Research on Spam Filtering Technology Based on New Mutual Information Feature Selection Algorithm

Kunfu Wang\textsuperscript{1,*}, Wanfeng Mao\textsuperscript{1}, Wei Feng\textsuperscript{1} and Hui Wang\textsuperscript{2}

\textsuperscript{1} System Engineering Research Institute of China State Shipbuilding Corporation, Beijing, China. 100036
\textsuperscript{2} School of computer science and technology, Henan Polytechnic University, Jiaozuo, 454000

\textsuperscript{*}Corresponding Author Email: wangkunfu1166@163.com

Abstract. Aiming at the deficiency of traditional mutual information algorithm in feature selection, this paper proposes a weighted naive Bayesian algorithm based on improved mutual information, called imi-wnb algorithm. In the feature selection stage, the word frequency factor and the difference factor between classes are introduced to improve the traditional mutual information algorithm to achieve feature dimension reduction. In the process of classification, the value of IMI is introduced to weight the attributes of naive Bayes algorithm, which partly eliminates the influence of conditional independence assumption of naive Bayes algorithm on classification, and improves the efficiency and stability of spam classification.

1. Introduction
The birth of e-mail has brought unprecedented convenience to users' communication, but spam has also been produced. According to the global spam data in 2018 released by Kaspersky information security website securelist, China has become the world's first source of spam, accounting for 11.69% of the global spam sources [1]. Spam occupies a large number of network bandwidth and mailbox space, resulting in network congestion. Spam contains some malicious software and phishing websites, which brings huge economic losses to users. Therefore, the research on spam filtering is of great significance.

At present, the main filtering technologies include black and white list filtering technology, behavior pattern recognition technology and content-based filtering technology. Content based filtering technology is more feasible and less expensive, and has become the mainstream research direction of spam filtering methods [2-3]. Content based filtering technology mainly includes support vector machine (SVM), k-nearest neighbor (KNN), naive Bayes (NB), etc. [4]. Naive Bayes classifier is simple to implement and has high accuracy, which has become a widely used classification method for spam filtering [5]. Naive Bayes classification is based on conditional independence hypothesis, that is, the hypothesis conditions are completely independent, which affects the accuracy of classification results to a certain extent.

The advantages and disadvantages of feature selection algorithms before spam classification will affect the classification effect. Several common feature selection algorithms are document frequency (DF), information gain (Ig), TF-IDF, square root fit test (test) and mutual information (MI). Among them, mutual information has poor effect, but the method complexity is low and easy to understand, which is a widely used feature selection method [6-7]. However, the traditional mutual information method does not calculate the frequency of feature words, and sometimes the mutual information value of low-frequency words is higher, which affects the classification accuracy [8-9].
Aiming at the problem of feature redundancy and independence assumption, many scholars have improved the feature selection and classification algorithm to improve the accuracy of mail classification. Mishra et al. [10] applied naive Bayesian, random tree and random forest machine learning algorithms to spam data sets, and the classification accuracy was higher than that of pure Bayesian classifier. Elssied et al. [11] proposed a mail classification algorithm combining support vector machine algorithm with K-means clustering algorithm, which improves the classification accuracy and reduces the training time.

Wu Jianjun et al. [12] proposed to apply mutual information to weighted naive Bayes. The influence of naive Bayes conditional independence assumption on classification effect was eliminated by weighting part, and the text classification effect of naive Bayes was improved. However, the traditional mutual information algorithm was still used for weighting in this method, and the traditional mutual information algorithm was not improved. Yang Lei et al. [13] proposed a tsvm-nb algorithm, which uses naive Bayes algorithm for initial training, and then uses support vector machine algorithm to construct optimal classification hyperplane to reduce the dimension of feature items. Finally, naive Bayes algorithm is used to generate classification model again, which improves the speed and accuracy of spam filtering. However, this algorithm is suitable for the corpus with large overlapping attribute vectors, and its efficiency improvement is limited on the corpus with weak aliasing. Tao Yongcai et al. [14] introduced the idea of entropy and combined with map reduce technology, proposed an improved mutual information text feature selection mechanism based on map reduce, which improved the accuracy of text classification. Li et al. [15] proposed a parallel feature selection method based on MapReduce, using the maximum mutual information theory to select the feature variable combination with the most abundant information. However, these methods only improve the feature selection algorithm in the classification process, and do not combine with the classification algorithm to comprehensively improve the classification.

