News Classification

Akhilesh Pothuri
1610110049
Department of Computer Science
Shiv Nadar University
Greater Noida, UP 201314
Email: ap354@snu.edu.in

Rahul Goyal
1610110274
Department of Computer Science
Shiv Nadar University
Greater Noida, UP 201314
Email: rg294@snu.edu.in

October 30, 2019

Abstract

This research investigates the problem of news articles classification. The classification is performed using TD-IDF textual features extracted from text. The application domain is news articles written in English that belong to five categories: Business, politics, entertainment, Science-Technology and Sports down-loaded from the well-known website BBC. Various classification experiments have been performed with the Random Forests machine learning method using TD-IDF textual features. Using the TD-IDF textual features led to much better accuracy results (96%). The main contribution of this work is the introduction of a news article classification framework based on Random Forests and, as well as the late fusion strategy that makes use of Random Forests operational capabilities.

Keywords: Text Classification, Supervised Learning, News Articles, TD-IDF

1 Introduction

Text classification is one of the most important task in Natural Language Processing. It is the process of classifying strings or documents into different categories, depending upon the contents of the string. Text classification methods have drawn much attention in recent years and have been widely used in many programs. These techniques are essential because the textual data is swiftly rising with the passage of time. Text mining tools are required to perform indexing and retrieval of this rapidly growing text data. Text mining is the finding of some information which is previously unknown by extracting that information from large sets of unstructured text. Today, this un-structured data is growing as mostly the information is available in an electronic form such as emails, on World Wide Web, electronic publications and other documents. The term un-structured mean, the type of data in which the text is occurring in a natural free form or a sequence that may include word and sentence ambiguity. This un-structured information cannot be used for further processing by computers. The computers typically handle text as simple sequences of character string and are unable to provide useful information from the given text, without any process performed on it. Therefore, specific processing and preprocessing methods are required in order to extract useful patterns and information from the unstructured text.

News article classification is considered a Document classification (DC) problem. DC means labeling a document with predefined categories. This can be achieved as the supervised learning task of assigning documents to one or more predefined categories. Using machine learning (ML), the goal is to learn classifiers from examples which perform the category classifications automatically. DC is applied in many tasks, such as: clustering, document indexing, document filtering, information retrieval, information extraction and word sense disambiguation. Current-day DC for news articles poses several research challenges, due to the large number of multimodal features present in the document set and their dependencies.

2 Literature Survey

Since in this study we are dealing with supervised machine learning in Document Classification (DC) in general and with news articles classification in particular, we report previous work related to these two fields. Furthermore, since Random Forests (RF) is the machine learning method we use for our proposed classification framework, we provide the theoretical background and related work for this method.

Random Forests (RF) is an ensemble learning method for classification and regression. The basic notion of the methodology is the construction of a group of decision trees. RF employs two sources of randomness in its operational procedures:

1. Each decision tree is grown on a different bootstrap sample drawn randomly from the training data.
2. At each node split during the construction of a decision tree, a random subset of m variables is selected from the original variable set and the best split based on these m variables is used.

For an unknown case, the predictions of the trees that are constructed by the RF are aggregated (majority voting for classification / averaging for regression). For a RF consisting of N trees, the following equation is used for predicting the class label l of a case y through majority voting:

$$l(y) = argmax_c(\sum_{n=1}^{N} I_{h_n}(y) =_c) \tag{1}$$

where I is the indicator function and hn is the n^{th} tree of the RF. RF has an internal mechanism that provides an estimation of its generalization error, called out-of-bag (OOB) error estimate. For the construction of each tree, only 2/3 of the original datas cases are used in that particular bootstrap sample. The rest 1/3 of the instances (OOB data) are classified by the constructed tree and therefore, used for testing its performance.

In this study, we investigate the application of RF for news articles classification. Although the RF have been successfully applied to several classification problems, to the best of our knowledge they havent been applied to news article classification problems. Moreover, an important motivation for using RF was the application of late fusion strategies based on the RF operational capabilities.

