Feature Selection using Improved Mutual Information for Text Classification

Jana Novovičová[1,2], Antonín Malík[1,3], and Pavel Pudil[1,2]

[1]Institute of Information Theory and Automation,
Department of Pattern Recognition, Academy of Sciences of the Czech Republic,
Prague, Czech Republic
[2] The University of Economics, Faculty of Management,
Prague, Czech Republic
[3] Czech Technical University, Faculty of Electrical Engineering,
Prague, Czech Republic
{novovic,amalik,pudil}@utia.cas.cz

Abstract. A major characteristic of text document classification problem is extremely high dimensionality of text data. In this paper we present two algorithms for feature (word) selection for the purpose of text classification. We used sequential forward selection methods based on improved mutual information introduced by Battiti [1] and Kwak and Choi [6] for non-textual data. These feature evaluation functions take into consideration how features work together. The performance of these evaluation functions compared to the information gain which evaluate features individually is discussed. We present experimental results using naive Bayes classifier based on multinomial model on the Reuters data set. Finally, we analyze the experimental results from various perspectives, including F_1-measure, precision and recall. Preliminary experimental results indicate the effectiveness of the proposed feature selection algorithms in a text classification problem.

1 Introduction

The goal of text document classification is to assign automatically a new document into one or more predefined classes based on its contents.

An increasing number of statistical classification methods and machine learning algorithms have been explored to build automatically a classifier by learning from previously labelled documents including *naive Bayes*, *k-nearest neighbor*, *support vector machines*, *neural network*, *decision trees*, *logistic regression* (see e.g. [7], [9], [5], [11], [12] and the references therein).

In text classification, usually a document representation using a *bag-of-words* approach is employed (each position in the feature vector representation corresponds to a given word). This representation scheme leads to very high-dimensional feature space. *Feature selection* is a very important step in text classification, because irrelevant and redundant words often degrade the performance of classification algorithms both in speed and classification accuracy.

Methods for feature subset selection for text document classification task use an evaluation function that is applied to a single word. All words are independently evaluated and sorted according to the assigned criterion. Then, a predefined number of the best features is taken to form the best feature subset. Scoring of individual words can be performed using some of the measures, for instance, *document frequency*, *term frequency*, *mutual information*, *information gain*, *odds ratio*, χ^2 *statistic* and *term strength* [10], [8], [3]. Yang and Pedersen [10] and Mladenic [8] give experimental comparison of the above mentioned measures in text classification. The information gain (IG) and a very simple frequency measures were reported to work well on text data. Forman in [3] presents an extensive comparative study of twelve feature selection criteria for the high-dimensional domain of text classification.

In this paper we propose to use sequential forward selection methods based on improved mutual information introduced by Battiti [1] and Kwak and Choi [6], who introduced these criteria for non-textual data. To our knowledge, the improved mutual information has not yet been applied in text classification as a criterion for reducing vocabulary size. We use the simple but effective naive Bayes classifier based on multinomial model.

2 Naive Bayes Classifier

According to the bag-of-words representation, the document d_i can be represented by a feature vector consisting of one feature variable for each word w_t in the given vocabulary $V = \{w_1, \ldots, w_n\}$ containing n distinct words. Let $C = \{c_1, \ldots, c_{|C|}\}$ be the set of $|C|$ classes. Note, that $|C|$ classes are pre-defined and that document always belongs to at least one class. Given a new document d, the probability that d belongs to class c_j is given by Bayes rule

$$P(c_j|d) = \frac{P(c_j)P(d|c_j)}{P(d)}. \tag{1}$$

If the task is to classify a new document into a single class, simply select the class $c^\star$ with the highest posterior probability.

Assuming a multinomial model [7, 9] and class-conditional independence of words yields the well-known naive Bayes classifier, which computes the most probable class for d as

$$c^\star = \operatorname*{argmax}_{j=1,\ldots,C} P(c_j|d) = \operatorname*{argmax}_{j=1,\ldots,C} P(c_j) \prod_{t=1}^{n} P(w_t|c_j)^{N(w_t,d)} \tag{2}$$

where $N(w_t, d)$ is the number of occurrences of word w_t in document d. The word probability $P(w_t|c_j)$ are usually estimated using Laplacean prior:

$$P(w_t|c_j) = \frac{1 + \sum_{d_i \in c_j} N(w_t, d_i)}{|V| + \sum_{r=1}^{|V|} \sum_{d_i \in c_r} N(w_t, d_i)} \ . \tag{3}$$

The class priors $P(c_j)$ are estimated by maximum-likelihood estimates the fraction of documents in each class.

