Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2020.010201
ISSN 2790-0983 Vol. 1, Issue 2: 1-9

Filtering Spam Messages Based on Improved Naive Bayesian Algorithm

Xiaolei Zhang*
Personnel Department, Liaoning Police College, Dalian 116036, Liaoning, China
lnpcgkzp@163.com
**corresponding author*

Keywords: Machine Learning, Naive Bayes, SMS Filtering, Text Classification

Abstract: The relative lack of system and supervision has caused many negative impacts on the "black industry" around wireless communication, such as the spam messages of mobile phones, which have always troubled people's lives. This paper focuses on the research of spam short message filtering based on the improved naive Bayes algorithm. This paper introduces the background of spam short message, and summarizes the status quo of spam short message filtering technology and filtering system construction at home and abroad. The intelligent filtering technology is optimized, and a new algorithm combining Bayesian network classification algorithm with artificial intelligence algorithm is used to filter junk short messages. The final experimental results show that the improved naive Bayesian short message filtering method proposed in this paper can improve the accuracy, recall and overall efficiency of short message filtering, while the filtering efficiency also increases with the expansion of the cluster size.

1. Introduction

With the rapid development of technology, the cost of mobile phones is decreasing day by day, and smart phones have gradually become a necessity for the public. People use mobile phones in large quantities, and junk messages have become a problem for users. However, short message service has brought trouble while providing convenience for everyone. What makes users most worried is that a large number of junk short messages have caused problems for people, even brought economic losses to some users [1-2]. Junk messages affect the life of the public all the time. 30% of the junk messages are advertisements, 30% of them produce illegal invoices and the rest are fraud junk messages that win prizes. It has begun to infiltrate into the financial sector. What's more, illegal people use it for extortion, causing heavy losses for people with weak security awareness. The widespread popularity of spam short messages has greatly affected the normal life of users, damaged the vital interests of citizens, caused a loss of economic benefits to a certain extent, and

even shaken social stability [3-4]. How to make the SMS environment of mobile phone users safe? The legal measures to deal with junk messages are still relatively few, the supervision of operators is unfavorable, the source of junk messages is difficult to locate, and the links for criminals to participate in junk messages are complex, which creates a negative situation for public security personnel to fight against crime. In order to reduce the negative impact of spam messages, operators have also taken certain measures. In order to realize this scheme, we must consider the SMS receiving and sending service center, blacklist, content, and user reports [5]. From the perspective of the current field, junk message recognition belongs to text classification, and the improvement of the recognition rate of text classification technology brings hope to the prevention of junk messages [6].

The identification of spam messages mainly uses the technology of text classification. The idea of this kind of classification method is to turn the preprocessed text into the corresponding vector through the network model, then reduce the dimension through pooling, and finally use the classification function for classification [7]. The famous algorithms in the academic circle include naive Bayes, K-nearest neighbor method, logical regression, support vector machine and neural network. After extensive research, researchers found that K-nearest neighbor method, naive Bayes and neural network are suitable for junk message recognition. The K-nearest neighbor method is suitable for calculating the similarity between the short message text to be classified and the short message training set. It is used to count K samples with high similarity, take a large number of samples as the standard, and compare the short message text to be classified with this standard to distinguish junk short messages [8-9]. Naive Bayesian method calculates and counts the probability of text features and attributes to distinguish categories. With the rapid development of text classification technology, scholars at home and abroad have conducted research on spam short message recognition methods [10]. With a large number of experimental results published, the learning ability of support vector machine and naive Bayes far exceeds that of KNN model [11]. Common text vector representations include vector representation and semantic understanding representation. More importantly, high-dimensional data is reduced to low dimensional data, greatly reducing the computational complexity [12].

Establishing a stable and healthy Internet order is a relatively important task in society. Among them, the more appropriate solution is to use intelligent algorithms to identify and identify short messages, so as to reduce the number of junk messages and reduce the corresponding economic losses.