3 Classification Methodology

Machine learning algorithms have been successfully utilized in news classification. Machine learning classifiers can be broadly classified as decision trees (such as C4.5, ID3 and Random Forest), rule-based methods (such as RIPPER, PART and genetic algorithms), perceptron-based methods (such as artificial neural networks, radial basis function networks), statistical learning methods (such as Bayesian Networks and Nave Bayes classifier), instance-based classifiers (such as k-nearest neighbour algorithm) and support vector machines. Random Forest Classifiers, Nave Bayes classifier, support vector machines and decision trees are widely employed for text classification problems.

3.0.1 Random-Forest Classifier

Random Forest algorithm (RF) is an ensemble of classification and regression trees induced from bootstrap samples of the training data. In the algorithm, the generalization error of the classifier depends on the power of the individual trees and the association between trees. In the tree induction process, a random feature selection is utilized, which enhances the ability of the model to deal with noisy or irrelevant data. The algorithm yields comparable results to AdaBoost algorithm. The advantages of random forest classifier are-(a) It is robust to correlated predictors, used to solve both regression and classification problems. (b)It can be also used to solve unsupervised ML problems.(c) It takes care of missing data internally in an effective manner.

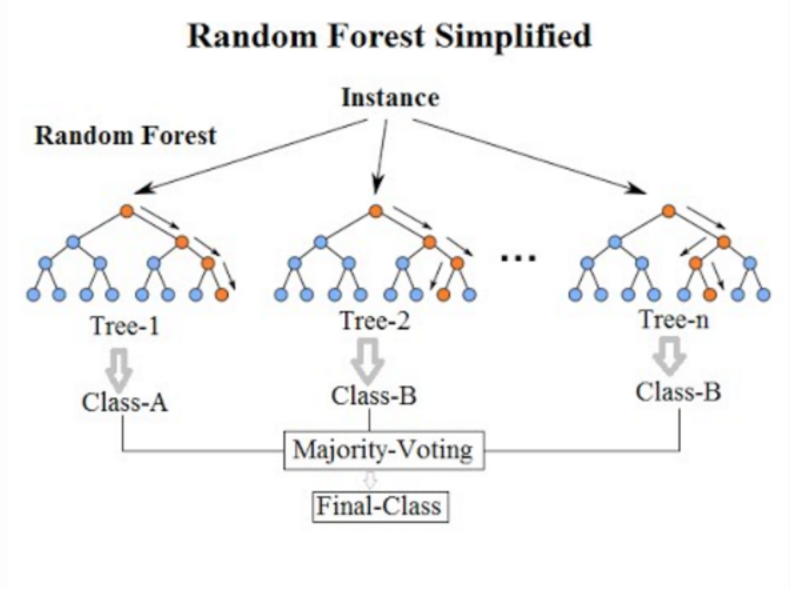


Figure 1: Random Forest Classification

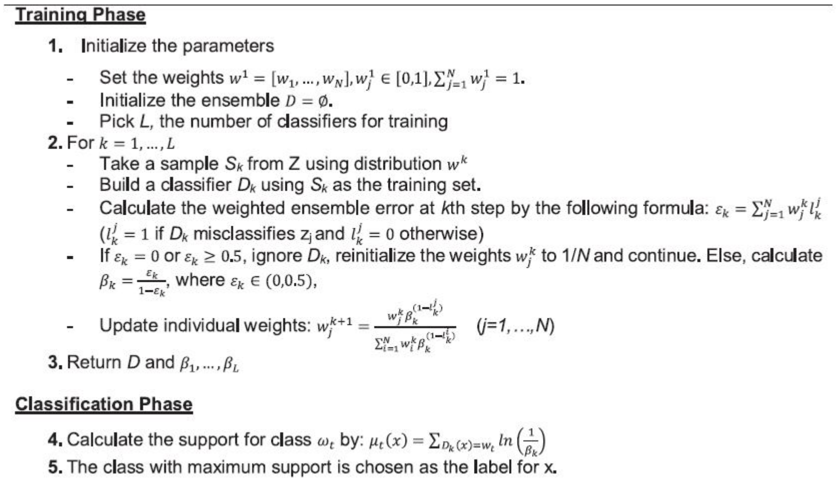
3.0.2 Ada-Boost Classifier

Ada-boost or Adaptive Boosting is one of ensemble boosting classifier proposed by Yoav Freund and Robert Schapire in 1996. It combines multiple classifiers to increase the accuracy of classifiers. AdaBoost is an iterative ensemble method. AdaBoost classifier builds a strong classifier by combining multiple poorly performing classifiers so that you will get high accuracy strong classifier. The basic concept behind Adaboost is to set the weights of classifiers and training the data sample in each iteration such that it ensures the accurate predictions of unusual observations. Any machine learning algorithm can be used as base classifier if it accepts weights on the training set. Adaboost should meet two conditions- (a)The classifier should be trained interactively on various weighed training examples. (b)In each iteration, it tries to provide an excellent fit for these examples by minimizing training error.

Boosting algorithm is a widely employed ensemble learning method to enhance the predictive performance of weak learning algorithms. In this method, a weak learning algorithm is ran recursively on the different sampling distributions of the training data. In this way, a single robust classification model can be built from the weak learning algorithms. AdaBoost (adaptive boosting) algorithm is an ensemble method which improves the boosting algorithm by an iterative process in which more focus is dedicated to the difficult patterns. First, all the patterns in the training set are assigned the same weight value. During the process, the weight values for misclassified instances are in- creased whereas the weights for correctly classified instances are decreased. In this way, the weak learning algorithm dedicates more iterations and classifiers to the patterns that are harder to classify. The stages of AdaBoost algorithm are outlined in Fig. 2

