Predicting the Possibilty of Appearance of a Cited Paper on YouTube Based on its Social Media Features

Srikanth Reddy Nagidi
Computer Science
Northern Illinois University
DeKalb Illinois U.S.A
z1836478@students.niu.edu

Krishnasree Mukkavilli
Computer Science
Northern Illinois University
DeKalb Illinois U.S.A
z1816567@students.niu.edu

Godwin Richard Thomas
Computer Science
Northern Illinois University
DeKalb Illinois U.S.A
z1838366@students.niu.edu

ABSTRACT

The possibilty of appearance of a cited paper on YouTube is predicted using the social media features. The entire Altmetrics dataset was taken into consideration and then sampling was performed to obtain the training set. The above sample was subjected to normalization to ensure that the data was in the optimal range. Features were selected one at a time after which a total of twelve features were selected for the classification. Seventy percent of the dataset was divided into the training set and the rest of the data was divided into test set. The class imbalance problem was dealt by using Synthetic Minority Oversampling Technique(SMOTE). Classifiers like K-Nearest Neighbour, Logistic Regression, RandomForest, Decision Trees and Logistic Regression with Smote were used for classification. The metrics on which the classifiers were tested was selected to be precision and recall. For our project, recall was considered to be the primary metric. The classifier having the highest recall was selected. The confusion matrix was obtained and the ROC curve was plotted.

KEYWORDS

Youtube, Social Media, Prediction, Altmetrics, Citations, Logistic Regression, SMOTE, Recall, Precision, Normalization, Class Imbalance.

1 Introduction

The number of citations of a paper is impacted by a lot of different factors. Currently, especially after the digitalization of the papers, it is easier to track not only the number of citations but also how they are influenced. In the era of digital searches, the users rely upon keyword searches or online browsing [1]. Upon investigating various sources which keep track of different features in which the citation counts were affected, we stumbled upon something which was detailed and intriguing. Altmetrics are metrics and qualitative data that are complementary to traditional, citation-based metrics. They provide this in the form of varioustypes of features. It is to be noted that, some publishers have turned to Altmetrics because they appear more rapidly than citations [2,3].

We wanted to select a paritcular field which is currently popular and has a big impact on people all over the world. social media was narrowed down because of its enormous user base and the amount of different platforms in which it is available. Some of the famous social media platforms we took into consideration were Facebook, Twitter and YouTube. The rise of popularity of social media has been gradual and now it is at a stage where it is present everywhere. In today's world, it is undeniable that social media plays an important role in impacting our culture [4].

We wanted to integrate both the disciplines and hence ultimately decided to find the impact social media has on the number of citations of the paper. Instead of directly correlating the different social media features like Facebook and Twitter with the citation count, we thought of using a social media platform as our target variable and then based on the list of social media features we have for that particular paper, the appeareance of the paper on the target variable could be determined. The target variable had to be something which was very popular currently. Figures showed that over 1.9 Billion logged in users visit YouTube each month and every day people watch over a billion hours of video and generate billions of views [5]. Based on the above statistic, we took Youtube as the target variable.

2 Dataset

The Altmetrics dataset was chosen as our dataset. Also, it was easier to collect the entire data. As altmetrics focuses on social media platforms that often provide free access to usage data through Web APIs, data collection is less problematic [6]. The dataset was merged from different JSON files and then integrated into a single ".csv" file.

The entire dataset was not taken at once and sampling was performed on the dataset. Sampling helps reduce the time taken to process the data and also gain information based on the subset of the data [14]. Around 100,000 tuples were considered.

Initially, we noticed that the data did not contain any missing values or tuples. But, some of the values were not in the same range. Also, we figured that the true negative dominated the true positive resulting in the class imbalance problem which had to be taken care of.

The dataset consisted of twelve features being present as the columns and the tuples containing the data. For our target variable YouTube, we did not want to see the number of counts or the number of views for the particular paper. We assumed 1 as the value when the paper will be displayed on YouTube and 0 as the

value when the paper will not be displayed on YouTube. All the values in the YouTube column were converted to either 0 or 1.

3 Preprocessing

One of the observations we found out was that the values in one of the features "Mendeley" was in a completely different range compared to the other features. Normalization was performed as one of the very first steps in preprocessing the data in Data Mining [8]. Upon further reading, we decided to normalize the values and found out that values were now in the correct range.

