Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230786

Comparison of spam classification methods based on machine learning

Chengrong Wu[1, 3, †], Jianlin Wang[2, †]

[1]Hongshan high school, Wuhan, China
[2]Cambridge foreign language middle school, Shanghai, China

[3]Tom (LZCWCR@outlook.com)
[†]These authors contribute equally to the work.

Abstract. Wapid development of the Internet, people's use of mail continues to expand, anti-spam has become a top priority. According to the statistics of relevant departments, in 2006, the total amount of spam mails received by netizens was 50 billion, which caused an economic loss of about 10.431.5 billion yuan to the national economy. In 2007, netizens received 69.4 billion pieces of junk mail, with a loss of 18.84 billion yuan. The growth rate was 38.8 percent. Spam is the main culprit that consumes network resources. Of course, preventing spam is a long way to go. Among the types of spam received by users, the top three are online shopping spam, online money-making spam and sex toys spam, which account for 17.57%, 12.55% and 9.21% respectively. Followed by attack spam, spam containing viruses, etc. At present, anti - spam main technology and method means.

Keywords: Classification.

1. Introduction

First of all, the topic of this study is the comparison of machine learning methods for spam determination. Generally speaking, people's daily life can not do without email, and spam has always been a public annoyance. In a report on the European E-mail market [1], a group predicts that the cost to businesses of spam and viruses will increase sharply in the coming years. It estimates that 46% of all E-mail received in Europe in 2004 was spam, a proportion that will rise to 71% by 2008. They will have a considerable impact on business. Spam will cost businesses in Europe 85 billion euros over the next four years. Among them, the security of mail is the main concern, and spam as a kind of invasion of people's privacy. People often encounter junk mail in the mailbox. Although it is easy to put junk mail into the trash can, it is very inefficient for people to directly judge junk mail and normal mail with the naked eye. A large number of statistics and research reports show that spam accounts for more than 50% of the world's mail, which brings interference to people's life and work, and wastes a lot of network bandwidth [2]. Before, some scholars used machine learning technology to classify spam, such as X for example in 2021 Junjun Feng use machine learning to classify the spam [3]. What will study is to compare the application of machine learning classification methods and improve the accuracy of spam classification. This will make it easier and more efficient for people to sort these spam messages. Finally, through our research, using machine learning technology to improve the

DOI: 10.54254/2755-2721/6/20230786

accuracy of spam classification and show the important role of machine learning in it.

In order to facilitate the discussion of the prevention and control measures of spam, spam is divided into three categories, namely, relay mail Pieces, low frequency spam, high frequency spam [4]. First a relay message is a message that is forwarded to another mail server using the forward function of a mail server. This is the mail that travels beten mail servers. A relay message does not relate directly to the end user, but a large number of relay emails affect the server performance and blacklist the server from other servers. Therefore, control and prevention measures are required.

Second, low junk mail is the wholesale delivery of a relatively small number of junk mail. It does not affect the performance of the mail server, but affects the processing of the mail by users. Users need to spend extra effort and cost to remove the junk mail. Third is high frequency spam which is basically all very similar messages sent in bulk by spamming tools. It is sent at a high frequency, which can affect the performance of the server, as ll as the end user. It is a disaster. It is this type of spam that is most prevalent on the Internet at present. The focus of the discussion is also on this ground.

2. Theoretical methods

2.1. Neural network learning

The process of neural network learning can also be called training. It refers to the stimulation of the external environment (data), which leads to the continuous optimization and adjustment of the free parameters of the neural network to achieve the optimal results. The main method of neural network is to discover the rules through data processing and learning, so that our problems can be solved, which is the main feature of neural network. Hover, the algorithm of neural network learning needs to continuously consider new factors and comprehensively consider them, such as the topology structure of neural network and the link mode beten neurons.

2.2. Support vector machine model

Relative than previous multi-layer perceptron model or simple neural network model, this article USES the support vector machine (SVM) model and the difference beten the above two methods of support vector machine (SVM) model will make sample points by kernel function into a higher dimensional space, then use support vector to describe the corresponding nonlinear dependency relationship beten variables nuclear oblique.

