Text Classifier Model with Decision Tree Learning

CMPT 310
Simon Fraser University
Fall 2018
Group name: ILikeKNN
Group members: Site Li(301244297),
Tianyang Zhou(301244114),
Liheng Ou(301270058)

Introduction

In this project, a machine learning model is built to estimate the validity of news articles with the decision tree method.

Method

Dataset

The dataset consists of 3000 news articles, which includes 1500 real news and 1500 fake news. We import the real and fake news dataset in WEKA to train the model.

Text Preprocessing

The first thing to do is preprocess the text from dataset. In WEKA, we used the StringToWordVector filter to preprocess the text file. We chose the Snowball stemmer, stop words from file, and the default tokenizer. Applying the filter, we got 1238 attributes of the relation. To remove the attributes with low correlation, we applied the AttributeSelection filter which reduced the number of attributes to 60.

Classification

WEKA provides the J48 algorithm for building the decision tree. We applied this algorithm to build the decision tree.

Evaluation

The 10-fold cross-validation method is used for the evaluation. We used the build-in 10-fold cross-validation tool in WEKA for this project.

Result

To get the best prediction, we built the models with different confidence factor. We tested the classifier with confidence factor from 0.1 to 0.9. We found the model with 0.2 of confidence factor has the highest accuracy, which 85.5667% of instances is correctly classified in the testing. Figure 1 shows that the accuracy has the peak when the confidence factor is around, and it trends to 84.9% as the confidence factor gets close to 1. The relationship of the tree size and confidence factor is indicated in Figure 2. As the confidence factor rises, the tree size is increasing. When the confidence

factor is up to 0.6, the tree size trends to the maximum, which is 377.

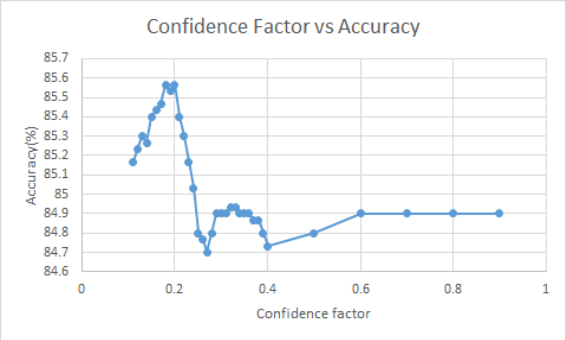

Figure 1: Accuracy with different confidence factors and confidence factor

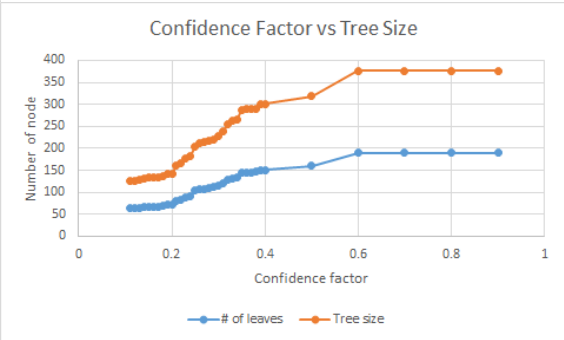

Figure 2: The relationship between tree size and confidence factor

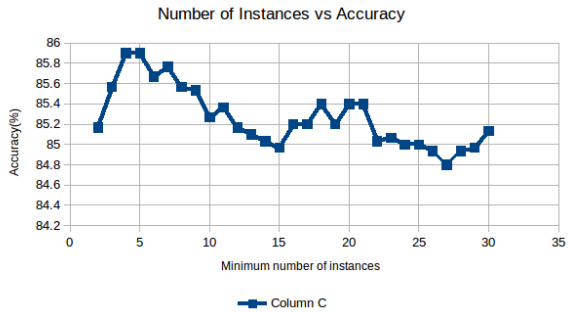

Figure :3 Accuracy affected by minimum number of instances

Discussion

As shown as Figure 2, the larger the confidence factor is, the larger the tree is. If we allow larger error in each node, then less nodes will be pruned. As the result, the tree size will increase with large confidence factor. As for the accuracy, we could see the performance goes up and reaches the peak when confidence factor increases from 0.1 to 0.2 in Figure 1. Since the tree size increases, the decision tree has more space to reduce the entropies for each feature. However, when the confidence factor is over 0.2, the accuracy decreases significantly. This is likely caused by overfitting. Since the decision tree grows too large, it splits many leaves to fit the training dataset. However, the performance of the tree might decrease when it is tested with unseen dataset.

When looking at our decision tree, the root of the tree is occurance of the word “share”. If “share” shows up in the article, the tree will classify it to fake news. We believed that the word “share” is unreliable in the news articles because people usually like to share their opinions instead of facts. In addition, we found that many features related to politics have low entropies, for example, “lawmakers”, “nomination”, and “president”. When these words show up in the articles, the news are usually real according to our decision tree. We thought that these words sound authoritative, and it is easy to trace the source of political news. Therefore, it is difficult for the fake political news to spread in the media.