Literature Review

Team: The Miner League

SRN: PES2201800065

Name: Rishab Kashyap

PAPER3:

Chanseung Lee, Mina Jung, 2016, PREDICTING MOVIE SUCCESS FROM SEARCH QUERY USING SUPPORT VECTOR REGRESSION METHOD, International Journal of Artificial Intelligence & Applications (IJAIA), https://www.aircconline.com/ijaia/V7N1/7116ijaia01.pdf

This paper takes a slightly different approach into explaining how the querying via Google Trends about the movie titles, actors and various other attributes allows in the prediction of a successful movie. Their basic idea is that Query data from search engines can provide many insights about human behavior. Therefore, massive data resulting from human interactions may offer a new perspective on the behavior of the market.

For the model the first step, like in the previous two papers, is to extract and clean the data that they have retrieved. The data sources include the following:

- Movie Query
- Rate (R)
- Number of Theatres (T)
- Number of words in movie title (W)
- Income of the movie (I)
- The prediction process is divided into three parts:
- Data Collection
- Data Pre processing
- Movie Prediction

The queries come into picture when users type in movie names on search engines. Popular movies come up more often as suggestions rather than the less popular ones, such as RoboCop versus Refuge.

In this paper, they have made use of Linear and ridge regression models to predict movie incomes. Then they make use of the SVM/SVR technique to do the classification and summary of regression models. Multivariate regression model is used to predict multiple dependent variables from the independent variables with the help of the sklearn library in python. The MSE and R square values are calculated along with the residuals.

The next thing is Ridge regression , which is made use of to improve the performance of the already existing model by adding in the correction factor (This is also known as regularized regression, or RLR). Unfortunately, this resulted in a worse model compared to linear regression with a higher MSE and presence of more outliers than in the previous case.

Hence, a kernel technique is made use of to solve the issue of non-linearityThe basic idea of the kernel method is to expand original dimension into higher dimension, and find a linear line in the expanded dimension.Suppose Φ is the mapping from original dimension to expanded dimension. If we transform our points to feature space, the scalar product form of two original data x1 and x2 looks like this:

<Φ(x1), Φ(x2)>

SVM in this case is the most popular kernel regression method and is once again made use of in this case for classification purposes. It provides the best accuracy compared to the previous regression methods. Its MSE is 4.85, which is a significant improvement than the two other regression methods. Polynomial kernel generates the importance/weight of each input variable.