Online video prediction is also possible using some api but that is not handled in this project. It can be considered as future scope.
- All video file will be in mp4 format and kept inside Data folder of current directory.
- All video file name and its youtube url will be present in an csv file from where it will pick up video file name and corresponding youtube link.

3 Analysis on Dataset

A dataset is a collection of discrete items of related data. Dataset is a statistical data matrix, where every column of the table represents a particular feature, and each row corresponds to a new member of data item. Following datasets are used for building model to predict rating of video.

I. Key Frame dataset
II. Audio dataset
II. Youtube metadata dataset

1. **Key Frame Dataset** : We did analysis on video frame to evaluate number of slides used, number of lines used, font size of content and transition time between two key frame. These statistic was used as features to model our system. To begin with we segmented videos into chunks of frames and then found key frames from them. Based on key frame we extracted above mentioned statistical values which was useful in modelling the system.

We used OpenCV tool for getting key frames and content font size and number of line of content.

2. **Audio Dataset** : We processed video file and converted it into audio file which was useful to get dataset for getting statistical feature from audio. An example of a relevant feature for audio is the frequency of speaker. We analyzed pitch rate average, pitch changes, loudness of speaker from audio signal.

3. **Youtube Dataset** : We are taking youtube video metadata from youtube using web crawling done by regular expression matching. For taking metadata value we are using pafy. We can download youtube video also using pafy but that we have not done in this project. We have made an assumption that video is already available to be processed and its realted url is also stored in an excel file. Thus all videos should be in mp4 format and should be present inside current directory in Data/ folder with same name which is specified in the excel beside its youtube url link. For using pafy for extracting youtube metadata it is necessary to have good internet connection.

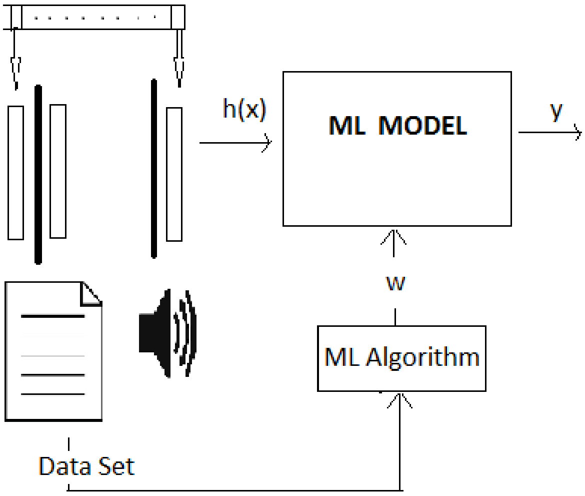


Figure 1: Model Building using DataSet

4 Feature Extraction

1. **Features from Frames** : We are using Ocropus for getting number of lines, and font size from image. We could have also used Tesseract but ocropus is more transparent and hackable. Since we are not doing analysis on content of video semantically, we used features which represent good presentation skill. For eg. if there is too much content on slide that video will be probably disliked video.

   First of all, as a part of pre-processing binarization of image is done, with the help of Ocropus library command stated below:
   ocropus-nlbin -n filename -o book

   Then after, features from key frames are being extracted total number of lines, font size, number of slides using the below command:

ocropus-gpageseg -n –maxcolseps 0 book/imagename

Also, image mining is done to extract number of figures from key frames:
extract-figures.py –interactive filename

The transition time between the key frames is also captured.

Since we are not dealing with content we need to check lectures based on its appearance and time taken. Because of this above mentioned features are used for predicting rating of a video.

2. **Audio Features** : Corresponding to the video being analyzed, its audio is generated in the form of .wav file using the ffmpeg command:
   ffmpeg -i Data fileName out.wav

   We have used PRAAT for speech analysis. There are numerous speech features that can be extracted from the audio like pitch rate, loudness of speaker etc. using the following command:
   praat –run speechRateV2.praat -25 2 0.3 1 filename

   We have also used mfcc features which is generated using OpenSmile command line statement.

   Video rating depends on the way speaker is speaking. Like if a speaker is speaking very low or very fast that video should be having low rating. Modulation in voice also is one of the features extracted out from the video.

3. **Youtube Features** : Given youtube video URL, corresponding metadata is extracted. Using PAFY library we have taken out the features like - video duration, view count, number of likes, rating etc. These features are very important features for our model creation. This features provide as data based on which we build our model. Our prediction value is extract from youtube metadata. Number of views, number os lifes and view count are important factors for rating a video.

Table 1: Features Used in Model creation

Number of lines of text in a frame
Font size of the text in frame
Transition Time between two key frames
Number of figures in a frame
Voice Pause feature
Speech duration
Phonotation time
Speech rate
Articulation rate
Speaking time
MFCC feature vector
Number of view Count of video
Number of likes of video
Duration of video

5 Algorithmic approach

We will be using supervised learning for making predictions,it will use a known dataset (called the training dataset),the training dataset includes input data and response values From it, the supervised learning algorithm seeks to build a model that can make predictions of the response values for a new dataset. As objective is to predict rating from 1 to 5 for the video based on its feature, we can use multiclass logistic regression for the same.