On the basis of previous studies, this paper proposes a weighted naive Bayes algorithm based on improved mutual information - IMI-wnb algorithm. In the feature selection stage, the word frequency factor and the difference factor between classes are introduced to improve the traditional mutual information algorithm to achieve feature dimension reduction. In the process of classification, the value of IMI is introduced to weight the attributes of naive Bayes algorithm, which partly eliminates the influence of conditional independence assumption of naive Bayes algorithm on classification, and improves the efficiency and stability of spam classification.

2. Improved Mutual Information — IMI Algorithm

2.1. Mutual Information Algorithm

After text preprocessing, spam will introduce a large number of features, among which a large number of features are meaningless for classification and belong to noise features. If the dimension reduction is not carried out, the classification effect of spam filtering will be affected. Mutual information algorithm is a kind of feature selection algorithm. Mutual information value represents the correlation degree between feature and category. The larger the mutual information value is, the closer the relationship between feature and category is. The formula of mutual information is:

\[
MI(w,C) = \log\frac{P(w,C)}{P(w)P(C)} = \log\frac{P(w|C)}{P(w)}
\]

(1)

\(w\) is the feature item, \(C\) is the category, \(P(w,C)\) is the probability that the feature item \(w\) and category \(C\) appear together, \(P(w)\) is the probability of the feature item appearing in the whole training text, \(P(C)\) is the probability of the category appearing in the training text, and \(P(w|C)\) is the probability of the feature item \(w\) appearing in the class \(C\).

For training texts with \(m\) categories, the formula of MI is as follows:

\[
I_{M}(w,C) = \sum_{m=1}^{m} \log\frac{P(w,C)}{P(w)P(C)} = \sum_{m=1}^{m} \log\frac{P(w|C)}{P(w)}
\]
\[ MI(w_i) = \sum_{j=1}^{m} P(C_j) * MI(w_i, C_j) \]
\[ = \sum_{j=1}^{m} P(C_j) * \log \frac{P(w_i | C_j)}{P(w_i)} \quad (2) \]

When the MI value is larger, it is generally considered that the feature item contributes more to the text classification, while the smaller or even negative value of the calculation value is generally considered to be unimportant to the classification. Through the MI value calculated by formula (2) and selecting the appropriate threshold, some features which are not important for classification can be filtered to realize feature selection.

2.2. Word Frequency Factor
The calculation method of mutual information algorithm only considers the text frequency of feature words, but not the word frequency, which will affect the classification accuracy to a certain extent. For example, two feature items \( w_j \) and \( w_q \) have the same text frequency, in which the word frequency of feature J is several times of that of feature term \( w_q \), i.e. \( tf(w_j) >> tf(w_q) \). Generally speaking, the feature \( w_j \) with higher word frequency is more relevant to this category. However, according to the traditional calculation method of mutual information, the mutual information values of the two features are the same, which is obviously inconsistent with the actual situation. The word frequency factor \( \alpha \) is introduced to describe the difference of word frequency among different features:

\[ \alpha_j = \frac{tf_{C_j}(w_i)}{df_{C_j}(w_i)} \quad (3) \]
\[ \alpha_i = \frac{tf_{C_{spam}}(w_i)}{df_{C_{spam}}(w_i)} + \frac{tf_{C_{ham}}(w_i)}{df_{C_{ham}}(w_i)} \quad (4) \]

Among them, \( tf_{C_{spam}}(w_i) \) and \( tf_{C_{ham}}(w_i) \) are spam and non spam word frequency of feature \( w_i \), \( df_{C_{spam}}(w_i) \) is spam category text frequency of feature \( w_i \), and \( df_{C_{ham}}(w_i) \) is non spam text frequency of feature \( w_i \).

By introducing word frequency factor, the traditional mutual information is improved to IMI (improved mutual information) algorithm:

\[ IMI(w_i) = \alpha_i * \sum_{j=1}^{m} P(C_j) * \log \frac{P(w_i | C_j)}{P(w_i)} \quad (5) \]

When the word frequency of a feature item is higher than the text frequency, the greater the weight of the word frequency factor is, the stronger the ability of the feature item for mail classification.

2.3. Inter Class Difference Factor
If the feature items are evenly distributed in both categories, it is not conducive to the judgment of the category, and it appears more in one category and rarely in the other category, it is generally considered that the feature item plays a greater role in distinguishing mail categories. For example, the words "company", "invoice" and "manage" appear more in spam than in non spam. In probability statistics, the standard deviation reflects the dispersion degree of the data set, and the characteristic items with larger standard deviation are more conducive to the judgment of mail category. We improve the mutual information model by calculating the standard deviation of \( w_i \) frequency between spam class \( C_{spam} \) and non spam class \( C_{ham} \). If the frequency of feature \( w_i \) in spam class \( C_{spam} \) is \( tf_{C_{spam}}(w_i) \), in non spam class \( C_{ham} \), the frequency is \( tf_{C_{ham}}(w_i) \), and the average frequency is \( tf_{avg}(w_i) \), then there is:
\[ t_{\text{avg}}(w_i) = \frac{1}{2} (t_{C_{\text{spam}}}(w_i) + t_{C_{\text{ham}}}(w_i)) \] (6)

