Features Selection Effect on Predicting the Popularity of Online News

Wedyan Alswiti, Ali Rodan
King Abdullah II School of Information Technology, The University of Jordan
Amman 11942 Jordan
E-mail: Wedyanmhmd89@gmail.com, a.rodan@ju.edu.jo

Abstract

News popularity occupied space in modern mining problems where it's became important to predict the audience for a specific news or journal for many parties include journal itself, advertisers and many others. In this paper we study the impact of feature selection on the quality of popularity prediction, by applying different features count on different classification models and different attribute ranking models. Classifiers like J48 and AdaBoost (J48) shows a noticeable sensitivity to the selected features for training and testing with different ranker, where less features gives better prediction accuracy that ranges from 60 to 64 %. For Random Forest classifier more used features give better accuracy and the best accuracy accomplished by using all the data set features, other classifiers shows variant sensitivity to feature selection.

Keywords: Classification, Features Selection, Genetic Search, Random Forest, News popularity, Support vector Machine, Information Gain.

1. INTRODUCTION

News importance in today's world measured by people interaction with news site, where social networks provide channel that people can share information by posting news, links of news or sharing their opinion [1]. News articles with its dynamic nature and time sensitivity became an area for researchers to predict and dedicate its popularity in social network.

News popularity may be defined in different ways, where some consider popularity as the number of times the article is clicked; others measure it by the number of shares. As the definition vary and the threshold that consider the article to be popular or not also vary from region to region, and from domain to domain where it strongly depends on the article area of interest and triggered audience. Due to the time sensitivity of the news articles, it became important to journalists, content providers, advertisers, and news recommendation systems to accurately estimate the extent to which the articles will spread [2]. Predicting the popularity of news is an important issue, but it get challenging due to the variety of population and resources taken into consideration the network properties that affect the social networks structure.

Classification is an important aspect of data mining which refers to the process of discovering a class for a specific entity with unknown label or class using a set of rule. Classification as a prediction technique used in many research areas like medicine, social media and other daily life aspects. Several classification methods have grown widely and now a lot of them are available for use. One of the used classifiers is Random Forests (RF) [7]. RF defines a new methodology for constructed classification trees by introducing new technique for splitting each node using the best among a subset of predictors randomly chosen at that node [7], Other classifier like J48 was developed by Ross Quinlan as a decision tree algorithm that represents an implementation of the C4.5 algorithm where it is used in data mining tools like Weka [8] along with Naive Bayes classifier which rely on simple probabilistic using Bayes theorem. Naive Bayes assumes that all attributes are independent given the value of the class variable [10]. One of the oldest classifiers is k-nearest neighbor (kNN), which considers as the simplest classification method since it uses the simple Euclidean distance to measure the dissimilarities between examples represented as vector inputs [9]. Moreover, an approach of classification was investigated in [11] that is called classification via regression which aims to binaries Class then one regression model is built for each class value [12].

In this work we focus on the study of features selection to identify the key features affected the prediction accuracy using many well-known classifiers for online news data, where every news article is represented by group of features, mainly the key feature which is the content of the article, other factors that are also used in classification study focused on the time manners of population and the post sources. In our paper we will study and compare many algorithms based on the accuracy of the model and the recall value.

The paper organization is as follows. Section 2 provides a related work. In Section 3, the dataset is described with its characteristics and features. The results of the prediction methods are presented in Section 4. Finally, in Section 5 our work is concluded with a discussion of future works.

2. RELATED WORK

Classification use for popularity prediction has grown recently due to the expansion of social network, Flavio Figueiredo in [13] tries to predict trends and hits in user generated videos, by analyzing public features for video provided by YouTube on a database of 24,482 videos, first by extraction patterns of popularity evolution using K-Spectral Clustering (KSC), then he build the prediction model based on extremely randomized ensemble trees and evaluate its F1 and Macro-F1 metrics using 5-fold cross validation which achieved 95% Confidence [13].

Prediction was also used for ranking as in [14] where user comments was used to predict the popularity of news articles by analyzing 260,000 articles collected from 2007 to 2011. Three prediction models where tested, linear regression, linear regression on a logarithmic scale, and constant scaling model. Those models where then evaluated and tested based on heuristics and Random method, Time of publication method, and a weighted method between the time of publication and the number of comments. The results show the most appropriate prediction method which is the simple linear regression.

In [15] a model for predicting the long-time popularity prediction of online content defined by user's access using YouTube and Digg, for predication LN model: linear regression on a logarithmic scale; least-squares absolute error, CS model: constant scaling, relative squared error, GP model: growth profile models were used and evaluated using the absolute squared error QSE and the relative squared error QRE.

