Journal of Physics: Conference Series

Predicting and Evaluating the Online News Popularity based on Random Forest

Predicting and Evaluating the Online News Popularity based on Random Forest

Yuchao Zhang[1, a, *, †], **Kun Lin**[2, b, *, †]

[1]Jinling Institute of Technology, Nanjing, China

[2]Nankai University, Tianjin, China

[*]Corresponding email: [a]yuchao_zhang1229@163.com, [b]1911433@mail.nankai.edu.cn

[†]These authors contributed equally

Abstract. With the rapid development of science and technology, the internet has become a large media of information spread. There is a large quantity message on this platform. And online articles are the main form of information propagation. If the press can know what kind of articles will be more popular, they can construct an article that can help them spread the information they want to spread. Therefore, it's very important to predict the popularity of these articles. Some models in machine learning could be applied to this problem. In this paper, it will introduce an approach based on Random Forest. To avoid too much calculation, the experiment first uses PCA to make dimension reduction. Then the model evaluation uses the ROC area values to assess the accuracy of the model. Its performance is better than CART and C4.5.

1. Introduction

Due to the rise of major network platforms and the expansion of the world wide web, considered a universe of information, are popular at home and offices[1]. In this enormous search engine, there is much information reflecting in the form of articles. Suppose an article is popular, which means that it was among the "most read" articles shown online[2]. If it is possible to predict the popularity of an online article, it may help people to know the situation of the information propagation. With this prediction model, people can even use it construct information that is easily spread. Therefore, it is very important to predict the popularity of online news.

Based on the classification of online articles' popularity, Elena Hensinger proposes a standpoint in Modelling and predicting the news popularity that the popularity is not an absolute property, like yes or no, which has something in common with our view. But our experiment did meticulous division, and they think it's relevant to the article's competitive environment. It is related to the attributes of other articles on the same web page[2]. In predicting the popularity of online news using classification methods with feature filtering techniques, the researcher used Random Forest, CART, C4.5, and got an accuracy is mainly between 60%~70%[3]. In our model, the experiment can acquire a more accurate forecast result. Choudhary in Genetic Algorithm Based Correlation Enhanced Prediction of Online News Popularity selected least optimum attributes by genetic algorithm to maximize prediction accuracy. The results suggest that Naïve Bayes can achieve the best prediction rate of 93.46% by choosing 32 attributes, and Neural Networks perform 91.96% accuracy when the attributes are 18[9]. But in our study, the accuracy of sour experiment is higher.

In our study, first the online articles are divided more meticulously. The popularity is divided into three levels, high, middle, and low. Then make dimension reduction to the articles by PCA. The best dimension is selected by using the elbow method. the experiment creates a model based on Random Forest, and assesses its accuracy by the ROC value area, which has a better performance than CART and C4.5. Last, the experiment prunes the decision tree to avoid overfitting[8].

2. Methodology

2.1 Decision tree

The decision tree is a basic classification and regression method. It can be considered as the conditional probability distribution defined in the feature space and class space. The idea of the decision tree is actually to find the purest division method. The decision tree is to process a single feature, each step is to find an optimal feature for division, and each step is to divide by the optimal feature until the leaf node. The learning process of decision trees mainly includes three steps: feature selection, decision tree generation, and decision tree pruning. There are two commonly used decision tree algorithms: C4.5 and CART. The main difference lies in the use of different formulas for feature selection. C4.5 uses the split rate of Information Gain [4].

$$SplitInfo_A(S) = -\sum_{j=1}^{m} \frac{|S_j|}{|S|} \log_2 \frac{|S_j|}{|S|} \tag{1}$$

$$InfoGainRation(S,A) = \frac{InfoGain(S,A)}{SplitInfo_A(S)} \tag{2}$$

While CART uses the Gini coefficient[5].

$$Gini(p) = 1 - \sum_{k=1}^{k} p_k^2 \tag{3}$$

$$Gini(D,A) = \frac{D_1}{D} Gini(D_1) + \frac{D_2}{D} Gini(D_2) \tag{4}$$

2.2 Random Forest

Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest)[6].

