Pattern Recognition

Course Project Updated Report

ONLINE NEWS POPULARITY

Group-12

V Banu Theja - S20190020258

Y Dileep Chandra - S20190020262

T Surya Satvik – S20190020254

Overview

With the help of Internet, the online news can be instantly spread around the world. Most of peoples now have the habit of reading and sharing news online, for instance, using social media like Twitter and Facebook. Typically, the news popularity can be indicated by the number of reads, likes or shares. For the online news stake holders such as content providers or advertisers, it's very valuable if the popularity of the news articles can be accurately predicted prior to the publication. Several feature extraction techniques such as analysis description, PCA, log transformation, etc. are implemented in available datasets to calculate and compare them. The work of this project includes data extraction, data observation, data cleansing, feature extraction (including feature selection) and classification.

Introduction

Online news is accelerating day by day due to the growth of social media. Big number of articles are published daily on various platforms such as Medium, GeeksforGeeks, etc. Articles fall into various categories such as sports, technology, politics, etc. and are published on different days. Online news content includes many key attributes: Number of images, number of videos, content size, etc. Based on these properties, the popularity of new articles can be predicted.

On the other hand, Unpopular news can be observed in terms of its content whether it is positive or negative. The length of the content i.e., number of words, images, videos, advertisement also have an impact on its popularity. Based on these features (input), we can analyse an article in terms of shares (output). The main aim of the project is to explore various features in the dataset and analyse them to give a model for predicting the popularity of an online news article before publishing.

Problem Statement

The aim of the project is to solve a binary classification problem i.e., to predict if an online news article is popular or not by using machine learning techniques learnt in pattern recognition. The popularity is characterized by the number of shares. **If the number of shares is higher than a pre-defined threshold, the article is labelled as popular, otherwise it is labelled as unpopular.**

Thus, the problem is to utilize a list of article's features and find the best machine learning model to accurately classify the target label (popular/unpopular) of the articles to be published. As the problem can be formulated as a binary classification problem, we have implemented using classification learning algorithms including Logistic Regression, Linear SVM, Fine Gaussian SVM, Logistic, K-Nearest Neighbours (KNN).

Data Observation - Raw data

This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years. Data sets used for analysis (www.mashable.com), which contains information on 40,000 articles published between 2013 and 2015. This dataset is publicly available in the UCI Machine Learning Repository. Total Features/Attributes: 61, Unexpected Attributes: 2 (Article URL, time delta), Predicted Attributes: 58, Target Attributes: 1 (number of shares).

Data Extraction

The data is originally loaded from the Mashable repository file is in .csv format (comma Separated values including all article URLs along with attributes). The downloaded CSV data is extracted as numeric matrix or tabular data (as shown below) so that it is compatible with software for cleaning and processing purposes like training and implementing algorithms.

	url	timedelta	n_tokens_title	n_tokens_content	n_unique_tokens	n_non_stop_words	n_non_stop_unique_tokens	num_hrefs	num_self_hrefs	num_imgs	nu
0	http://mashable.com/2013/01/07/amazon-instant-...	731.0	12.0	219.0	0.663594	1.0	0.815385	4.0	2.0	1.0	
1	http://mashable.com/2013/01/07/ap-samsung-spon...	731.0	9.0	255.0	0.604743	1.0	0.791946	3.0	1.0	1.0	
2	http://mashable.com/2013/01/07/apple-40-billio...	731.0	9.0	211.0	0.575130	1.0	0.663866	3.0	1.0	1.0	
3	http://mashable.com/2013/01/07/astronaut-notre...	731.0	9.0	531.0	0.503788	1.0	0.665635	9.0	0.0	1.0	
4	http://mashable.com/2013/01/07/att-u-verse-apps/	731.0	13.0	1072.0	0.415646	1.0	0.540890	19.0	19.0	20.0	

num_videos	average_token_length	num_keywords	data_channel_is_lifestyle	data_channel_is_entertainment	data_channel_is_bus	data_channel_is_socmed	data_channel_is_tech
0.0	4.680365	5.0	0.0	1.0	0.0	0.0	0.0
0.0	4.913725	4.0	0.0	0.0	1.0	0.0	0.0
0.0	4.393365	6.0	0.0	0.0	1.0	0.0	0.0
0.0	4.404896	7.0	0.0	1.0	0.0	0.0	0.0
0.0	4.682836	7.0	0.0	0.0	0.0	0.0	1.0