2.3. Random forests

Random Forest algorithm is a learner that includes a large number of decision trees. It can process high-dimensional data and omit the step of feature selection. Random forest algorithm also has many advantages, such as fast learning speed, not easy to overfit, and strong generalization ability. The main steps are to collect raw data and random sampling; The decision regression tree was constructed and the regression model was established for each sample. The final model is obtained after repeated training.

2.4. KNN (K-Nearest Neighbor)

KNN algorithm was first proposed by Cover and Hart in 1968 [5]. KNN algorithm first converts the test dataset and training dataset into the same pair of feature vectors, and then uses the distance function to calculate the distance beten the test dataset and the feature vectors of each training dataset. Then find out K samples that are closest to the samples in the training set, and then determine the type of test samples. KNN algorithm has the advantage that the idea is simple in KNN algorithm is compared with other algorithms are easier to implement, but also in the process of the classification of the algorithm does not need additional data sets, and due to its associated only with the surrounding a few samples, in the case of even category is difficult to determine also have very good classification effect, At the same time, it will successfully avoid the problem of unbalanced number of algorithms.

DOI: 10.54254/2755-2721/6/20230786

2.5. Naive Bayes algorithm

2.5.1. What is Naive Bayes algorithm. An advertising platform received the needs of two clothing stores, Xiaoming and Li, and prepared two online channels, A and B Run an AD. Since both Xiao Ming and Xiao Li stores sell women's clothes, they are advertisements of the same industry and category. Therefore, only one of the two channels will be shown in front of different users of A and B. A month later, from According to the click rate, Xiaoming's clothing store accounts for 65% of the total traffic of A and B channels, while Xiaoli's clothing store accounts for the rest.35% of the traffic. Only 30% of the total traffic of Xiaoming clothing store is obtained in the B channel, which is small The traffic obtained by Li Clothing Store through channel B accounts for 75% of the total traffic. Now because of the advertising plat form Due to the expiration of the cooperation with Channel A, there is only one channel left to release. Channel B To a Which clothing store has a higher flow rate beten Xiao Ming and Xiao Li? After learning the decision tree algorithm, intelligent readers will be eager to try. Some readers think just looking for To the audience of these two clothing stores and the characteristics of the crowd in this channel, this can construct a decision tree to solve this A problem. doing it with this way, it need to collect a lot of sample data and feature dimensions before Can build a more reliable decision tree. What if our goal was just to figure out which store's traffic would be Is there an easier way to solve this problem just with the information will have?

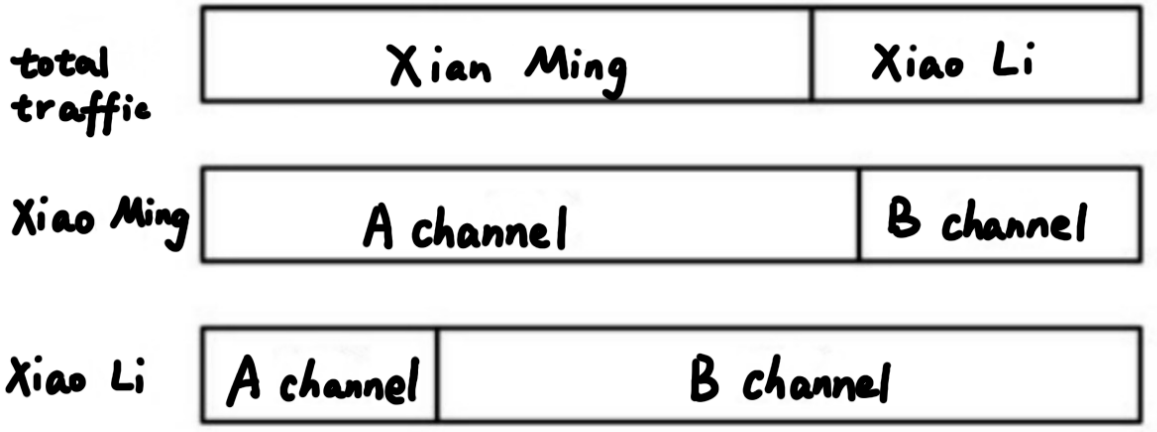

Figure 1. The example for Bayes algorithm.