2. Application of Improved Naive Bayesian Algorithm in Short Message Filtering

2.1. Message Content Pre-Processing

With the development of science and technology and the popularity of network language, the content of short messages is also varied, mainly including words, pictures, expressions and websites, among which words include not only written words, but also living words and network words. Short messages are different from common texts in daily life. The content of short messages is more lifelike and less logical. The characteristics of these short messages will have a greater impact on subsequent filtering. Therefore, it is necessary to preprocess the content of short messages to meet the classification requirements [13].

Common non-standard contents and special contents are as follows:

Special words: The content of spam messages will contain more special words, which are diverse, including deformed words and expression packs.

Sensitive words: SMS content contains a large number of sensitive words, including QQ number, mobile phone number, WeChat, express bill number, bank card number, etc.

Anti interception behavior: many junk message manufacturers will do special treatment to some words to avoid junk message filtering.

Network, life style and traditional characters: SMS, as a kind of emotional communication carrier for people to people communication, has rich and diverse contents, including not only network and life style terms, but also traditional characters.

Regular expression matching is required for special words and sensitive words, such as replacing "^ _ ^" with "happy" and matching mobile phone numbers. For anti interception, you need to use regular matching keywords and replace them. For words such as network and lifestyle, you can build a related thesaurus and then use regular matching and replacement. For text pre-processing, it not only includes the sensitive words and special words mentioned in this section, but also mainly includes word segmentation, stop word processing and removing invalid characters [14]. Before word segmentation, in order to avoid the impact of sensitive words, special words and life oriented words on word segmentation, it is necessary to match these words and replace them with reasonable words.

Word segmentation has always been one of the research focuses of many scholars studying natural language processing. Word segmentation plays an important role in the search field, data mining and text classification. The quality of word segmentation directly affects the accuracy of the following models [15]. Word segmentation, also known as word segmentation, is a process of combining the necessary types of continuous self sequenced vouchers from scratch into word sequences, and is also a key step in transforming text from unstructured data to structured data. At present, the popular word segmentation algorithms are mainly divided into mechanical word segmentation and non mechanical word segmentation algorithms. Mechanical word segmentation mainly matches Chinese characters with dictionary information, while non mechanical word segmentation is mainly based on rules, grammar and now popular in-depth learning methods. This paper selects the Python third-party library jieba. The jieba word segmentation algorithm uses the prefix dictionary based method to complete effective word scanning. At present, jieba word segmentation mainly includes three word segmentation modes: precise mode, full mode and search engine mode. In order to improve the accuracy of word segmentation, it is necessary to customize the user dictionary CustomDictionary.txt according to the content of the short message, such as adding words such as "siege lion", "fist company" and "eye-catching" to the user-defined user dictionary [16-17].

Removal of stop words refers to the removal of meaningless words before word segmentation. These stop words do not contain specific and effective semantic feature information. Stop words are used to connect words and enhance tone. Therefore, removal of stop words will not adversely affect the classification of SMS. After removing the stop words, it is necessary to select the feature words to lay a solid foundation for the subsequent feature expansion. For feature selection, an improved TF-IDF algorithm is adopted, which can extract key feature words well and focuses on the distribution of feature words among different categories in a category [18].

2.2. Improved Naive Bayesian Classification Algorithm

The spam short message cloud filtering platform can conduct full text fusion analysis and identification based on the content of short messages, maximize the accuracy and efficiency of system filtering through machine self-learning technology, reduce the workload of manual

processing, and improve the efficiency of judgment. Fusion analysis is a multi-dimensional fusion analysis of structured and unstructured data formed by short message feature analysis and classified data formed by clustering and archiving. The fusion analysis service can build various data analysis models based on the business needs of operators. Several types of models mainly focus on the analysis of massive SMS feature data, including keyword analysis, frequency analysis, similar content analysis, and time-space collision analysis.