Naïve Bayesian Classifier was used in [16] to model a user's general preferences for news stories and predict the short term stores, news stories was presented as features vector where each feature indicates the presence or absence of a word, for evaluating the model a data set of 3,000 total rated news stories was used.

By applying classification via regression, the authors of [2] achieved 84 % accuracy in predicting ranges of popularity on twitter, four different characteristics used for each article 1)The news source, 2)The category of news, 3)The subjectivity of the language and 4)Named entities mentioned in the article. They compared the result of four classification methods, Bagging, J48 Decision Trees, SVM [5], and Naive Bayes. The best method was bagging with 83.96% accuracy.

Using different dataset the authors of [3] uses Ranking Support Vector Machines and Pattern recognition by training the model with articles collected for one year. The authors tried to predict whether the next article will be popular or not by processing the data set that represents set of articles grouped by outlet. Then they performed standard text mining pre-processing techniques, namely, stop word removal, stemming by using different classification methods. The accuracy ranges from 58.6 to 75 % and by improving the model an accuracy of 86% was reached.

In [4] the authors build an Intelligent Decision Support System to analyzes articles, using dataset of 39,000 articles from the Mashable website, first the authors Perform Data Acquisition and Preparation which represent all articles published for two years from Mashable, then they extract a 47 feature from the articles HTML code, finally they classify the attributes into: number {integer value; Ratio {within [0; 1]}; Boolean {2 f0; 1g}; and nominal, the second pre-processing step take care of scaling the unbounded numeric features. By using 70% of the data set as training data, the prediction models gives the following accuracies {Random Forest (RF)- 0.67% , Adaptive Boosting (AdaBoost)- 0.66, Support Vector Machine (SVM) -0.66, K-Nearest Neighbors (KNN)- 0.62 , Naive Bayes (NB)- 0.62}.

To enhance the performance the Author of [4] uses features ranking using Random Forest model and tested the optimization and reach with an overall area under the Receiver Operating Characteristic (ROC) curve of 73%.

3. DATASET

The used data set contains 39,000 records each represents article collected for two years, the data is for news article was collected and pre-processed by [4]. The processed data represented by 61 features (58 predictive attributes, 2 non-predictive, 1 goal field) as shown in Table 1.

The target feature is the number of shares for each article where the type of this feature is number and ranges from 1 to 843300.

As mentioned in the dataset source the article considers being popular if the number of shares exceeded 1400 else it's classified as un-popular. In this work we randomly used 70% of the data as training and the rest as testing data.

Table 1: List of attributes by category [4]

Feature	Type(#)
Words	
Number of words in the title	number (1)
Number of words in the article	number (1)
Average word length	number (1)
Rate of non-stop words ratio	ratio (1)
Rate of unique words ratio	ratio (1)
Rate of unique non-stop words	ratio (1)
Links	
Number of links	number (1)
Number of Mashable article links	number (1)
Minimum, average and maximum of shares of Mashable links	number (3)
Digital Media	
Number of images	number (1)
Number of videos	number (1)
Time	
Day of the week	nominal (1)
Published on a weekend?	bool (1)
Keywords	
Number of keywords	number (1)
Worst keyword (min./avg./max. shares)	number (3)
Average keyword (min./avg./max. shares)	number (3)
Best keyword (min./avg./max. shares)	number (3)
Article category (Mashable data channel)	nominal (1)
Natural Language Processing	
Closeness to top 5 LDA topics	ratio (5)
Title subjectivity	ratio (1)
Article text subjectivity score and its absolute difference to 0.5	ratio (2)
Title sentiment polarity	ratio (1)
Rate of positive and negative words	ratio (2)
Pos. words rate among non-neutral words	ratio (1)
Neg. words rate among non-neutral words	ratio (1)
Polarity of positive words (min./avg./max.)	ratio (3)
Polarity of negative words (min./avg./max.)	ratio (3)

4. EXPERIMENT SETUP AND RESULTS

Random Forest model achieves the best accuracy after applying some optimization techniques to reduce number of used feature, based upon [4,6]. In this part we examine the effect of feature selection on the accuracy, and by studying the selected attributes count on the accuracy and what attributes are selected, this involve testing the same features selection model to generate different features count, then use those attribute to evaluate different classifiers, so to test what attribute we use same classifier with different features selection method. For classification [18, 19], we examine different classifiers starting with **Random Forest**, **KNN**, **SMO**:algorithm for training a support vector classifier, **AdaBoost** (**J48**):Class for boosting a nominal class classifier using the AdaBoost M1 method, where only nominal class problems can be tackled and it often dramatically improves performance, but sometimes over fits, which can be handled by **Bagging**-Class for bagging a classifier to reduce variance. All these models can do classification and regression depending on the base learner.