Multinomial(class) logistic regression is a particular solution to the classification problem that assumes that a linear combination of the observed features and parameters, that can be used to determine the probability of each particular outcome of the dependent variable. It is a very efficient classifier that learns to separate all data points of one class from those of the other class to minimize the probability of misclassification. The feature distribution in our data sets is so complicated that different classes may have overlapping or interwoven areas. The features that we will be using to perform classification tasks can be -
- Colour based features, such as colour histograms, texture and edge information.
- Shot-based information, such as shot length and transition type.
- Object-based features,such as faces or text boxes.
- Motion-based features,such as optical flow, frame difference, motion vectors.

- Audio,text and visual information can be exploited all together to improve the results on the single classifier,while the different features can also be combined for further classification.
There will be two basic phases involved in the process -
1) Training phase - which is used to let the system "get to know" about the features.
2) Testing phase - where we either compare unidentified features with the training data or compute the likelihood between the both,a decision will then be made based on comparison results.

6 Results and Observations

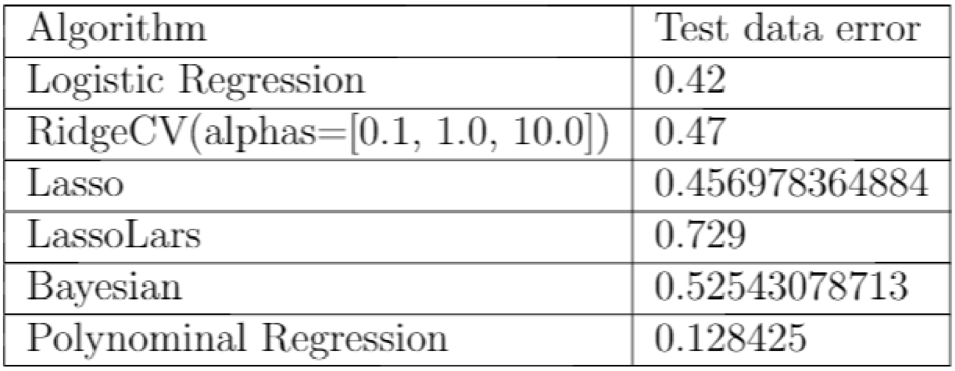
As a part of dataset, 50 videos were taken and for each video. Frame feature extraction takes a lot of time for big videos,so we can limit key frame count to 100, so that modeling takes lesser time. So we did analysis on sampled 80 frames and they were analyzed for extracting frame related features.

To train the model, we have applied 80-20 random split, out of which 80% would be used to train the model of prediction and 20% would be used as testing data.

We have tried various algorithms for finding the best, giving lowest test data error. Refer figure 2.
Logistic Regression is basically used as a binary classification algorithm, which can be extended to multiclass classification or one-vs-all approach. Another, algorithm is RidgeCV where different values of alpha are taken and is based on imposing a penalty on the size of coefficients to minimize the squared error. The Lasso is a linear model that estimates sparse coefficients, and because of its tendency to prefer solutions with fewer parameter values, effectively reducing the number of variables upon which the given solution is dependent. Polynomial Regression is using linear models trained on nonlinear functions of the data. The least-squares used in this method minimizes the variance of the unbiased estimators of the coefficients. It also depends on the degree of the model.

We found logistic regression was giving better result than rest all other regression model. We found 0.42 error using logistic regression. Using ridge regression(which is like logistic regression but with imposed penalty on the size of coefficients which is helpful for regularization) we obtained error of 0.47 by selecting best model using alpha values as 0.1, 1.0, 10.0. By using polynomial regression using degree of 2 we got best result. We have got square error as 0.128425 on whole data set and 0.00635625 on test data alone.

Table 2: Results



Algorithm	Test data error
Logistic Regression	0.42
RidgeCV(alphas=[0.1, 1.0, 10.0])	0.47
Lasso	0.456978364884
LassoLars	0.729
Bayesian	0.52543078713
Polynominal Regression	0.128425

7 Future scope

1) Can use youtube url directory and download video and then do processing on it. This feature is already avaialble in pafy.
2) Can do content analysis on video using transcript.

References

[1] About Ocropus[Online]
Available:http://www.danvk.org/2015/01/09/extracting-text-from-an-image-using-ocropus.html/

[2] About Different modeling algorithm
Available: http://scikit-learn.org/stable/index.html

[3] Youtube metadata Libraries[Online]
Available:http://pythonhosted.org/Pafy/

[4] Praat code
Available: https://github.com/timmahrt/praatIO

[5] Finding figures in image,
Available: http://www.pyimagesearch.com/2014/10/20/finding-shapes-images-using-python-opencv/
Available: https://github.com/acdha/image-mining