In order to improve the interception efficiency of the system, manual review shall be reduced for spam message filtering, and automatic interception of the system shall be realized through intelligent algorithm. The design process of intelligent algorithm is as follows:

When users send short messages, they first enter the short message gateway, and the care gateway sends the short messages to the spam short message filtering platform for identification;

The spam short message filtering platform makes the initial judgment according to the white list and blacklist strategies. If it is a white list user, it will be released directly, and if it is a blacklist user, it will be blocked directly; If none of the above is true, the algorithm will be judged automatically;

The algorithm monitors the content of short messages according to the data model accumulated by big data, calculates according to the sending content, sending frequency, sending quantity, sending range and other parameters of the same short message, and quickly classifies them;

Classified short messages are automatically filtered according to the policy set by the system. For example, illegal short messages are directly intercepted and blacklisted;

For SMS that cannot be classified automatically by the system, the system will automatically alert and promote manual approval;

After manual approval, the system will automatically incorporate the data into the calculation model. In the future, similar scenarios can be automatically filtered by the system.

Bayesian classification algorithm deals with three main problems: first, representing documents by attribute values, second, estimating the text probability required by Bayesian classification and third, selecting decision strategies.

To implement Bayesian classification algorithm, there is an important premise that the attribute features of the text to be classified must be independent of each other, but this condition cannot be met in some cases. If a class in a specific text training set does not contain an eigenvalue, the conditional probability of the class corresponding to the eigenvalue is 0, which is an underestimate. In order to better classify text messages, it is necessary to optimize the class conditional probability estimation method.

Use PH (xi|yi) to represent the probability of the occurrence of the eigenvalue xi in the yi class. PH (xi|yi) can be verified by the following formula:

$$PH(x_i \mid y_i) = \frac{N_{x_i y_j + 1}}{N_{y_j} + M + |C|} \tag{1}$$

In the above formula, Nxiyj represents the number of documents containing the characteristic value xi in the training set category yj, Nyj represents the total number of documents in the training set category yj, | C | represents the total number of categories, and M represents the adjustment item.

SMS can be divided into legal SMS and spam SMS. Legal SMS can be used to represent legal SMS and spams SMS can be used to represent spam SMS. Through the naive Bayesian formula, we can find the problem of low experimental recall. The following is the improved formula:

$$PH(Legal\ SMS \mid T) * \lambda - P(Spam\ SMS \mid T) \geq \varepsilon \quad (2)$$

In the above formula λ represents the magnification factor ε Represents a very small constant. If and only if λ= 1, ε=0, the above formula will be converted into naive Bayesian classification formula. When the above formula is valid, the SMS text is classified as Legal SMS. If the above formula is not valid, the SMS text is classified as Spam SMS.

3. Simulation Experiment Settings

3.1. Data Preparation

The experimental data of this paper is selected from the real short messages of China Mobile Communication Research Institute. In order to verify the filtering efficiency, the manually classified short message samples are selected as the training set and test set. During the experiment, the original SMS text data set will be read, and the relevant information of each SMS text will be formed into a line, which will be merged and stored in a single file.

3.2. Experimental Evaluation Criteria

In this paper, the performance evaluation indexes of spam message filter are: Precision, Recall, F-Measure, and Speed up Ratio. As shown in Table 1, SMS is divided into spam SMS and legal SMS in this paper, so the evaluation indicators are listed according to binary classification.

Table 1. Binary classification contingency table

	Judge it as spam message	Judge it as Legal messages
Spam messages	A	B
Legal messages	C	D

4. Analysis of Experimental Results

4.1. Comparison Experiment of Different Thresholds

In the spam short message filtering system, in order to ensure the accuracy of classification, it is necessary to set appropriate thresholds in the classification phase. Through the comparative experiment of setting different threshold ranges in the classification stage process, to find the most appropriate threshold range. The following two main stages are selected for the threshold comparison experiment.