2.1 Classifier estimation

For evaluating the multi-label classification accuracy we use the standard multi-label measures: precision, recall and F_1 measure. Precision and recall are computed as

$$precision = \frac{\#\ classes\ found\ and\ correct}{\#\ total\ classes\ found}$$

$$recall = \frac{\#\ classes\ found\ and\ correct}{\#\ total\ classes\ correct}$$

where "classes found" means classes c_k with $P(c_k|d) \geq h$. To obtain the single number measure of classification performance we compute the F_1 measure that combine both the precision p and recall r

$$F_1 = \frac{2pr}{p+r}. \tag{4}$$

The closer are the values of precision and recall, the higher is the F_1 measure.

In the case of multi-label classification the document d is classified in the class c_k if the probability $P(c_k|d) \geq h$. The threshold h is estimated to maximize F_1 measure (4) on the training data set. The threshold h_j shifts from 0 to 1 and for each potential value of h_j we make the classification process on the training data set. All training documents d_i are classified according to the equation $P(c_k|d_i) \geq h_j$ and the F_1 measure is computed for each h_j. Then the threshold h with the highest F_1 value is selected.

Given a new document d, the probability $P(c_j|d)$ is computed by applying Bayes rule (1). If the probability $P(c_j|d) > h$, than the document d is assigned to the class c_j. Therefore, the document d can be assigned to one or more classes. If $P(c_j|d) < h$ for each class c_j, the document d is classified in the class with highest probability $P(c_j|d)$.

While the word independence assumption is false in practice with real-world data, there is empirical evidence that the naive Bayes yields surprisingly good classification performance on text data.

3 Feature Selection

Feature subset selection is commonly used when learning on text data, since text documents are characterized by high-dimensionality feature vector.

The focus of this paper is the comparison between *best individual features* and *sequential forward selection* methods. Both methods are based on mutual information $I(C, w_i)$ between classes C and word w_i that is commonly named *information gain* (IG) in text classification

$$I(C, w_i) = \sum_{k=1}^{|C|} P(c_k, w_i) \log \frac{P(c_k, w_i)}{P(c_k)P(w_i)} + \sum_{k=1}^{|C|} P(c_k, \overline{w}_i) \log \frac{P(c_k, \overline{w}_i)}{P(c_k)P(\overline{w}_i)} \quad (5)$$

where $P(w_i)$ is the probability, that the word w_i occurred, $\overline{w_i}$ means, that the word not occurred, $P(c_j)$ is the probability of the class c_j, $P(c_j, w_i)$ is the joint probability of the class c_j and the occurrence of the word w_i.

3.1 Best individual features

Best individual features (BIF) methods [4] evaluate all the n words individually according to a given criterion, sort them and select the best k words.

Since the vocabulary has usually several thousands or tens of thousands of words, the BIF methods are popular in text classification because they are rather fast, efficient and simple. However, they evaluate each word separately and completely ignore the existence of other words and the manner how the words work together. In [2] it has been proven that the best pair of features need not contain the best single features.

Scoring of individual features can be performed using some of the measures, for instance, *document frequency*, *term frequency*, *mutual information*, *information gain*, χ^2 *statistic* or *term strength.* Yang and Pedersen [10] give experimental comparison of the above mentioned measures in text classification. They found information gain and χ^2 statistic most effective in word selection.

In our comparison we include BIF method with information gain criterion (BIF IG) defined in (5).

3.2 Sequential forward selection

Sequential forward selection (SFS) methods firstly select the best single word evaluated by given criterion. Then, add one word at a time until the number of selected words reaches desired k words. However SFS methods do not result in the optimal words subset but they take note of dependencies between words as opposed to the BIF methods. Therefore SFS often give better results than BIF. The similar strategy is *sequential backward selection* that starts with all n words and successively remove one word at a time.

SFS are not usually used in text classification because of their computation cost due to large vocabulary size. However, in practice we can often both employ calculations from previous steps and make some pre-computations during the initialization. Since feature selection is typically done in an off-line manner, the computational time is not as important as the optimality of words subset or classification accuracy.

We propose two SFS methods based on mutual information (SFS MI) introduced by Battiti [1] and Kwak and Choi [6]. They sufficiently applied these two methods for non-textual data compared with BIF. In contrast to BIF IG, SFS MI uses not only mutual information $I(C, w_i)$ between the set of classes C and a word w_i but also mutual information $I(w_i, w_j)$ between the words w_i and w_j. The SFS MI algorithm is described in the following steps:

1. **Initialization:**
   the set of selected words $S = \emptyset$,
   the set of unselected words $U =$ 'all n words'.

