An Efficient Method for Automatic Text Categorization

Mohammad Behrouzian Nejad¹*, Iman Attarzadeh² and Mehdi Hosseinzadeh³

¹Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Kerman, Iran
²Department of Computer Engineering, Dezful Branch, Islamic Azad University, Dezful, Iran
³Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

*Corresponding Author's E-mail: m.behrouzian@srbiau.ac.ir

Abstract

Automatic Text Categorization refers to assigning uncategorized text documents to one or more predefined categories. Texts categorization generally divided into two main sections: feature selection and learning algorithm. For Feature selection and learning algorithms techniques, various methods have been proposed. The purpose of the proposed techniques, increasing the accuracy of classification and achieve optimal performance. In this paper a hybrid method is proposed which uses Filtering feature selection technique to reduce the complexity and works on combining classifiers outputs. The proposed method is homogeneous and uses uniform classifiers with different sampling with replacement from the training set. The results show the superiority of the proposed method compare to Naïve Bayes and j48 classifier and some related works according to the criteria of accuracy, precision, recall, F1 and classification error.

Keywords: Data Mining, Text Mining, Automatic Text Classification, Feature Selection, Learning Algorithm.

1. Introduction

Data mining is one of the most recent advances to serve data management technologies. Text mining is the one of the most important areas of data mining [1,2]. Main operation of text mining is consists of extracting knowledge from text. The
most important issues facing the text mining is automatic text categorization. Automatic text categorization means of assigned text documents available to pre-
define several categories that documents that belong to them [3]. To do this work,
must first identify the categories that usually it is done by experts, then documents
must be determined for each category. The main purpose finding real category of
collected text documents. With the increasing use of the Internet and electronic
documents, automatic text categorization has become imperative. Several methods
have been proposed for the text categorization.

Among the proposed methods can be cited text categorization based on
unorganized data with extracted information [4], text classification based on features
[5], Text Categorization using PDDP with Support Vector Machines (SVM) [6], text
categorization using K-Nearest Neighbor (KNN) [7], Improved KNN Algorithm by
Optimizing Cross-validation [8], Improved KNN based on clustering [9], improved
KNN using ant colony [10], Naïve Bayes method [11, 12], text categorization using
association rules [13], … that their purpose are improving the accuracy and efficiency
of classification.

The purpose of this paper providing a method for automatic text categorization that
can classify the text documents with high performance. The results show the
superiority of the proposed method compared heterogeneous classifier and single
classifiers according to the criteria of accuracy, precision, recall, F1 and classification
error. The reminder of this paper is organized as follows. In Section 2, we review the
related works. Process of text categorization is discussed in sections 3. Section 4
describes the proposed method. Sections 5 and 6 describe the implementation and
evaluation of proposed method. Finally, Sections 7 and 8 contains the conclusion and
references of the paper.
2. Related Works

KNN algorithm is one of the approaches widely used in text classification because is easy to implement. One of the problems in K nearest neighbor algorithm is determine the proper value of the parameter K, which is necessary for the performance of a classification. The parameter K is usually more than the number of classes and a odd number is considered. According to the KNN is a lazy learning, which means that hold all the training samples to the end of classification. With increasing parameter K, increases the complexity of KNN which is not desirable. In order to solve this problem in [14] a hybrid classification method using two classifiers KNN and SVM are presented. In order to reduce training time, for each category first SVM is used for classification.

Then support vectors of different categories as training data are given with KNN classifier. To calculate the average distance between the test data and each support vector, the Euclidean distance is used. The final decision will be based on batch which its vector have minimum distance supporter with test data. Extracted results show the efficiency of this method. In [15] an improved KNN algorithm for text categorization is proposed. In this method, classification model is built using a clustering algorithm which for clustering uses from minimum distance for divides learning textual examples into hyper Koreas with the same radius.