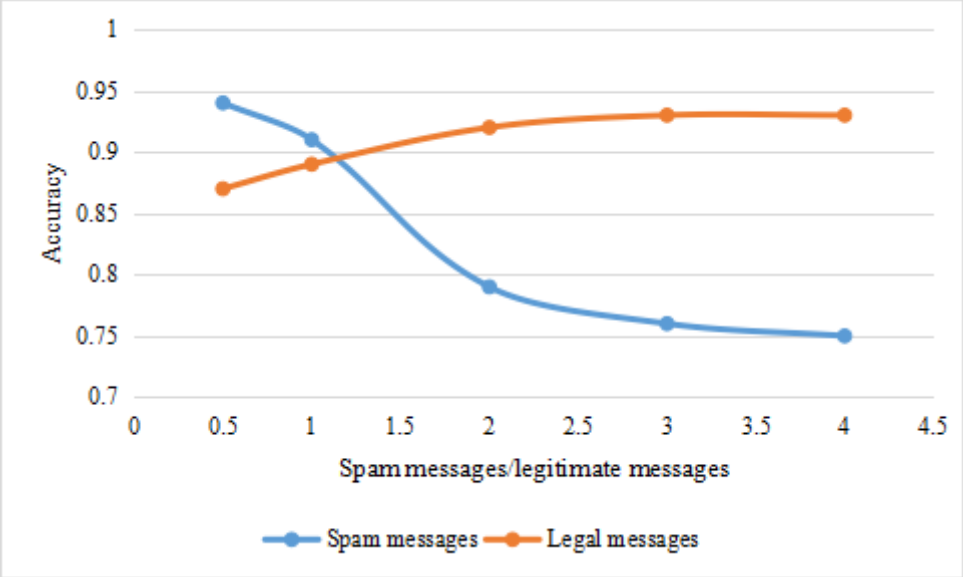

Figure 1. Experimental results of different spam messages and legitimate messages in the training set

The abscissa in Figure 1 shows the proportion of spam messages to legitimate messages. The experimental results show that the classification accuracy of the category that accounts for a large proportion in the training set sample is low. Therefore, for the distribution proportion of different types of samples in the training set, it is necessary to consider the focus of specific classification problems. In the spam message filtering problem studied in this topic, the accuracy rate of spam messages is the focus of consideration, so the number of spam messages in the training set should be appropriately less than the number of legitimate message samples.

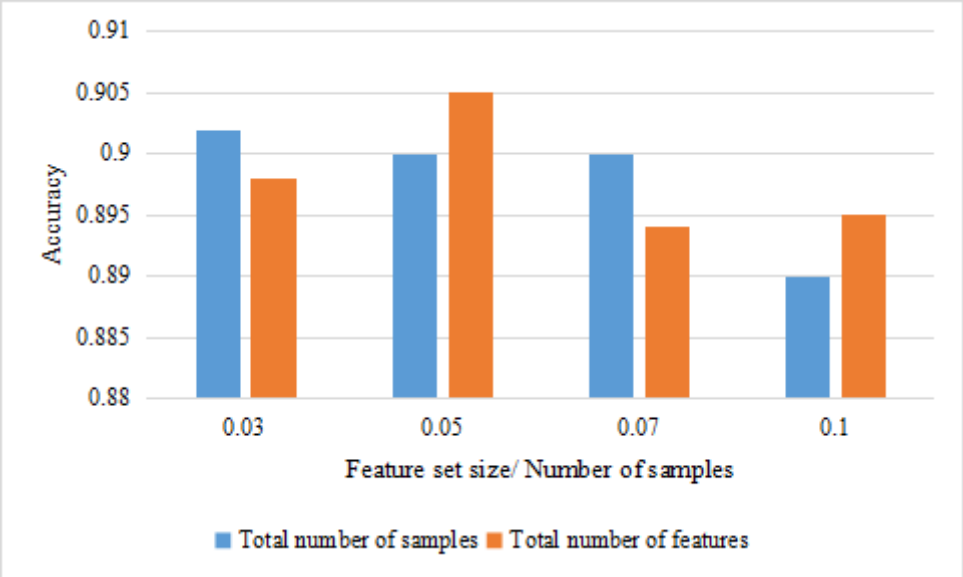

Figure 2. Experimental results based on feature sets of different sizes

In Figure 2, the abscissa represents the proportion of the feature set size to the total number of samples or features, the blue polyline represents the accuracy change chart of the different

proportions of the feature set size to the total number of features, and the black polyline represents the accuracy change chart of the different proportions of the feature set size to the total number of samples.

4.2. Comparison of Naive Bayes before and after Improvement

Set up two groups of experiments. Experiment A is based on the traditional naive Bayesian short message filtering algorithm for short message filtering; in experiment B, the improved naive Bayesian short message filtering algorithm is used for short message filtering.