In the first experiment **Bagging** with **Random Forest** (RF) was used as classification model to be tested with different number of features ranked upon the importance of each feature in specifying the popularity of an article, we use **Gain Ratio feature evaluator** to evaluates the worth of an attribute by measuring the gain ratio with respect to the class which distribute counts for missing values.

After applying the feature selection with the ranker search method the features achieves the highest ranked and selected to be used for farther evaluation are shown in Table 2.

Data_channel_is_socmed - Is data channel 'Social Media- this feature is of type Keywords achieved the best rank by 0.02974 , the day the article published also was of top ranked feature where ***weekday_is_Saturday*** and is weekend occupied the 2nd and 3rd place in top ranked features by 0.02687 and 0.02647 respectively. Moreover, ***weekday_is_Sunday*** was in the 4th place with 0.01441, many of the top ranked feature was categorized as keyword features like Avg. keyword (max. shares), Avg. keyword (avg. shares), Best keyword (max. shares) and Worst keyword (max. shares), features related to the article subject considered as affected features in the popularity prediction like ***Is data channel 'World'?*** , ***Is data channel 'Entertainment'?*** And ***is data channel***

Technology? For the next 20 features the rank ranges from 0.004591 to 0.001949 after the 52 feature the rank reaches zero.

Table 2: Top 20 ranked features

Ranke	Attribute
0.02974	data_channel_is_socmed
0.02687	weekday_is_saturday
0.02647	is weekend
0.02303	data_channel_is_world
0.01441	weekday_is_sunday
0.01428	kw_max_avg
0.01417	kw_avg_avg
0.01391	data_channel_is_entertainment
0.01119	data_channel_is_tech
0.01099	kw_min_avg
0.01073	self_reference_avg_sharess
0.00998	self_reference_min_shares
0.00874	self_reference_max_shares
0.00847	LDA_02
0.00565	n_unique_tokens
0.00545	kw_max_max
0.00523	num_imgs
0.00518	kw_max_min
0.00503	global_subjectivity
0.00501	LDA_01

After applying feature selection using **Gain Ratio Attribute Evaluation,** we examine the accuracy of **Random Forest** model with **bagging** and **100 tree sensitivity** where the number of selected features with their results is shown in Table 3.

Table 3: RF evaluation with different features.

No. of features	Accuracy	Recall	ROC
10	62.3476 %	0.623	0.665
20	66.2322 %	0.662	0.722
40	65.988%	0.660	0.722
58	**66.3415 %**	**0.663**	**0.722**

From Table 3 we might conclude that attribute count might affect the accuracy if the number of selected attribute was bellow specific count. Furthermore, other models were tested with different count of features as shown in Table 4.

We found depending on Table 4 that each algorithm respond differently to the change of features number using the same selection method, for J48, and AdaBoost with J48 methods, classification results shows Inverse relationship between the features count and the accuracy, on the other hand, Naïve Bayes, KNN, and SVM gave best result for 40 feature where Classification via regression (SYNOPSIS) shows best accuracy for features count 20 and worst for 10 features.

Table 4: Different models accuracy % with different features

Features	10	20	40	58
Naïve Bayes	59.867	58.580	61.582	60.884
Classification via regression	63.995	**64.676**	**64.449**	**64.079**
J48	63.93	63.55	60.90	58.256
AdaBoost(J48)	**64.161**	62.213	61.385	60.658
KNN	55.851	57.105	57.339	56.761
SVM	60.444	60.442	63.123	62.551

Random Forest with bagging score the best accuracy of all used models for features count (20, 40 and 58), but **AdaBoost (J48) for 10 features achieves the better accuracy** comparing to all other models of 64.1613%. Other classifiers like SMO shows stable behavior and almost no change of accuracy with the change of features count until it reach the total number of attributes where the accuracy increase by almost 2.3%.

In the next step we use the random forest as classification model to test different features selection methods. Table 5 shows the accuracy of AdaBoost (J48) model with different feature selection methods.

The used Selection models are:

1. **Correlation Based**: selection that calculate correlation evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class we use it to select top 20 feature using ranker search method.
2. **Information Gain:** Evaluates the worth of an attribute by measuring the information gain with respect to the class. This model also used to select top 20 best features.
3. **Learner based:** Evaluates attribute sets by using a learning scheme. Cross validation is used to estimate the accuracy of the learning scheme for a set of attributes. To test the model different learner algorithms was used like (**J48, RandomForest, SVM [17], and KNN**) and select the best first attributes.