Then used KNN method for classifying sets of experiments based on the model. This method significantly reduces the computational complexity and dynamically updates the classification model. The results indicate that this method works best from methods such as KNN, NB and SVM. In [16] the combination of two classifiers are used as serial categories, that the first classifier are candidates categories for new document, then second classifier, select final class of new documents between classes. This method with 2850 documents (training and testing) is investigated. The results show that the performance
improvement which due to the low number of documents obtained, the reliability is not high enough.

3. Process of Text Categorization

The main steps of the process of text categorization can be classified into three main stages of preprocessing, feature selection and classification stage [17].

3.1. Preprocessing

In the preprocessing stage, usually on the input data operations are separating words, removal of redundant words (stop words), stemming and term (Feature) weighting.

3.2. Feature Selection

This step refers to select a subset of features of the text. In general, feature selection techniques are classified into two general categories: filtering and wrapper methods. Filtering methods are independent of the learning algorithm. These methods Regardless of learning algorithm and using statistical methods to do feature selection and have time complexity lower than the wrapper methods. Wrapper methods uses from learning algorithm as the evaluation function. These methods have higher time complexity and accuracy than filter methods [18]. With the increasing size of the features in text classification, generally these methods could not be used because of the high complexity. Some filtering methods that can be used in many texts classification techniques such as Document Frequency (DF), Information Gain (IG), and Mutual Information (MI) [19]. One of the most versatile and popular methods is information gain method which used in many studies and has good results. Information gain value measures \"the number of bits of information obtained for category prediction by knowing presence of absence of a term in a document\". Information gain values were calculated as (1) which $P(t, c)$ shows number of
text documents in category $c$ which have term $t$ and $p(\bar{t}, c)$ shows number of text documents in category $c$ which have nor term $t$ [19]:

$$
\lg(t_k, c_l) = \sum_{c \in [c_l \cup c]} \sum_{t \in (t_k, \bar{t}_k)} p(t, c) \log_2 \frac{p(t, c)}{p(t)p(c)}
$$

(1)

3.3. Learning Algorithm

In this step, classifiers from preprocessed text, act to learning [17, 20].

3.3.1. Naïve Bayes: One of the most commonly methods for text classification is Naïve Bayes algorithms. This algorithm is based on conditional probabilities and uses the Bayes theorem. In this approach to classify a new example, the probability of all categories for new example calculated and then category with the highest probability is selected as a category of new example. It is one of the fastest methods.

3.3.2. Decision Tree: decision tree is a tree which internal nodes represent features, outputted edges are feature selection criteria and leaf nodes represent classes. Construct of decision tree has two phases: growth and pruning. In growth phase a decision tree is built from the training data. In The pruning phase, section of the tree is pruned so that the test is not done on the branch. Decision tree classification based on feature selection criteria are divided into two categories CART and C4.5. CART is classification and regression algorithm. In this study, we used from the j48 decision tree that construct the pruning or not pruning tree of C4.5.

3.3.3. Support Vector Machine: This algorithm is one of the most popular classification algorithms which in recent years had a pretty good performance. This algorithm is based on classification of linear data. In the dividing line data, try to choose a line that is more confident border than the other lines. Optimum choice for line data, can done QP methods
which related to solve restrictions problems. In [21] linear and nonlinear methods are tested which linear method is little better than nonlinear.

4. Proposed Method

One way to improve classification performance is use combination of classifiers. Use of combination of classifiers by combining multiple alone classifier can improve performance. Conducted Researches shows by use of combination of classifiers, can improve performance [22]. The proposed method combines the outputs of classifiers and because they do not need to know the structure of classifiers and its feature vectors, it is more common [23]. The proposed method is homogeneous and uses uniform classifiers with different sampling with replacement from the training set. For this work we use from bootstrapping method [24-27]. Fig. 1 shows the proposed method.
In the proposed method, input documents are in the preprocessing. Feature selection is carried out at next stage. Then we use of different sampling with replacement from the training set, several new training set with the initial size obtained and use by each classifiers. Then classifiers outputs are combine. In combine stage we use majority vote. Last stage is performance evaluation by different criteria.
5. Implementation

5.1. Implementation Tools

In this paper for implement proposed method we use from Rapid Miner version 5.2 [28]. This software is an open source data mining tools and written by Java language. Evaluation runs on a system with windows operating system, 2 GHz of CPU and 2 GB of memory.