Table 2. Comparison of experimental results

Group	Precision	Recall	F-Measure
A	84.91	88.56	84.90
B	92.17	92.75	92.54

By comparing the experimental results in Table 2, it can be found that the improved Bayesian short message filtering algorithm and the traditional naive Bayesian short message filtering algorithm have improved in accuracy, recall and F-Measure.

4.3. Parallelization Performance Test

The experimental evaluation index is used to accelerate the comparison of Naive Bayesian classification algorithm. In order to better demonstrate the parallel performance of the naive Bayesian algorithm, the text message data used in this experiment is expanded to 1G, 2G, 4G and other data sets of different sizes, which are run on the cloud clusters with 1, 3, and 5 cluster nodes respectively, and the algorithm execution time is recorded respectively to calculate the speedup ratio.

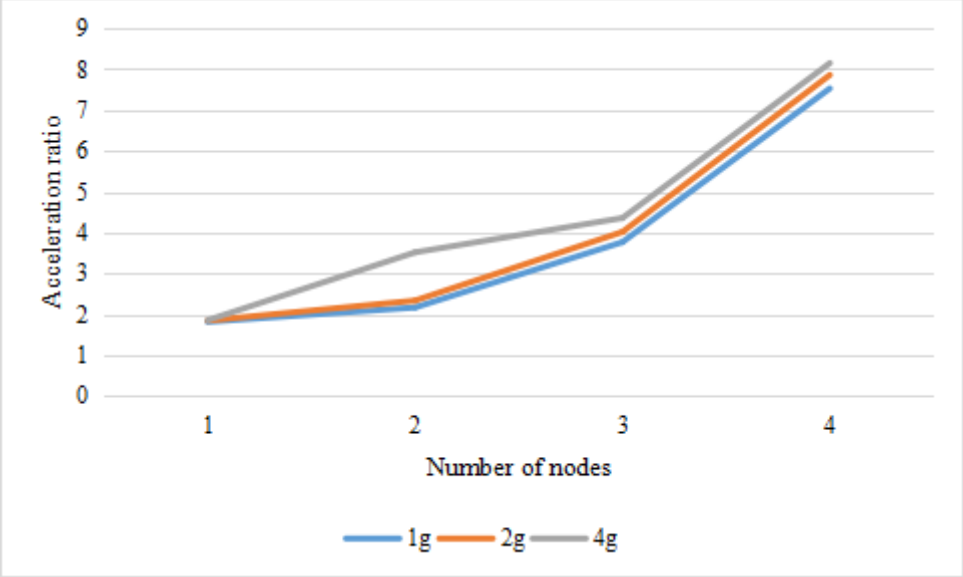

Figure 3. Ratio of speedup ratio to the number of cluster nodes

The abscissa in Figure 3 represents the number of clusters. It can be seen from the figure above

that, with the increase of data volume, the acceleration ratio of the naive Bayesian algorithm increases almost linearly, and there is no significant difference from the broken line of the acceleration ratio of smaller data sets, indicating that the processes in each stage have been distributed and run in parallel, with high concurrency and running efficiency.

5. Conclusion

In the new era of rapid information development, new technologies bring explosive growth of information. SMS, as the main means of communication in today's society, also faces enormous challenges. All kinds of advertising fraud SMS have poured into people's daily life, which has a great negative impact on citizens' privacy, social stability, etc. Therefore, the filtering of spam SMS has played an increasingly important role. Aiming at the problems of low efficiency and single strategy faced by traditional content-based SMS malicious identification methods, this paper proposes to use improved naive Bayesian machine learning algorithm to identify malicious SMS, which can improve the recognition rate by expanding the training set. The test results show that the system has the ability to identify fraudulent short messages accurately.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Mussa D J, Jameel N G M. Relevant SMS spam feature selection using wrapper approach and XGBoost algorithm. Kurdistan Journal of Applied Research, 2019, 4(2): 110-120. https://doi.org/10.24017/science.2019.2.11

[2] Adel H, Bayati M A. Building bi-lingual anti-spam SMS filter. International Journal of New Technology and Research, 2018, 4(1): 263147.