The flowchart of the proposed classification framework (training phase) is depicted in Fig 2. Next, the different steps of the framework are described in detail.

First, all the necessary data is collected in the form of text from news article web pages, as well as images associated to each web page. We note that given the fact that the majority of the web pages contain several images including banners and advertisement logos, it was decided to keep only the biggest image of each site which would most probably be the main image of the article. One other important thing to note is that in this study the features of each modality are treated independently. Hence, two different



```
Training Phase
  1. Initialize the parameters
     - Set the weights $w^1 = [w_1, ..., w_N], w_j^1 \in [0,1], \sum_{j=1}^{N} w_j^1 = 1$.
     - Initialize the ensemble $D = \emptyset$.
     - Pick L, the number of classifiers for training
  2. For $k = 1, ..., L$
     - Take a sample $S_k$ from Z using distribution $w^k$
     - Build a classifier $D_k$ using $S_k$ as the training set.
     - Calculate the weighted ensemble error at kth step by the following formula: $\varepsilon_k = \sum_{j=1}^{N} w_j^k l_k^j$
       ($l_k^j = 1$ if $D_k$ misclassifies $z_j$ and $l_k^j = 0$ otherwise)
     - If $\varepsilon_k = 0$ or $\varepsilon_k \geq 0.5$, ignore $D_k$, reinitialize the weights $w_j^k$ to 1/N and continue. Else, calculate
       $\beta_k = \frac{\varepsilon_k}{1-\varepsilon_k}$, where $\varepsilon_k \in (0,0.5)$,
     - Update individual weights: $w_j^{k+1} = \frac{w_j^k \beta_k^{(1-l_k^j)}}{\sum_{i=1}^{N} w_i^k \beta_k^{(1-l_k^i)}}$   (j=1,...,N)
  3. Return D and $\beta_1, ..., \beta_L$

Classification Phase
  4. Calculate the support for class $\omega_t$ by: $\mu_t(x) = \sum_{D_k(x)=w_t} ln\left(\frac{1}{\beta_k}\right)$
  5. The class with maximum support is chosen as the label for x.
```

Figure 2: The general structure for AdaBoost algorithm

feature vectors (one for each modality) are formulated. In the training phase, the feature vectors from each modality are used as input for the construction of a RF. From the two constructed RFs (one for the textual and one for the visual features), we compute the weights for each modality, in order to apply a late fusion strategy and formulate the final RF predictions. In this study, two different approaches for the computation of the modality weights are followed:

1. From the OOB error estimate of each modalitys RF, the corresponding OOB accuracy values are computed. These values are computed for each class separately. Then, the values are normalized (by dividing them by their sum) and serve as weights for the two modalities.