   Pre-computation:
   $I(C, w_i)$ for $i = 1, \dots, n$,
   I_{ij}, for $i, j = 1, \dots, n$ and $i \neq j$
   - Battiti: $I_{ij} = I(w_i, w_j)$
   - Kwak-Choi: $I_{ij} = I(w_i, w_j) I(C, w_j) / H(w_j)$
2. **First word selection:**
   Find the word $w^\star$ with maximal $I(C, w_i)$,
   $w^\star = \arg\max_{i=1,\dots,n} I(C, w_i)$,
   set the sets $S = \{w^\star\}$, $U = U \setminus \{w^\star\}$.
3. **One step:**
   Repeat until the demand k words are selected ($|S| = k$).
   Choose the best word $w^\star$ from the set U.
   $w^\star = \mathrm{argmax}_{i=1,\dots,|U|} \{I(C, w_i) - \beta \sum_{j=1}^{|S|} I_{ij}\}$
   Set the sets $S = S \cup w^\star$ and $U = U \setminus w^\star$.

$H(w_j)$ is the entropy of the word w_j. The variable $\beta \geq 0$ is typically set to 1. The higher β the stronger impact of the mutual information between words. On the other hand if $\beta = 0$, then the mutual information between words is not considered and the algorithm coincides with the BIF IG selection.

4 Experimental Results

In our experiments we compared the performance of three feature selection methods. The first method is standard BIF algorithm using the IG criterion. Each word is evaluated by IG criterion and then are selected the first best k words. The other two SFS methods Battiti SFS MI [1] and Kwak-Choi SFS MI [6] are based on mutual information criterion between the set of classes C and the word w_i as well as BIF IG method. Moreover SFS MI consider the mutual information between each pair of words and add one word in each step.

All experiments were tested for different number of words on the common used Reuters[1] data set. Since Reuters documents are multi-labelled we employed

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578.

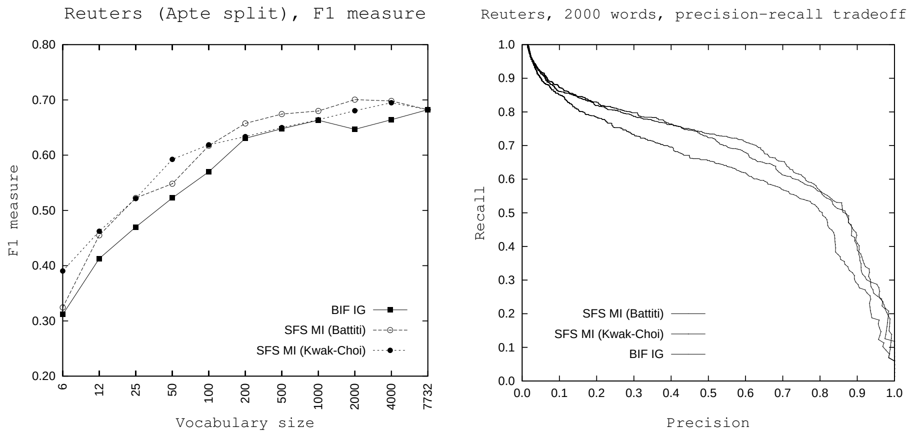

Fig. 1. (a) F_1 measure of BIF IG, SFS MI (Battiti) and SFS MI (Kwak and Choi) on Reuters data set with Apte split. (b) Identical feature selection methods with (a) but precision-recall tradeoff on 2000 selected words. The highlighted points in (b) show the maximal F_1-measure.

the *micro-average F_1-measure* and the *precision-recall tradeoff* for evaluating performance.

First, we displaced all unlabelled documents from the data. Second, we removed all uninformative words occurring in stop-list, such as prepositions, conjunctions or articles. Then, Porter stemming algorithm[2] was used. Finally we deleted the words that occurred only once or twice. The data resulted in 7732 words and 11280 documents in 118 classes. We divided this data set in the training and the testing set according to the usually used Apte split.

For classification was used the naive Bayes classifier based on the multinomial model. In addition to training standard parameters, the threshold h for multilabel classification was made to maximize F_1 measure on the training data set.

Figure 1 (a) shows the comparison of all three FS methods on F_1 measure. We can see that both observed SFS MI methods significantly overcome the BIF IG algorithm on the Reuters data. Compared with the BIF IG, the F_1 value of SFS MI algorithms is with some vocabulary sizes even greater than with the full number of 7732 words.

The highest value of F_1 is achieved on 2000 words with the Battiti SFS MI algorithm. The precision-recall tradeoff on 2000 words is depicted on the Figure 1 (b). Figures 2 (a) and (b) presents the similar result like Figure 1 (a) but on the precision and recall measure.