[3] Novo-Lourés M, Ruano-Ordás D, Pavón R, et al. Enhancing representation in the context of multiple-channel spam filtering. Information Processing & Management, 2020, 59(2): 102812.

[4] Vidhya K. A Machine Learning Approach to Prevent Malicious Calls over Telephony Networks. Turkish Journal of Computer and Mathematics Education (Turcomat), 2020, 12(9): 1767-1771.

[5] Barushka A, Hajek P. Spam filtering using integrated distribution-based balancing approach and regularized deep neural networks. Applied Intelligence, 2018, 48(10): 3538-3556. https://doi.org/10.1007/s10489-018-1161-y

[6] Anitha P U, Rao C V G, Babu D S. Email Spam Filtering Using Machine Learning Based Xgboost Classifier Method. Turkish Journal of Computer and Mathematics Education, 2020, 12(11): 2182-2190.

[7] Irawan D, Perkasa E B, Yurindra Y, et al. Perbandingan Klassifikasi SMS Berbasis Support Vector Machine, Naive Bayes Classifier, Random Forest dan Bagging Classifier. Jurnal Sisfokom (Sistem Informasi dan Komputer), 2020, 10(3): 432-437. https://doi.org/10.32736/sisfokom.v10i3.1302

[8] Ouni S, Fkih F, Omri M N. BERT-and CNN-based Tobeat approach for unwelcome tweets detection. Social Network Analysis and Mining, 2020, 12(1): 1-19. https://doi.org/10.1007/s13278-022-00970-0

[9] Mohammed M A, Ibrahim D A, Salman A O. Adaptive intelligent learning approach based on visual anti-spam email model for multi-natural language. Journal of Intelligent Systems, 2020, 30(1): 774-792.

[10] Okunade O A. Improved Electronic Mail Classification Using Hybridized Root Word Extractions. Fudma Journal of Sciences-ISSN: 2616-1370, 2019, 3(1): 56-71.

[11] Chirra V R R, Maddiboyina H D, Dasari Y, et al. Performance Evaluation of Email Spam Text Classification Using Deep Neural Networks. Journal homepage: http://iieta. org/journals/rces, 2020, 7(4): 91-95. https://doi.org/10.18280/rces.070403

[12] Rajendran P, Tamilarasi A, Mynavathi R. A Collaborative Abstraction Based Email Spam Filtering with Fingerprints. Wireless Personal Communications, 2020, 123(2): 1913-1923. https://doi.org/10.1007/s11277-021-09221-5

[13] Othman N F, Din W. Youtube spam detection framework using naïve bayes and logistic regression. Indonesian Journal of Electrical Engineering and Computer Science, 2019, 14(3): 1508-1517. https://doi.org/10.11591/ijeecs.v14.i3.pp1508-1517

[14] Kumar P S, Gowri D. Spam E-Mail Detection with Proabablistic Data Structure Using Java. IJRAR-International Journal of Research and Analytical Reviews (IJRAR), 2020, 8(3): 906-911-906-911.

[15] De Mendizabal I V, Basto-Fernandes V, Ezpeleta E, et al. SDRS: A new lossless dimensionality reduction for text corpora. Information Processing & Management, 2020, 57(4): 102249. https://doi.org/10.1016/j.ipm.2020.102249

[16] Mahabub A, Mahmud M I, Hossain M F. A robust system for message filtering using an ensemble machine learning supervised approach. ICIC Express Letters, Part B: Applications, 2019, 10(9): 805-812.

[17] Tuncer I, Kara K C, Karakas A. Determining abbreviations in Kariyer. net domain. New Trends and Issues Proceedings on Advances in Pure and Applied Sciences, 2020 (12): 01-07. https://doi.org/10.18844/gjpaas.v0i12.4980

[18] Odukoya O H, Adedoyin O B, Akhigbe B I, et al. An architectural-based approach to detecting spim in electronic means of communication. Nigerian Journal of Technology, 2018, 37(3): 770-778. https://doi.org/10.4314/njt.v37i3.28