Bayesian classification is a general term for a class of classification algorithms, which are based on Bayes' Theorem Based on the assumption of independent characteristics and conditions. And naive Bayes classification is the most common Bayesian classification It is also one of the most classic machine learning algorithms. In a lot of scenes? Place The problem is direct and efficient, so it has a wide range of applications in many fields, such as spam filtering, text Classification and spelling correction. Naive Bayes classification is a very simple classification algorithm, and it is very simple because of its solution The idea is very simple, that is, for a given term to be classified, solve the problem under some conditions? Various categories appear the probability, which is the largest, is considered to belong to the category. To give a visual example, if the Walking in the street to see a dark foreign friend, guess where this foreign friend from, ten times out of ten 'd guess it's from Africa, because Africans make up the most of the dark-skinned people, even though they're dark-skinned Eighty percent of foreigners are also likely to be American or Asian. But where there is no other information available? Going to pick the category with the highest probability, and that's the basic idea of Naive Bayes.

$$P(Y_k\,|\mathrm{X}) = \frac{P(XY_k)}{P(X)} = \frac{P(Y_k)P(X|Y_k)}{\sum_j P(Y_j)P(X|Y_j)}$$

The basic formula made by Naïve Bayes.

DOI: 10.54254/2755-2721/6/20230786

Naive Bayes algorithm is a typical statistical learning method, and its main theoretical basis is a Bayesian formula. The basic definition of Bayesian formula is as follows. The right-hand side of the formula is the prior probability The left-hand side of the formula is the predicted probability If X Y as a category, as a characteristic, P (Yk | X) is in the case of known characteristics X Yk category of probability, and P (Yk | X) the calculation of all the into the category of Yk feature and distribution. The representation of Naive Bayes classifier:

$$y = f(x) = \arg\max_{c_k} P(Y = c_k \mid X = x) = \arg\max_{c_k} \frac{P(Y = c_k)\prod_j P\left(X^{(j)} = x^{(j)} \middle| \mathrm{Y} = c_k\right)}{\sum_k P(Y = c_k)\prod_j P(X^{(j)} = x^{(j)} \mid \mathrm{Y} = C_K)}$$

Naive Bayes' formula work like this.

When the feature is x, the conditional probabilities of all categories are calculated, and the category with the largest conditional probability is selected as the category to be classified. Since the denominator of the above formula is the same for each class, the calculation can be done without considering the denominator,

$$y = f(x) = \arg\max_{c_k} P(Y = c_k)\prod_j P\left(X^{(j)} = x^{(j)} \middle| \mathrm{Y} = c_k\right)$$

The calculation can be done without considering the denominator.

2.5.2. Application to text classification. There are many applications of text classification, such as spam mail and spam SMS filtering is a 2-classification problem, news classification, text sentiment analysis. So it can be regarded as text classification problem, classification problem consists of two steps: training and prediction, to establish a classification model, at least one training dataset is required. The Bayesian model can be naturally applied to text classification: Now there is a Document d (Document), and to determine which category ck it belongs to. People only need to calculate the probability that document d belongs to which category is the largest:

$$c = \arg\max_{c_k} P(c_k \mid \mathrm{d})$$

One may calculate the probability.

In the classification problem, people do not use all the features. For a document d, people only use some of the feature items t1,t2... D, TND (and said the total number of entries), because a lot of words for classification is of little value, such as some stop words ", is, in the "will appear in each category, the term is also fuzzy classification decision, about the selection of key words. Then people use of key items after the said document, calculate the document d categories into:

$$c = \arg\max_{c_k} P(c_k \mid \mathrm{d}) = \arg\max_{c_k} P(c_k) \prod_{1 \le j \le n_d} P(t_j \mid c_k)$$

People use of key items after the said document, calculate the document d categories.

Researchers only need to be calculated from the training data set, each of the categories appear probability P (ck) and the probability of each key items in each category P (tj | ck), and the probability value calculated using maximum likelihood estimation. Thus it is essentially statistics the number of occurrences of each word in each categories and the number of each category of documents.

$$P(c_k) = \frac{N_{c_k}}{N}$$

$$P\left(t_j \middle| c_k\right) = \frac{T_{jk}}{\sum_{1 \le i \le |V|} T_{ik}}$$

The essentially statistics the number of occurrences.