3. Experiment

The Experiment is divided into three steps. The experiment consists of the following steps in order: data acquisition, followed by data preprocessing, and constructing the classification models.

3.1. Data Acquisition

The dataset of online news popularity is obtained by UCI Machine Learning Repository, the center for Machine Learning and Intelligent Systems [7]. This data set summarizes a heterogeneous set of features about articles published by Mashable in a period of two years. The goal is to predict the number of shares in social networks. The dataset consists of 39644 samples, with a total of 60 attributes and one label "shares". The features and labels of the online news data set are all numerical.

3.2. Data Preprocessing

3.2.1. Data Label. After loading the CSV file, a new column was created for the popularity after the "shares". Samples whose shares number is below 3000 are classified as low popularity. Samples whose shares number is greater than 3000 but below 10000 are classified as medium popularity. Samples whose shares number is greater than 10000 are classified as high popularity. What's more, URLs and time delta were removed because they were meta-data and cannot be used as characteristics. Also, the data preprocessing removed the shares column and assigned the popularity to be the class label, which finally led to 59 attributes. After that, the data set is divided into the training set and testing set. The experiment randomly selected 75% of the data as the training data set and the remaining 25% of the data as the testing data set. Table 1 shows the diving criterion of the three different popularity categories and the number of each category. It also shows that the number of the training set and testing set, respectively.

Table 1. Diving Criterion

Category	Range	Training set	Testing set	Total
High	shares>10000	1661	554	2215
Medium	3000<shares<10000	9131	3044	12175
Low	shares<3000	20203	5051	25254

3.2.2. Data Dimension Reduction. Dimension reduction is a method of data preprocessing for high-dimensional features. It is to preserve some of the most important features of high-dimensional data and remove noise and unimportant features to improve the speed of data processing. In this paper, the experiment uses the PCA algorithm to reduce the dimension of features by using the sklearn package in python to reduce the amount of calculation. Then the experiment chooses some points and applies these points to the Random Forest algorithm to find out the best components of dimension reduction by comparing the performance of the RF model.

3.3. Constructing the Classification Models

The experiment uses three methods, C4.5, CART, and Random Forest, to construct the classification models with the assistant of the sklearn package. They were trained on the same data set. Then pruning the decision tree that has been generated to optimize the model and avoid overfitting. What's more, the experiment applies the gridseachcv algorithm, which uses parameter regulation to prune the depth of the decision tree. The hyperparameter is set from 2 to 50, and the step size is 1. The experiment also record the running time of these three algorithms by using the time package.

4. Results and Discussion

4.1. Beat feature's dimension

The experiment chooses the dimensions from 1 to 30. Table 2 below shows the changes of accuracy with the dimension increasing. And it can be found that when the feature's dimension is 25, the accuracy reaches 1.00, reducing the feature to 25 dimensions.

Table 2. The accuracy of the different dimensions by using RF

Dimension	1	3	5	10	15	20	25	30
Accuracy	0.942	0.966	0.969	0.977	0.977	0.978	1.00	1.00

4.2. The performance of different classification

4.2.1. Accuracy. The results of the accuracy of each classifier are illustrated in the tables below. From Table 3, it was found that the accuracy of the training set of C4.5, CART, and RF is all 1.00, and the accuracy of the testing set of C4.5, CART, and RF is 0.996, 0.996, and 0.997. Also, the running time of C4.5, CART and RF are 1.247, 1.439, and 1.342. They all perform well both in the training set and testing set. However, it still may happen to overfit the data set. Hence, the experiment applies the gridsearchcv method to adjust the parameter of tree depth. From Table 4, the performance of these three classifiers almost does not have any difference. But the performance of RF is a little higher than the performance of the other two classifiers. Therefore, with respect to the classification accuracy as a performance measure, RF is the best among all other models, followed by C4.5 and CART.