self_reference_max_shares	self_reference_avg_sharess	weekday_is_monday	weekday_is_tuesday	weekday_is_wednesday	weekday_is_thursday	weekday_is_friday	weekday_is_saturday
496.0	496.000000	1.0	0.0	0.0	0.0	0.0	0.0
0.0	0.000000	1.0	0.0	0.0	0.0	0.0	0.0
918.0	918.000000	1.0	0.0	0.0	0.0	0.0	0.0
0.0	0.000000	1.0	0.0	0.0	0.0	0.0	0.0
16000.0	3151.157895	1.0	0.0	0.0	0.0	0.0	0.0

ositive_polarity	avg_negative_polarity	min_negative_polarity	max_negative_polarity	title_subjectivity	title_sentiment_polarity	abs_title_subjectivity	abs_title_sentiment_polarity	shares
0.7	-0.350000	-0.600	-0.200000	0.500000	-0.187500	0.000000	0.187500	593
0.7	-0.118750	-0.125	-0.100000	0.000000	0.000000	0.500000	0.000000	711
1.0	-0.466667	-0.800	-0.133333	0.000000	0.000000	0.500000	0.000000	1500
0.8	-0.369697	-0.600	-0.166667	0.000000	0.000000	0.500000	0.000000	1200
1.0	-0.220192	-0.500	-0.050000	0.454545	0.136364	0.045455	0.136364	505

Data Cleaning

Data cleaning or pre-processing involves removing any outliers which damage the accuracy of the model in any terms of mean.

Feature Extraction

Feature extraction is a dimensionality reduction process applied to transform raw data into managed data or to remove extraneous data components. It causes the performance of the model to degrade. In this section, conforming to the same definition, it focuses on techniques such as general feature extraction techniques with comparison of resulting accuracy.

Raw Data Analysis

Raw Data is data without any external transformations.

Classification Models and Results

Models tested on and worked with: Linear SVM, Fine Gaussian SVM, Logistic, K-Nearest Neighbours

Table 1: Results obtained from modelling excluding outliers

Classification Model	Time Elapsed	Accuracy	Remarks
Logistic	13 s	68.2%	Good performance
KNN	26.18 s	56.98%	Decent performance
Linear SVM	86.34 s	68.20%	More time consuming
Gaussian SVM	346.89 s	68.19%	More time taking than linear

These results show that for a predefined threshold and by deleting few features, around 68% of the articles present in the dataset is labelled as popular by the above-mentioned classification models.

Table 2: Results obtained from modelling including outliers

Classification Model	Time Elapsed	Accuracy	Remarks
Logistic	32.89 s	65.11%	Good accuracy & performance
KNN	66.27 s	57.31%	Less accuracy compared to logistic
Linear SVM	709.86 s	64.02%	Good accuracy and more time
Gaussian SVM	1015.39 s	60.59%	Too slow and good accuracy

Log Transformation Technique

In general, machine learning algorithms work best on normally distributed data. In either regression or classification, some algorithms implicitly assume that the data is normal. Some features are visualised. This affects the accuracy of the modelling.

So, the two main advantages of log transformations are,

1) removes this skewness and makes the data as "normal" as possible.
2) converts complex multiplication relation between the data to linear addition relation.

Transformation: $Y = \log(y)$

Classification Models and results

Table 3: Results obtained from modelling including outliers- Log Transformation

Classification Model	Time Elapsed	Accuracy	Remarks
Logistic	48.83 s	70.44%	Well and Excellent performance
KNN	122.33 s	62.42%	Not as accurate as logistic

Table 4: Results obtained from modelling excluding outliers – Log Transformation

Classification Model	Time Elapsed	Accuracy	Remarks
Logistic	2.57 s	68.20%	Very fast and quick performance
KNN	12.13 s	56.99%	Quick performance
Linear SVM	77.95 s	68.20%	Good accuracy
Gaussian SVM	264.06 s	68.19%	Comparatively more computation