5.2. Dataset

In this paper we use from reuters-21578 dataset [29]. We use from R (8) subset of this dataset [30]. This data set contains 8 main category and have 7674 text documents. Each document related to a category. Details of R (8) subset of reuters-21578, shows in table 1.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Train documents</th>
<th>Test documents</th>
<th>Total documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acq</td>
<td>1596</td>
<td>696</td>
<td>2292</td>
</tr>
<tr>
<td>Crude</td>
<td>253</td>
<td>121</td>
<td>374</td>
</tr>
<tr>
<td>Earn</td>
<td>2840</td>
<td>1083</td>
<td>3923</td>
</tr>
<tr>
<td>Grain</td>
<td>41</td>
<td>10</td>
<td>51</td>
</tr>
<tr>
<td>Interest</td>
<td>190</td>
<td>81</td>
<td>271</td>
</tr>
<tr>
<td>money-fx</td>
<td>206</td>
<td>87</td>
<td>293</td>
</tr>
<tr>
<td>Ship</td>
<td>108</td>
<td>36</td>
<td>144</td>
</tr>
<tr>
<td>Trade</td>
<td>251</td>
<td>75</td>
<td>326</td>
</tr>
</tbody>
</table>

5.3. Preprocessing

This step in the proposed method contains: Transform Case, Tokenization, Filter Stop Words, Stemming and Generate n-Gram.
Transform Case: in this step, all characters of text are converted to the same form. At this stage, all the characters are converted to lowercase.

Tokenization: At this stage, the whole text is divided into separate successive words.

Filter Stop Words: This step will eliminate redundant words like and, the, for and ….

Stemming: At this stage for eliminate prefixes and suffixes of words, Porter algorithm are used.

Generate n-Gram: At this stage for indexing and reduce the dimension text n-gram method is used. With using n-gram we can make the text as a series of consecutive words with length n. This model was originally proposed for speech processing issues. But now many different versions of this model have been proposed for text classification problems in natural language processing [31-34]. The experiments performed on different values of n and to avoid increasing the complexity of the n-gram with n = 2 was used.

After these steps, the weighted features are performed. The Tfidf Term weighting method used in this paper [3].

5.4. Feature Selection

In this study, for increasing the efficiency and reducing the complexity, information gain feature selection method used in the feature selection stage.

5.5. Learning Algorithm

The proposed method using libsvm, is implemented. Libsvm which Library Support Vector Machines also say that, based on support vector machines and Java
libraries which have been developed by [35]. The proposed methods are compared with naïve bayes and j48.

5.6. Evaluation Criteria

This section will examine the evaluation criteria for the text classification [3]. Different status of categories and documents according to input dataset to classification with TP, FP, TN, FN values for two Negative and Positive categories shows as table 2.

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Negative</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

Were FP is the number of documents which incorrectly classified under Positive ; TN number of documents which correctly classified under Negative, TP is number of documents which correctly classified under Positive, and FN is number of documents which incorrectly classified under Negative. According to parameters in table 2., different evaluation criteria like Accuracy, Classification Error, Precision, Recall and F1 are presented. How to calculate of these measures are shown (2) to (6) respectively.

\[
Accuracy = \frac{TP + TN}{TP + FP + FN + TN}
\]  

\[
ER = \frac{FN + FP}{TP + FP + FN + TN} = 1 - Accuracy
\]
Were index represent that these parameters should be calculate for each category $i$.

### 6. Evaluation of Proposed Method

At this step, the proposed method has been evaluated with naïve bayes and j48 using different criteria. All result are per percent. Table 3. shows the results of proposed method, Table 4. shows the results of naïve bayes classifier and Table 5. shows the results of j48 classifier.