Table 3. The accuracy of the different algorithms

Classifier	Accuracy of the training set	Accuracy of the testing set	Predicting time
C4.5	1.00	0.996	1.247
CART	1.00	0.996	1.439
RF	1.00	0.997	1.342

Table 4. The accuracy of the different algorithms after pruning

Classifier	Pruning the Max_depth	Adjustable parameter method	Accuracy of the training set	Accuracy of the testing set	Running time
C4.5	11	GridsearchCV	1.00	0.997	45.870
CART	24	GridsearchCV	1.00	0.997	53.068
RF	15	GridsearchCV	1.00	0.998	46.645

4.2.2. ROC. Under each category, it can get the probability of test samples being the category (the column in matrix P). Therefore, according to each column of probability matrix P and label matrix L, it can calculate the false positive rate (FPR) and true case rate (TPR) under each threshold and draw a ROC curve. In this way, three ROC curves can be drawn. Finally, the final ROC curve can be obtained by averaging the three ROC curves. Figure 1 shows that the ROC of different algorithms. It is found that the area of RF is 0.81. However, the area of C4.5 and CART are only 0.78 and 0.64, respectively. Therefore, the RF algorithm is the best among all other algorithms, followed by C4.5 and CART.

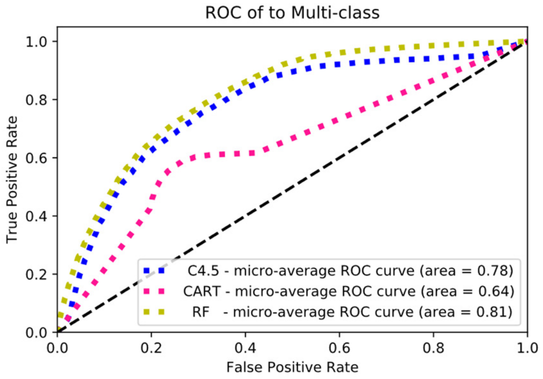

Figure 1. The ROC of different algorithms

5. Conclusion

The experiment use C4.5, CART, and RF to classify the article into three categories by using the sklearn package to forecast the online news popularity before publishing. This experiment reveals that the RF is the best among other algorithms, followed by C4.5 and CART. At this time, the experiment has many deficiencies. First of all, the experiment only uses three algorithms and two model evaluations. What's more, the experiment does not consider whether the data set has abnormal data or missing values. If the subject can be continued, the data will be collected by the domestic website so that the experiment can get the result which is line with the actual situation of China. In addition, more algorithms such as logistic regression, neural network, and other common classification models, and the more methods of model evaluation would be used in the experiment.

References

[1] Gordon M, Pathak P. Finding information on the World Wide Web: the retrieval effectiveness of search engines[J]. Information processing & management, 1999, 35(2): 141-180.Fkkfjkfj

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

[3] OBIEDAT R. PREDICTING THE POPULARITY OF ONLINE NEWS USING CLASSIFICATION METHODS WITH FEATURE FILTERING TECHNIQUES[J]. Journal of Theoretical and Applied Information Technology, 2020, 98(08).

[4] Singh S, Gupta P. Comparative study ID3, cart and C4. 5 decision tree algorithm: a survey[J]. International Journal of Advanced Information Science and Technology (IJAIST), 2014, 27(27): 97-103.

[5] Sharma H, Kumar S. A survey on decision tree algorithms of classification in data mining[J]. International Journal of Science and Research (IJSR), 2016, 5(4): 2094-2097.

[6] Biau G, Scornet E. A random forest guided tour[J]. Test, 2016, 25(2): 197-227.

[7] Information from: **https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity**

[8] Bramer M. Avoiding overfitting of decision trees[J]. Principles of data mining, 2007: 119-134.

[9] Choudhary, Swati K. et al. "Genetic Algorithm Based Correlation Enhanced Prediction of Online News Popularity." (2017).