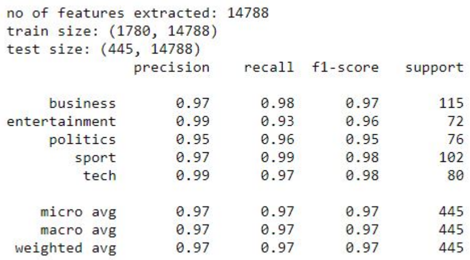
2. For the second weighting strategy, the same procedure as in 1 is applied.

However, instead of employing the OOB accuracy values from each RF, the ratio values between the inner-class and the intra-class proximities (for each class) are used.

First, for each RF the proximity matrix between all pairs of data cases P= $\{p_{ij}, i, j = 1, , w\}$ (w=number of data cases) is constructed and then, the aforementioned ratio values are computed.

4 Results and Conclusions

We have built a number of models to predict the category of news from its headline and short description. Different classification models resulted in different accuracy measure for the same dataset. After preprocessing of data i.e. preprocessing and cleaning of tweets, two feature extraction approach were followed i.e. bag of words and tf-idf method. Our best model achieves on the data set 96.85% accuracy, if considering all the labels. It is interesting how this news dataset is extremely hard to classify for even the most complex models. We attribute this to the subjectivity in category assignment in the data.



```
no of features extracted: 14788
train size: (1780, 14788)
test size: (445, 14788)
               precision    recall  f1-score   support

     business       0.97      0.98      0.97       115
entertainment       0.99      0.93      0.96        72
     politics       0.95      0.96      0.95        76
        sport       0.97      0.99      0.98       102
         tech       0.99      0.97      0.98        80

    micro avg       0.97      0.97      0.97       445
    macro avg       0.97      0.97      0.97       445
 weighted avg       0.97      0.97      0.97       445
```

Figure 3: Results obtained with Random Forests

Future directions for research are:

1. Defining and applying additional various types of features such as: function words, key-phrases, morphological features (e.g.: nouns, verbs and adjectives).

2. Applying various kinds of classification models based on textual and visual features for a larger number of documents that belong to more than four categories in the news articles area, as well as in other areas, applications and languages

5 References

1. Schneider, K. M.: Techniques for improving the performance of naive Bayes for text classification. In Computational Ling

2. Ensemble of keyword extraction methods and classifiers in text classification by Aytu g Onan a , , Serdar Koruko glu b , Hasan Bulut b a Celal Bayar University,

3. Toutanova, K.: Competitive generative models with structure learning for NLP classification tasks. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, (pp. 576-584) (2006)

4. Klassen, M., Paturi, N.: Web document classification by keywords using random forests. In Networked Digital Technologies, pp. 256-261, Springer Berlin Heidelberg (2010)

5. Caropreso, M. F., Matwin, S., Sebastiani, F.: Statistical phrases in automated text categorization. Centre National de la Recherche Scientifique, Paris, France (2000)

6. Aung, W. T., Hla, K. H. M. S.: Random forest classifier for multi-category classification of web pages. In Services Computing Conference, 2009. APSCC 2009. IEEE AsiaPacific, pp. 372-376, IEEE (2009)

7. Gamon, M., Basu, S., Belenko, D., Fisher, D., Hurst, M., Knig, A. C.: BLEWS: Using Blogs to Provide Context for News Articles. In ICWSM (2008)

8. $https://www.kaggle.com/shineucc/bbc-news-dataset$

9. Scikit-learn 0.20.0 documentation. 1.10 Decision Trees. $https://scikit-learn.org/stable/modules/tree.html$

10. A Gentle Introduction to Gradient Boosting,College of Computer and Information Science,Northeastern University.

11. Random Decision Forests,Tin Kam Ho,AT & T Bell Laboratories