PCA Technique Modelling

The PCA (Principal Component Analysis) is a commonly used dimensionality reduction algorithm that could give us a less dimensional approximation to the original dataset, while preserving the possible variability. It is a type of pattern recognition in the data. PCA is a powerful tool for data analysis. Principal component analysis deals with the reduction of features/attributes present in data by representing linear combination of available variables with specific weights (eigenvalues and eigenvectors).

$$Y = a_1 X_1 + a_2 X_2 + a_3 X_3 + \ldots + a_n X_n$$

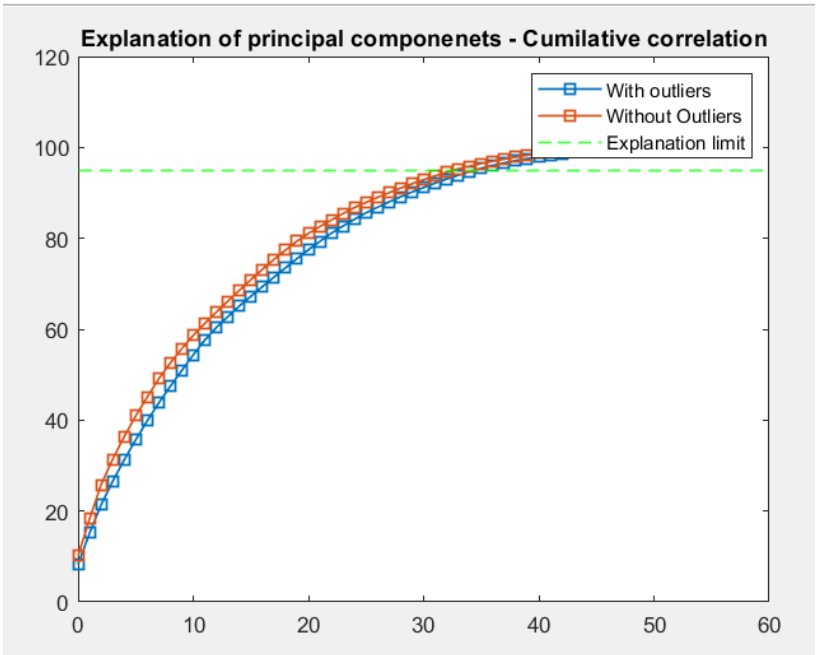
The first step in performing PCA is to normalize the available data to continue the analysis, adjusting all variable ranges between them to (-1 and 1).

$$Z = \frac{X - \mu}{\sigma}$$

The next step continues with eigenvalue searches. Eigenvalues are the actual components that help determine the coefficient of explanation.

$$|Z - \lambda I| = 0$$

Eigenvectors that can be extracted from the data and eigenvalues are actually principal components or representations of the original data in a low-dimensional feature space where the variance description is stored.



Classification Models and Results

Models tested on and worked with: Linear SVM, Fine Gaussian SVM, Logistic, K-Nearest Neighbours

Table 5: Results obtained from modelling including outliers- PCA Modelling

Classification Model	Time Elapsed	Accuracy	Remarks
Logistic	5.33 s	70.07%	Well and good with outliers
KNN	32.04 s	62.34%	Good but not as logistic

Table 6: Results obtained from modelling excluding outliers – PCA Modelling

Classification Model	Time Elapsed	Accuracy	Remarks
Logistic	2.9 s	68.02%	Very fast and more accurate
KNN	6.77 s	56.28%	Good performance and accurate
Linear SVM	93.47 s	68.20%	Accurate and good computation
Gaussian SVM	460.07 s	68.19%	More computation time

Conclusion

Feature extraction components/methods are very useful for improving model performance. Efficiency lies not only in improving accuracy, but also in reducing mathematical background computation and evaluation time. This is evident in the observation tables, the analysis on complete raw data without an analysis has taken very long time compared that of analysis with feature extractions. The accuracy of all methods is very similar, and the real improvement lies in the calculations.