The inter class difference factor \( \sigma \) is introduced to describe the word frequency difference between classes:

\[ \sigma_i = \sqrt{(t_{C_{i}}(w_i) - t_{\text{avg}}(w_i))^2} \] (7)

\[ \sigma_i = \frac{1}{2} \sqrt{(t_{C_{\text{spam}}}(w_i) - t_{\text{avg}}(w_i))^2 + (t_{C_{\text{ham}}}(w_i) - t_{\text{avg}}(w_i))^2} \] (8)

The IMI algorithm is further improved to:

\[ \text{IMI}(w_i) = \alpha_i * \sigma_i * \sum_{1}^{m} P(C_j) * \log \frac{P(w_i | C_j)}{P(w_i)} \] (9)

Formula (9) adds the weight factor of frequency difference between classes on the basis of formula (5), which reflects the influence of frequency difference between classes on mail classification, and improves the efficiency of mutual information algorithm for effective feature selection.

2.4. Improved Imi Algorithm Based on Word Frequency Factor and Inter Class Difference Factor

Algorithm 1: IMI algorithm

- Input: message feature vector set \( T = \{w_1, w_2, \ldots, w_n\} \)
- Feature subset dimension \( k \)
- Output: feature subsets \( F = \{w_1, w_2, \ldots, w_k\} \)

a) Calculate \( P(C_{\text{ham}}) \) and \( P(C_{\text{spam}}) \)
b) for \( i = 1 \) to \( n \)
c) Statistics of word frequency \( t_{C_{\text{spam}}}(w_i) \) and \( t_{C_{\text{ham}}}(w_i) \)
d) Statistics document frequency \( d_{C_{\text{spam}}}(w_i) \) and \( d_{C_{\text{ham}}}(w_i) \)
e) Calculate \( P(w_i | C_{\text{spam}}) \) and \( P(w_i | C_{\text{ham}}) \)
f) Calculate \( P(w_i) \)
g) Formula (2) calculates the mutual information value \( \text{MI}(w_i) \)
h) Formula (4) calculates the word frequency factor \( \alpha_i \)
i) Equation (8) calculates the difference factor \( \sigma_i \) between classes
j) The results of equation (2) (4) (8) are substituted into equation (9) to calculate the value of IMI
k) end
l) Sort (T) // arrange the eigenvectors in descending order of IMI values
m) for \( i = 1 \) to \( k \)
n) Add feature item to feature subset \( F \)
o) end

Algorithm 1 is an algorithm in the feature selection stage of imi-wnb algorithm. IMI algorithm improves mutual information algorithm, which only considers text frequency but not word frequency. It defines and introduces word frequency factor and inter class difference factor, which reflects the contribution of word frequency and word frequency difference between classes to classification. At the same time, it improves the expression ability of feature items.

3. Naive Bayesian Classification Algorithm Based on IMI
3.1. Naive Bayesian Classification Model

Naive Bayes classification is a classification method based on Bayes theorem and independent assumption of characteristic conditions. The event probability is obtained by calculating the existing event training set, and the event probability is predicted. Given category $C_j$ and text object $d$, Bayes formula can be expressed as:

$$P(C_j | d) = \frac{P(d | C_j)P(C_j)}{P(d)}$$  \hspace{1cm} (10)

Where $P(C_j)$ is the prior probability of $C_j$ class occurrence. For spam classification, category $C$ can be divided into spam and non spam, namely $C = \{C_{\text{spam}}, C_{\text{ham}}\}$. Let the characteristic term of text $d$ be $\{w_1, w_2, \ldots, w_n\}$, and under the assumption of conditional independence of naive Bayes, there is formula (11):

$$P(d | C_j) = P(w_1, w_2, \ldots, w_n | C_j) = P(w_1 | C_j) * P(w_2 | C_j) * \ldots * P(w_n | C_j)$$  \hspace{1cm} (11)

By substituting equation (11) into equation (10), formula (12) is obtained:

$$P(C_j | d) = \frac{P(C_j) \prod_{i=1}^{n} P(w_i | C_j)}{P(d)}$$  \hspace{1cm} (12)