DOI: 10.54254/2755-2721/6/20230786

2.5.3. The pros and cons of Naive Bayes.

Advantages:

- Naive Bayes model originated from classical mathematical theory, which has a solid mathematical foundation and stable classification efficiency;
- People perform ll on small-scale data;
- People can handle multiple classification tasks, suitable for incremental training;
- It is not very sensitive to missing data and the algorithm is relatively simple, which is often used for text classification

Disadvantages:

- People can only be used for classification problems;
- Prior probability should be calculated;
- There is error rate in classification decision;
- The system is sensitive to the expression form of input data.

2.5.4. Naive Bayes.

- Gaussian distribution Naive Bayes

The Gaussian distribution is the normal distribution and is used in general classification problems

- Polynomial distribution Naive Bayes

The distribution applies to textual data (features represent times, such as the number of occurrences of a word)

$$P(X_1 = n_1, \cdots, X_k = n_k) = \begin{cases} \dfrac{n!}{n_1! \cdots n_k!} P_1^{n_1} \cdots P_k^{n_k}, \sum_{i=1}^{k} n_i = n \\ 0, otherwise \end{cases}$$

Polynomial distribution Naive Bayes.

- Bernoulli distribution Naive Bayes

It applies to Bernoulli distribution, and also applies to text data (in this case, the feature represents whether the occurrence of a word is 1, for example, the occurrence of a word is 0). In most cases, the performance of the Bernoulli distribution is not as good as the polynomial distribution, but sometimes the Bernoulli distribution is better than the polynomial distribution, especially for small orders of magnitude text data.

$$\frac{n!}{r!\,(n-r)!} P^k (1-P)^{n-k}$$

The small orders of magnitude text data.

3. Data preprocessing

The data set used in this article is from the Apache Spam Assassin public data set [6]. There are 2500 normal emails and 500 spam emails in the data set, and a total of 3000 emails are normal. In the process of data processing, Python language is used to convert the data file type into readable CSV file. It mainly calls NTLK to process emails in natural language. This module can be downloaded directly from the Python official module library. It can slice and dice messages into natural sentences, then slice and dice them into single words and translate them into computer language. At the same time, mark normal mail as 1 and spam mail as 0 so that the computer can recognize it.

4. Intra-group study and analysis

Firstly, this paper uses the data downloaded from Kaggle for data preprocessing. The original data downloaded from Kaggle cannot be imported directly because the original data is a data file. open the file in a Python readable way and read Lines () reads each line (string) and saves it in the list, removes the last line break, and appends it to the new list. Finally, the list is written to a CSV file. Then the D Rap NA function is used to process the missing values of the data. Then, the NLTK function is used to

DOI: 10.54254/2755-2721/6/20230786

divide the paragraph sentence by natural language processing technology. Finally, it is divided into words and converted into computer language.

In this paper, the data set is first divided into training set and test set, and the machine learning SVC algorithm is used to learn and classify the data set. For the statistical mail filtering method [7], the KNN algorithm is more widely used when the number of categories is not clear enough. As a simple and easy to implement chance rule classification algorithm, it is often used in machine learning. Then, the same training set and test set are used to learn and classify the dataset using KNN algorithm. Finally, the random forest algorithm of neural network is used to learn and classify data sets. Finally, the accuracy of SVC algorithm is 94%, KNN algorithm is 92%, and random forest algorithm is 96%.[8]