For every feature selection method the accuracy and recall and ROC was measured using J48 classifier, the time here represent the needed time to build each model upon the selected classifiers for Correlation Based, Information Gain, and Gain

ratio. The model was forest to select top 20 feature for the learner based on the best first search that searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility.

Learner based features selection method achieve the best accuracy close to gain ratio and way better than other used model but with longest model building time of 6235.1 comparing with Gain ratio that needed only 0.25 % of the learner based consumed time and give accuracy of 97.25% of Learner based (J48).

Table 5: J48 Classifier evaluating with different features selecting models

Method	Accuracy	Recall	ROC	Time
Correlation Based	60.792	0.608	0.64	39.88
Information Gain	60.607	0.606	0.65	47.2
Gain Ratio	62.213	0.622	0.66	**16.14**
Learner based (J48)	**63.970**	**0.640**	**0.64**	6235.1

5. CONCLUSION

Every classification model tend to have different sensitivity when it came to attributes selection, where some shows static behavior to increase accuracy by increasing the number of included feature, other had a threshold with the features count increase. On the other side, the feature selection method also affect the accuracy of a model where the best result achieved was by using bagging with RandomForest and Gain Ratio feature selection with all set of feature of 66.3415%, the second best accuracy achieved by Classification via regression and Gain Ratio with 20 feature of 64.6767 %. Some models shows interesting results like AdaBoost with J48 where the less number of features lead to better accuracy, the Gain ratio also shows a close accuracy to the best feature selection method with less timely cost and can be adapted for such tasks along with J48.

We can conclude that it's a matter of choosing the right classifier, right selection method, and the best number of features, where there is no generic rule for models behavior keeping in mind the time complexity for each model.

6. REFERENCES

[1] Lerman, K., & Ghosh, R. (2010). Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. ICWSM, 10, 90-97.

[2] Bandari, R., Asur, S., & Huberman, B. A. (2012). The pulse of news in social media: Forecasting popularity. arXiv preprint arXiv:1202.0332.

[3] Hensinger, E., Flaounas, I., & Cristianini, N. (2013). Modelling and predicting news popularity. Pattern Analysis and Applications, 16(4), 623-635.

[4] Fernandes, K., Vinagre, P. & Cortez, P. (2015). A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. In Portuguese Conference on Artificial Intelligence (pp.535-546) Springer International Publishing .

[5] Rodan, A., & Faris, H. (2016). Credit Risk Evaluation Using Cycle Reservoir Neural Networks with Support Vector Machines Readout. In Asian Conference on Intelligent Information and Database Systems (pp. 595-604). Springer Berlin Heidelberg.

[6] Ren, H., & Yang, Q. (2017). Predicting and Evaluating the Popularity of Online News.

[7] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

[8] Gholap, J. (2012). Performance tuning of J48 Algorithm for prediction of soil fertility. arXiv preprint arXiv:1208.3943.

[9] Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research, 10(Feb), 207-244.

[10] Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International Journal of Computer Science and Applications, 6(2), 256-261.

[11] Chen, P. C., Lee, T. J., Lee, Y. J., & Huang, S. Y. (2010) Multiclass support vector classification via regression. Tech. Rep., Tech. Rep.

[12] http://weka.sourceforge.net/

[13] Figueiredo, F. (2013). On the prediction of popularity of trends and hits for user generated videos. In Proceedings of the sixth ACM international conference on Web search and data mining (pp. 741-746). ACM.

[14] Tatar, A., Antoniadis, P., De Amorim, M. D., & Fdida, S. (2012). Ranking news articles based on popularity prediction. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012) (pp. 106-110). IEEE Computer Society.

[15] Szabo, G., & Huberman, B. A. (2010). Predicting the popularity of online content. Communications of the ACM, 53(8), 80-88.

[16] Billsus, D., & Pazzani, M. J. (1999). A hybrid user model for news story classification. In UM99 User Modeling (pp. 99-108). Springer Vienna.

[17] Rodan, A., Faris, H., Alsakran, J., & Al-Kadi, O. (2014). A support vector machine approach for churn prediction in telecom industry. International Information Institute (Tokyo). Information, 17(8), 3961.

[18] Rodan, A., & Faris, H. (2016). Optimizing feedforward neural networks using biogeography based optimization for e-mail spam identification. International Journal of Communications, Network and System Sciences, 9(1), 19.

[19] Rodan, A., Fayyoumi, A., Faris, H., Alsakran, J., & Al-Kadi, O. (2015). Negative Correlation Learning for Customer Churn Prediction: A Comparison Study. The Scientific World Journal, 2015.