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acq</td>
<td>99.58</td>
<td>99.73</td>
</tr>
<tr>
<td>Trade</td>
<td>99.88</td>
<td>100</td>
</tr>
<tr>
<td>Ship</td>
<td>100</td>
<td>99.60</td>
</tr>
<tr>
<td>Interest</td>
<td>97.35</td>
<td>98.07</td>
</tr>
<tr>
<td>Grain</td>
<td>100</td>
<td>98.99</td>
</tr>
<tr>
<td>Crude</td>
<td>99.78</td>
<td>99.83</td>
</tr>
<tr>
<td>Earn</td>
<td>99.92</td>
<td>99.81</td>
</tr>
<tr>
<td>money-fx</td>
<td>97.45</td>
<td>97.31</td>
</tr>
</tbody>
</table>

**Avg-precision: 99.25**

**Avg-recall: 99.17**

**F1: 99.20**

**Accuracy: 99.63**

**classification_error: 0.37**
### Table 4: the results of naïve bayes classifier

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acq</td>
<td>84.62</td>
<td>90.35</td>
</tr>
<tr>
<td>Trade</td>
<td>77.19</td>
<td>70.12</td>
</tr>
<tr>
<td>Ship</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>Interest</td>
<td>73.96</td>
<td>74.74</td>
</tr>
<tr>
<td>Grain</td>
<td>17.72</td>
<td>34.15</td>
</tr>
<tr>
<td>Crude</td>
<td>62.59</td>
<td>72.73</td>
</tr>
<tr>
<td>Earn</td>
<td>99.31</td>
<td>90.88</td>
</tr>
<tr>
<td>money-fx</td>
<td>67.94</td>
<td>68.93</td>
</tr>
</tbody>
</table>

**Avg-precision:** 64.30  
**Avg-recall:** 68.90  
**F1:** 66.52  
**Accuracy:** 86.33  
**classification_error:** 13.67

### Table 5: the results of j48 classifier

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acq</td>
<td>85.59</td>
<td>88.97</td>
</tr>
<tr>
<td>Trade</td>
<td>81.93</td>
<td>81.27</td>
</tr>
<tr>
<td>Ship</td>
<td>36.26</td>
<td>30.56</td>
</tr>
<tr>
<td>Interest</td>
<td>77.20</td>
<td>78.42</td>
</tr>
<tr>
<td>Grain</td>
<td>13.64</td>
<td>7.32</td>
</tr>
<tr>
<td>Crude</td>
<td>83.53</td>
<td>84.19</td>
</tr>
<tr>
<td>Earn</td>
<td>95.59</td>
<td>94.61</td>
</tr>
<tr>
<td>money-fx</td>
<td>67.32</td>
<td>66.99</td>
</tr>
</tbody>
</table>

**Avg-precision:** 67.88  
**Avg-recall:** 66.55  
**F1:** 67.20  
**Accuracy:** 88.37  
**classification_error:** 11.63
Table 6. shows the results of proposed method compare to some related works. Result show that our method has better performance than Naïve Bayes and j48 Classifiers and related works according to the criteria of accuracy, precision, recall, F1 and classification error.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Precision</td>
<td>99.25</td>
<td>91.33</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Average Recall</td>
<td>99.17</td>
<td>91.15</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Average F1</td>
<td>99.20</td>
<td>91.23</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Accuracy</td>
<td>99.63</td>
<td>-</td>
<td>98.05</td>
<td>90.04</td>
</tr>
<tr>
<td>Cls.Error</td>
<td>0.37</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Conclusion

According to this note that one way to improve classification performance is use combination of classifiers; in this paper we proposed a hybrid method which uses filtering method for feature selection technique to reduce the complexity and combine classifiers outputs. The proposed method is homogeneous and uses uniform classifiers with different sampling with replacement from the training set. The results show the superiority of the proposed method compared naïve bayes and j48 classifiers and some references according to the criteria of accuracy, precision, recall, F1 and classification error.
References