Since $P(d)$ is the probability of obtaining this data under any assumption, it is a standardized constant and a constant. Therefore, the category $C_{\text{map}}$ of the maximum posterior probability to be calculated by naive Bayes is as follows:

$$C_{\text{map}} = \arg \max_{C_j \in C} P(C_j) \prod_{i=1}^{n} P(w_i | C_j)$$  \hspace{1cm} (13)

In order to avoid the overflow problem caused by multiplication of a large number of decimals, the logarithm of the product of formula (13) is taken to obtain equation (14):

$$C_{\text{map}} = \arg \max_{C_j \in C} \left[ \log P(C_j) + \sum_{i=1}^{n} \log P(w_i | C_j) \right]$$  \hspace{1cm} (14)

3.2. Weighted Naive Bayes Classifier Based on IMI

Naive Bayes classification algorithm is a classification method based on conditional independence assumption, which is usually not tenable in practical application. In order to partially eliminate the adverse impact of conditional independence assumption on classification, attribute weight value can be added to naive Bayes formula to distinguish the contribution of different features to classification. The IMI value can be used as attribute weight in Bayesian classification. When the calculation result of IMI value is large, the correlation between feature and category is high, and when the value is low or even negative, it means that the effect of feature term on classification is small. Mutual information value can represent the correlation between feature and category to a certain extent, and partially eliminate the adverse effect of conditional independence hypothesis on classification. The posterior probability in equation (13) is given the weight of mutual information, and equation (15) is obtained:
The formula for calculating the weight attribute value of feature \( w_i \) for category \( C_j \) is:

\[
IMI(w_i, C_j) = \alpha_{ij} \times \sigma_{ij} \times \log \frac{P(w_i | C_j)}{P(w_i)}
\]

(15)

The formula for calculating the weight attribute value of feature \( w_i \) for category \( C_j \) is:

\[
IMI(w_i, C_j) = \alpha_{ij} \times \sigma_{ij} \times \log \frac{P(w_i | C_j)}{P(w_i)}
\]

(16)

Put the attribute weight into the above formula and take logarithm:

\[
C_{map} = \arg \max_{C \in C} \left[ \log P(C_j) + \sum_{i=1}^{n} (\log P(w_i | C_j) \times \alpha_{ij} \times \sigma_{ij} \times \log \frac{P(w_i | C_j)}{P(w_i)}) \right]
\]

(17)

In order to avoid the case of zero probability, we use Laplace smoothing to deal with \( P(w_i) \) and \( P(w_i | C_j) \) in mutual information formula:

\[
P(w_i) = \frac{df(w_i) + 1}{df_{total} + 2}
\]

(18)

\[
P(w_i | C_j) = \frac{dfC_j(w_i) + 1}{df_{C_j} + 2}
\]

(19)

Among them, \( df(w_i) \) is the text frequency of feature \( w_i \) in the whole training set, \( df_{total} \) is the text frequency of the whole training set, \( dfC_j(w_i) \) is the text frequency of feature \( w_i \) in class \( C_j \) training set, and \( df_{C_j} \) is the text frequency of \( C_j \) like training set.

Algorithm 2 Imi-wnb algorithm:

Input: feature subsets \( F = \{w_1, w_2, \ldots, w_k\} \)

Mail sample set \( D = \{d_1, d_2, \ldots, d_l\} \)

Output: the category of each sample in the sample set \( C \)

a) Calculate \( P(C_{ham}) \) and \( P(C_{spam}) \)

b) for i=1 to k

c) Count word frequency \( tf_{spam}(w_i) \) and \( tf_{ham}(w_i) \)

d) Statistics document frequency \( df_{spam}(w_i) \) and \( df_{ham}(w_i) \)

e) Equation (19) calculates \( P(w_i | C_{spam}) \) and \( P(w_i | C_{ham}) \)

f) Equation (18) calculates \( P(w_i) \)

g) Formula (1) calculates mutual information values \( MI(w_i, C_{spam}) \) and \( MI(w_i, C_{ham}) \)

h) Formula (3) calculates the word frequency factor \( \sigma_{ij} \)

i) Equation (7) calculates the difference factor \( \sigma_{ij} \) between classes

j) The results of formula (1) (3) (7) are substituted into equation (16) to calculate the values of \( IMI(w_i, C_{spam}) \) and \( IMI(w_i, C_{ham}) \)

k) end

l) for i=1 to l

m) for each \( w_i \) in \( d_i \)

n) Calculate \( P(w_i | C_{spam}) \) and \( P(w_i | C_{ham}) \)

o) Substitute \( IMI(w_i, C_{spam}), IMI(w_i, C_{ham}), P(C_{ham}), P(C_{spam}), P(w_i | C_{spam}) \) into equation (17) for calculation
Algorithm 2 is a spam filtering and classification stage algorithm. After the feature subset is obtained by algorithm 1 in the feature selection stage, algorithm 2 applies the IMI value as the attribute weight value to naive Bayes classification, which reflects the difference of the contribution of different features to classification decision-making, partially eliminates the adverse impact of naive Bayes conditional independence assumption on classification, and improves the classification accuracy.