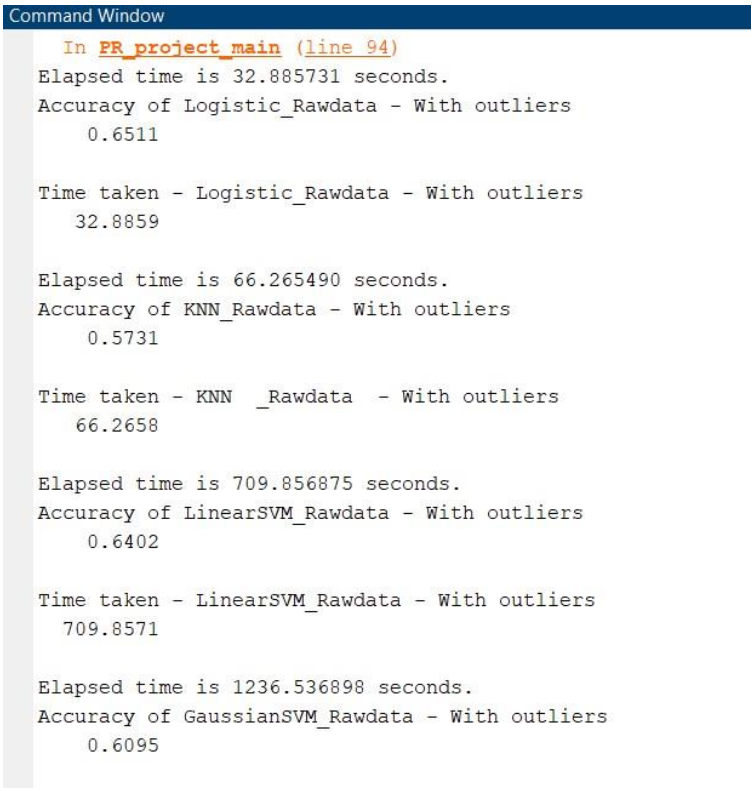
Contribution

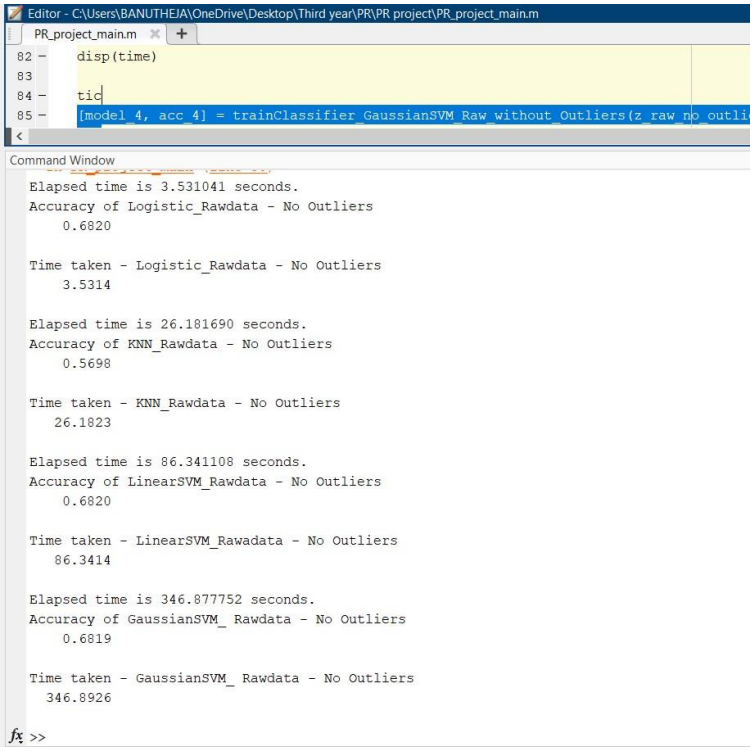
V Banu Theja – Feature Extraction - PCA and modelling, Raw Data Analysis.

Y Dileep Chandra – Feature Extraction, Log Transformation and modelling.

T Surya Satvik – Data extraction and modelling.

Codes and Outputs





Remaining images of all outputs are in the mentioned below drive link

Code and all model's outputs: link

https://drive.google.com/drive/folders/1Cm4anZv2KPnGtVOiNcxwEnhnkgT6WALv?usp=sharing

In the codes uploaded in the drive link , PR_project_main.m is the main code and rest are the defined functions sir

References

K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.