Initially, the values present in the Mendeley had values like 133, 194, 104, 105, 1, and 2. After normalization, the values were changed to 2.4, 2.5, 1.2 , -0.3 , -0.5 and -0.7. Basically, the range was completely normalized to make sure that all the other data was also in the same range as this particular column.

3.1 Class Imbalance Problem

The Class Imbalance problem is a case when the total number of class data that are positive is far less than the total number of class data that are negative [9]. In our case, we encountered a total number of 99,384 for 0's and a total number of 616 for 0's. To overcome this issue, we were looking for an oversampling technique.

When we looked at previous works to deal with the class imbalance problem, there was an attempt where the original population of the minority class was not changed and the majority class was under-sampled and also keeping the geometric mean as the performance measure for the classifier [10]. Another work was when the minority class was over-sampled and the majority class was under-sampled [11]. Some of the problems with the under-sampling approach was that it might lead to loss of valuable data and lead to bias.

After a lot of consideration, we finally decided to proceed with Synthetic Minority Oversampling Technique (SMOTE). Smote synthesizes new minority instances between existing minority instances. SMOTE causes the decision region of the minority class to become more general [12]. After using SMOTE, the number of 1's rose up to 69,569.

4 Feature Selection

Feature selection was applied to reduce the number of features in many applications where data has hundreds or thousands of features [7]. From the list of features present in the dataset, initially 11 features were taken into consideration where the feature "YouTube" was taken as the target variable. We did not take all the features at the same time due to the complications possibly involved. Therefore, we decided to perform the classification process with only one feature at the starting. The metrics on which the classifiers were judged here were recall and precision. After that, we wanted to add in one feature at a time. All the social media features like Facebook, Twitter, Linkedin etc., were taken as the features.

One interesting thing to note was there was a noticeable difference in the recall and precision when we added the feature "Altmetrics Score". There was an increase in both recall and precision with the addition of the new feature. Because of that, we decided to include that to the list of features making the total number of features to be 11.

4.1 Identical Features

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used as a way to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

To test whether the features were similar to each other, we constructed the correlation matrix for all the features against the target variable. From the matrix, the highest score out of all the features was 0.15 for the feature "blogs" and the lowest score was 0.011 for the feature "F100". Therefore, we decided not to drop any of the features and keep all of them.

Some other interesting details which could be inferred from the correlation matrix were that altmetrics score and news had the highest score of 0.86. As far as other features were concerned, apart from "Altmetrics Score" and "Blogs", all the other features were more or less of the same range when compared with YouTube.

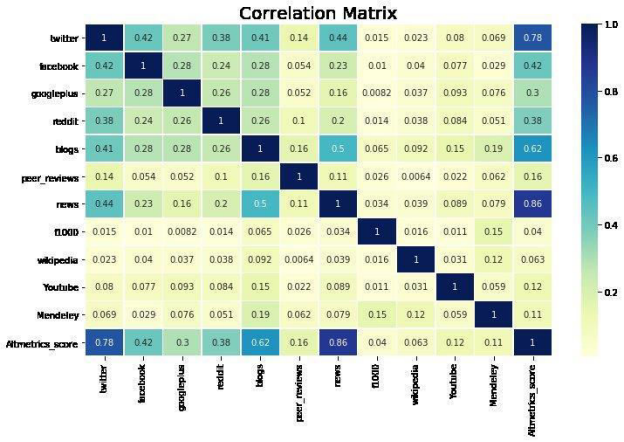

Figure 1: Correlation Matrix between features and YouTube.

5 Classification

The entire dataset was divided into training set and test set. Seventy percent of the data was divided into training set and thirty percent of the data was divided into test set. Five classifiers were used for the entire process.

5.1 K-Nearest Neighbor(KNN)

The very first classifier to be used was the K-Nearest Neighbor (KNN). In KNN, the basic idea is that an object is classified by a majority vote of its neighbors, with the object

being assigned to the class most common among its "k" nearest neighbors. In the first step of the algorithm, a positive integer "k" is specified, along with a new sample. After that we select the "k" entries in our database which are closest to the new sample. We find the most common classification of these entries.