3.3. Spam Filtering Process Based on Imi-wnb Algorithm
The general process based on imi-wnb is as follows:
First of all, in the e-mail preprocessing stage, the text is de stopped word processing, and then the text is segmented. The python Chinese word segmentation component Jieba is used to automatically segment the text.
Secondly, in the feature selection stage, the IMI algorithm is used to select the feature items in the text. Through the IMI algorithm, some features which are not related to classification can be filtered out.
Finally, the prior probability and conditional probability in the training phase are counted, and the classification is calculated by using the imi-wnb classifier in the application stage. When the spam probability is greater than the non spam probability, the classifier determines the email text as spam.

4. Conclusion
Aiming at the deficiency of traditional mutual information algorithm in feature selection, this paper proposes an improved mutual information algorithm, analyzes and improves the influence of word frequency of feature items in text and frequency difference between classes on classification, effectively utilizes frequency information in training set, and improves traditional mutual information algorithm only considers text frequency Rate of defects. In order to eliminate the negative influence of the independence assumption of naive Bayes algorithm on mail classification, this paper weighted the Bayesian algorithm and finally proposed an imi-wnb algorithm, which partially eliminated the adverse impact of the independence assumption of naive Bayes algorithm on the classification.

5. Reference
[1] Kaspersky 2018 Annual Report on Spam and Phishing Attacks[EB/OL]. https://securelist.com/spam-and-phishing-in-2018/89701/
[2] Delany S J, Buckley M, Greene D. SMS spam filtering: Methods and data[J]. Expert Systems with Applications, 2012, 39(10):9899-9908.
[3] Liu P , Zhao H H , Teng J Y , et al. Parallel naive Bayes algorithm for large-scale Chinese text classification based on spark[J]. Journal of Central South University, 2019, 26(1):1-12.
[4] Bennasar M, Hicks, Yulia, Setchi, Rossitza. Feature selection using Joint Mutual Information Maximisation[J]. Expert Systems with Applications, 2015, 42(22):8520-8532.
[5] Yang Y, Pedersen J O. A Comparative Study on Feature Selection in Text Categorization[C]// Proc International Conference on Machine Learning. 1997: 412-420.
[6] Wang, Yi, Zhao, et al. Maximum mutual information regularized classification[J]. Engineering Applications of Artificial Intelligence, 2015, 37(37):1-8.
[7] Wang W, Zhou D. A multi-level approach to highly efficient recognition of Chinese spam short messages[J]. Frontiers of Computer Science, 2018, 12(6):1-11.
[8] Sreekanth Madisetty, Maunendra Sankar Desarkar. A Neural Network-Based Ensemble Approach for Spam Detection in Twitter[J]. IEEE Transactions on Computational Social Systems, 2018, 5(4):973-984.
[9] LIU Hai-feng YAO Ze-qing SU ZhanP. Optimization Mutual Information Text Feature Selection Method Based on Word Frequency [J]. Computer Engineering, 2014, 40(7):179-182.
[10] Mishra R, Thakur R S. Analysis of Random Forest and Naïve Bayes for Spam Mail using Fea-
ture Selection Catagorization[J]. International Journal of Computer Applications, 2014, 80(3):42-47.

[11] Elsied N O F, Ibrahim O, Osman A H. Enhancement of spam detection mechanism based on hybrid k-mean clustering and support vector machine[J]. Soft Computing, 2015, 19(11):3237-3248.

[12] WU Jian-Jun, LI Chang-Bing, Mutual Information-Based Weighted Naive Bayes Text Classification Algorithm [J]. Computer Systems & Applications, 2017, 26(7):178–182.

[13] YANG Lei, CAO Cui-ling, SUN Jian-guo, ZHANG Li-guo, Study on an improved naïve Bayes algorithm in spam filtering [J]. Journal on Communications, 2017, 38(4):140-148.

[14] TAOYong-cai, ZHAOGuo-hua, SHILei, WEILin, Improved Map Reduce Mutual Information Text Feature Selection Mechanism [J]. Journal of Chinese Computer Systems, 2018, 39(3):433-438.

[15] Li Z, Lu W, Sun Z, et al. A parallel feature selection method study for text classification[J]. Neural Computing and Applications, 2016, 28(1):513-524.