The Kwak-Choi SFS MI has approximately higher value of F1, precision and recall than the Battiti SFS MI on the lower number of words. However, on the greater number of words the Battiti SFS MI overcome it with all three measures.

[2] http://www.tartarus.org/˜martin/PorterStemmer.

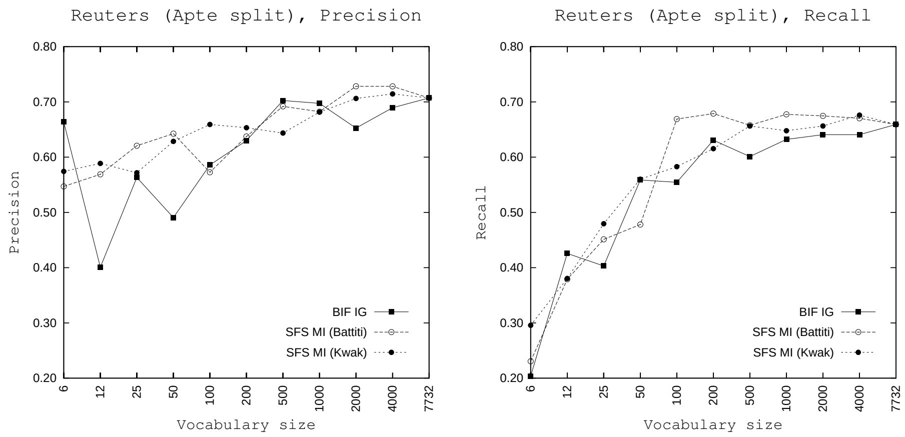

Fig. 2. (a) Precision and (b) recall of BIF IG, SFS MI (Battiti) and SFS MI (Kwak and Choi) on Reuters data with Apte split. The same threshold h was used as in the figure 1.

The time complexity of SFS algorithms is less than $O(kn^2)$ where k is the number of desired words and n is the total number of words. The algorithm adds step by step k words and in each step compute the mutual information between each word belonging to the set S (selected words) and each word from the set U (unselected words). The required space is $n^2/2$ because we need to store the mutual information between each pair of words. If we compare the BIF and SFS methods, the SFS methods are more time consuming but achieve significantly better results on the testing data.

5 Conclusions and Future Work

In this paper, we have presented sequential forward selection methods based on novel improved mutual information measure. The algorithms are new in the field of text classification and take into consideration how the features work together. These methods significantly overcome standard best individual features method based on information gain on the testing data set.

Many areas of future work remain. Ongoing work includes comparison on the other text classifiers, for example, support vector machines and k-nearest neighbor.

Acknowledgements

The work has been supported by grants of the Grant Agency of the Academy of Sciences of the Czech Republic No A2075302, the Grant Agency of the Czech Republic No 102/03/0049 and the Complex research project of the Academy of Sciences of the Czech Republic K1019101.

References

1. Battiti, R.: Using Mutual Information for Selecting Features in Supervised Neural Net Learning. IEEE Trans. Neural Networks **5** (1994) 537–550
2. Cover, T.M.: The Best Two Independent Measurements are not The Two Best. IEEE Trans. Systems, Man, and Cybernetics **4** (1974) 116–117
3. Forman, G.: An Experimental Study of Feature Selection Metrics for Text Categorization. Journal of Machine Learning Research **3** (2003) 1289–1305
4. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical Pattern Recognition: A Review. IEEE Trans. on Pattern Analysis and Machine Intelligence **22** (2000) 4–37
5. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Proceedings of the ECML'98 (1998) 137–142
6. Kwak, N., Choi, C.: Improved Mutual Information Feature Selector for Neural Networks in Supervised Learning. In: Int. Joint Conf. on Neural Networks (IJCNN '99) (1999) 1313–1318
7. McCallum, A., Nigam, K.: A Comparison of Event Models for Naive Bayes Text Classification. In: Proceedings of the AAAI-98 Workshop on Learning for Text Categorization (1998) 41–48
8. Mladenic, D., Grobelnik, M.: Feature Selection for Unbalanced Class Distribution and Naive Bayes. In: Proceedings of the Sixteenth International Conference on Machine Learning (1999) 258–267
9. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text Classification from Labelled and Unlabelled Documents Using EM. Machine Learning **39** (2000) 103–134
10. Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of the 14th ICML97 (1997) 412–420
11. Yang, Y.: An Evaluation of Statistical Approaches to Text Categorization. Journal of Information Retrieval **1** (1999) 67–88
12. Yang, Y., Zhang, J., Kisiel, B.: A Scalability Analysis of Classifier in Text Categorization. In: Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval (2003) 96–103