The simple Bayes algorithm is different from the regression algorithm and decision tree algorithm will learned earlier. Regression and decision tree algorithms are practical algorithms that can be directly applied. Although the naive Bayes algorithm is simple to implement, it has an important premise that attributes are independent from each other, which is often not true in practical applications. Because of the limitation of this premise, Bayesian algorithm can only be used in scenes with few features and small correlation beten features for a long time. Once the number of attributes increases or the correlation beten attributes becomes large, the classification effect will be sharp. This situation is very much like the current situation of some cutting-edge technologies, the same technology has the solution, but has not found the right application scenario, so it cannot be utilized. Thanks to the rapid development of modern natural language processing, people gradually find that naive Bayes algorithm is very suitable for processing text information, such as spam detection, community violation information detection and document classification. The main reason is that the correlation beten text words is very small, which can be assumed to be independent of each other, so the application of Bayesian algorithm in text has a significant effect. About 10 years ago, opening our mailboxes every day to find a flood of ads, swamping important emails, and making many users miserable. [9] Smart product managers spotted this pain point and immediately came up with rules to filter out emails with certain words in the subject line, a feature that still exists in many mailboxes today. Hover, cunning merchants always try to evade the detection of keywords, so the filtering effect of this method is not good. If the filtering rules are set for each new type of spam, the false detection rate of this filter will also increase, and it is possible to misjudge normal mail as spam. For most users, the consequences of missing a normal email are much more serious than receiving spam, so a good filter can't misjudge an email. It is a "better let go than kill" scenario. In this case? ", an engineer named Paul Graham proposed using "Naive Bayes" to filter spam, and his tests shod that it worked very ll, filtering 995 out of 1,000 spam messages without a single miscalculation. What's even more porful is that the filter is self-learning, constantly taking the model based on incoming emails, and the more spam the get, the better it gets. How does such a fantastic classifier work? In fact, Graham just built a classifier based on Naive Bayes. In the data preparation phase, he found 4,000 normal emails and 4,000 spam emails. First, all the emails re parsed and each word in the 8000 emails was extracted to build a vocabulary database. This database consisted of two tables, one of which recorded all the words that appeared in the emails, and the other of which counted the frequency of each of these words. Pick up? To calculate how often each word appears in normal mail versus spam. For example, if detect that 200 out of 4,000 spam messages contain the word "sex," the probability that the word will appear in the spam is 5%. With only two out of 4,000 normal emails containing the word, the probability of the word appearing in a normal email was 0.05%. [10] With this preliminary statistical result, the classifier is ready for use.

$$P(S|\mathrm{W}) = \frac{P(S,W)}{P(W)} = \frac{P(W|\mathrm{S})\mathrm{P}(\mathrm{S})}{P(W|\mathrm{S})\mathrm{P}(\mathrm{S}) + P(W|\mathrm{H})\mathrm{P}(\mathrm{H})}$$

How it works like in the computer.

DOI: 10.54254/2755-2721/6/20230786

$$P(S|\mathrm{W}) = \frac{5\% \times 50\%}{5\% \times 50\% + 0.05\% \times 50\%} = \frac{0.025}{0.02525} = 0.9901$$

How it is converted into the number.

5. Conclusion

According to the previous test results of different algorithms using the same training set and test set, the random forest algorithm has the highest accuracy of 96%, while KNN algorithm has the lost accuracy of 92% among the three algorithms. The main reason is that according to empirical studies, random forest algorithm is the most reliable method for classifying spam mails in our daily use.

References

[1] Androutsopoulos I.J. Koutsias, K.V. Chandrinos, G. Paliouras, and C.D. Spyropoulos. 2000a. An Evaluation of Naive Bayesian Anti-Spam Filtering. Proceedings of the Workshop on Machine Learning in the New Information Age, Barcelona, Spain, pages 9-17.

[2] ATENIESE G, BURNS R, CURTMOLA R, et al. Provable data possession at untrusted stores [C]// Proceedings of the 14th ACM conference on Computer and Communication Security. New York: ACM, 2007 :598-609.

[3] Feng Junjun, LI Li. Implementation of Machine Learning in Spam Filtering [J]. Computer Knowledge and Technology,2021,17(08):154-155.DOI:10.14004/j.cnki.ckt.2021.06

[4] Shen ichao, Design and Implementation of mail filter system. Information and Electronic Engineering, June 2003, Volume 1, Number 2, P18-21.

[5] T.M.Cover and PE. Hart(1968), Rates of convergence for Nearest Neighbor Procedures , inProc .HaWaii Int. Conf . on System Science

[6] https://www.kaggle.com/datasets/ozlerhakan/spam-or-not-spam-dataset HAKAN OZLER

[7] Hao Jie. Research on P2P Traffic Detection and Control Based on Dual Features [D]. Chengdu: University of Electronic Science and Technology of China, 2010, 25-26.

[8] Xue Jinqi Research on spam identification and processing scheme. Sun Yat sen University 20040508

[9] https://blog.csdn.net/tysonchiu/article/details/125485175 Google academic

[10] https://www.jianshu.com/p/7ddcf3f996f8 jianshu