One of the key features of KNN we were able to find out after using it was that KNN stores the entire training dataset which it uses as its representation. Also, KNN makes predictions just-in-time by calculating the similarity between an input sample and each training instance.

KNN yielded a precision of 0.72 and a recall of 0.07. We were quite happy with the precision but the recall was very low. Also, the precision wasn't completely indicative of what was happening on the inside because of the accumulation of True Negative.

5.2 Decision Tree[Bagging]

A decision tree is used to visually and explicitly represent decisions and decision making. Based on the decisions, we can either traverse to the left of the tree or to the right of the tree.

Two of the most popular techniques for constructing ensembles are Bagging [15] and the Adaboost family of algorithms which is also known as Boosting [16]. Both of these methods operate by taking a base learning algorithm and invoking it many times with different training sets.

In bagging, each training set is constructed by forming a bootstrap replicate of the original training set. In other words, given a training set S of m examples, a new training set S0 is constructed by drawing m examples uniformly (with replacement) from S [17].

The Adaboost algorithm maintains a set of weights over the original training set S and adjusts these weights after each classifier is learned by the base learning algorithm. The adjustments increase the weight of examples that are misclassified by the base learning algorithm and decrease the weight of examples that are correctly classified [17].

For our project, we proceeded to go with Bagging. After we performed the classification, we were able to obtain a precision of 0.82 which was the highest out of all the classifiers and a recall of 0.26 which we were pretty satisfied with.

5.3 Random Forests

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees [18] [19].

When we performed this classificatio, we were able to produce a precision of 0.77 and a recall of 0.22. This was almost similar to the Decision Tree with Bagging in terms of both precision and recall.

5.4 Logistic Regression

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

We did not want to get into much detail on Logistic Regression as the recall was the lowest out of all the other classifiers (0.02). We wanted to upgrade Logistic Regression by concatenating it with SMOTE.

5.5 Logistic Regression[SMOTE]

With SMOTE the objective is to find a new balanced dataset which includes all the majority class examples and a synthetic over-sampled replica of the minority class examples, such that the new set is balanced [12]. This classifier had the highest recall of 0.46. Therefore, for this project, the chosen classification algorithm is Logistic Regression with SMOTE.

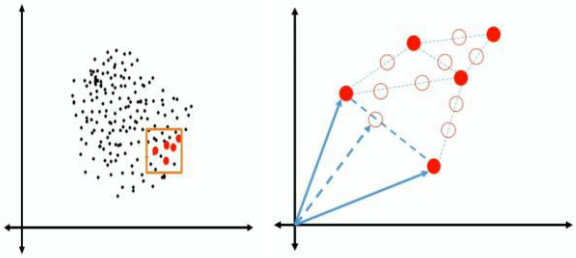

Figure 2: SMOTE

6 Metrics

Precision and Recall were used as the metrics for the classifiers. Machine learning algorithms were typically evaluated using predictive accuracy but this was not appropriate when the data was imbalanced and/or the costs of different errors very markedly [6]. Considering the above statement, we proceeded to take recall as the evaluation metric. The classifier with the highest recall was Logistic Regression with SMOTE and we decided to use that as our classifier.

7 Performance Measurement

Logistic regression model is a modeling procedure applied to model the response variable Y that is category based on one or more of the predictor variables X, whether it is a category or continuous [7]. Logistic Regression was used along with SMOTE to deal with the class imbalance problem.

7.1 Confusion Matrix

In a binary decision problem, a classifier labels examples as either positive or negative. The decision made by the classifier can be represented in a structure known as a confusion matrix or a contingency table. The confusion matrix has four categories: True Positives (TP) are examples correctly labeled as positives. False Positives (FP) are examples referring to negative values labelled

incorrectly as positive. True Negatives (TN) correspond to negatives correctly labelled as negative. Lastly, False Negatives (FN) refer to positive examples incorrectly labeled as negative [20]. The confusion matrix for our project is given below:

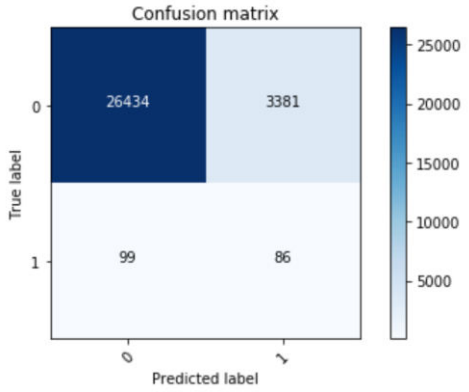

Figure 3: Confusion Matrix

7.2 ROC Curve

From the confusion matrix, we are able to plot the ROC curve. The False Positive Rate (FPR) is plotted on the x-axis and the True Positive Rate (TPR) is plotted on the y-axis. The FPR measures the fraction of negative examples that are misclassified as positive. The TPR measures the fraction of positive examples that are correctly labelled. The ROC curve is given below:

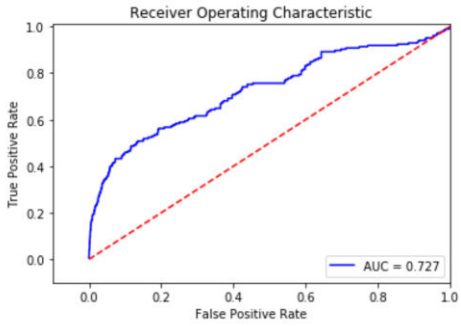

Figure 4: The ROC Curve

8 Future Work

Instead of taking only a portion of the Altmetrics dataset, the entire Altmetrics dataset could be taken as a whole. Though this would consume more time and space, the recall and precision could definitely vary. Also, under the different classifiers, Support Vector Machines (SVM) can be used to perform the classification. Different sampling techniques other than SMOTE could be experimented with for the class imbalance problem.

REFERENCES

[1] Thelwall M, Haustein S, Larivière V, Sugimoto CR (2013) Do Altmetrics Work? Twitter and Ten Other Social Web Services. PLoS ONE 8(5): e64841. https://doi.org/10.1371/journal.pone.0064841

[2] Adie E, Roe W (2013) Altmetric: enriching scholarly content with article-level discussion and metrics. Learned Publishing 26: : 11–17. Available: http://figshare.com/articles/Enriching_scholarly_content_with_article_level_discussion_and_metrics/105851. Accessed 2013 February 19.

[3] Seglen PO (1992) The skewness of science. Journal of the American Society for Information Science 43: 628–638.

[4] Amedie, Jacob, "The Impact of Social Media on Society" (2015). Advanced Writing: Pop Culture Intersections. 2. http://scholarcommons.scu.edu/engl_176/2

[5] Youtube, 2018. Retrieved from https://www.youtube.com/yt/about/press/.

[6] Priem J (2013) Altmetrics. In: Cronin B, Sugimoto C, editors. Bibliometrics and Beyond: Metrics-Based Evaluation of Scholarly Research, Cambridge: MIT Press, in press.

[7] Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. Journal of machine learning research, 5(Oct), 1205-1224.

[8] Shalabi, L.A., & Shaaban, Z. (2006). Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix. 2006 International Conference on Dependability of Computer Systems, 207-214.

[9] Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent data analysis, 6(5), 429-449.

[10] Kubat, M., & Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: OneSided Selection. InProceedings of the Fourteenth International Conference on MachineLearning, pp. 179–186 Nashville, Tennesse. Morgan Kaufmann.

[11] Ling, C., & Li, C. (1998). Data Mining for Direct Marketing Problems and Solutions. InProceedings of the Fourth International Conference on Knowledge Discovery and DataMining (KDD-98)New York, NY. AAAI Press.

[12] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357

[13] A. Agresti, Categorical Data Analysis, John Willey & Sons, Inc, New York, 2002.

[14] Marshall, M. N. (1996). Sampling for qualitative research. Family practice, 13(6), 522-526.

[15] Breiman, L. (1994). Heuristics of instability and stabilization in model selection. Technical Report 416, Department of Statistics, University of California, Berkeley, CA

[16] Freund, Y. & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In Proc. 13th International Conference on Machine Learning (pp. 148–146). Morgan Kaufmann.

[17] Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning, 40(2), 139-157.

[18] Ho, T.K. (1995) Random Decision Forest. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, 14-16 August 1995, 278-282.

[19]Ho, T.K. (1998). The Random Subspace Method for Constructing Decision Forests. IEEE Trans. Pattern Anal. Mach. Intell., 20, 832-844.

[20]Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240). ACM.