JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 39, 855-868 (2023)
DOI: 10.6688/JISE.202307_39(4).00XX

Classify Text-based Email Using Naive Bayes Method With Small Sample

YANJUN ZHU[1], TING ZHU[2,+], JIANXIN LI[3,8], WENLIANG CAO[4],
PENG YONG[5], FEI JIANG[6] AND JIE LIU[7]
[1,3,4,5,7]*School of Electronic Information*
Dongguan Polytechnic, Dongguan, 523808 China
E-mail: Zhuyanjun073@163.com[1]*; 279149042@qq.com*[3]*; caowl22@163.com*[4]*;*
pengy@dgpt.edu.cn[5]*; 1123261349@qq.com*[7]
[2]*Department of Information Engineering*
Gannan University of Science and Technology
Ganzhou, Jiangxi, 341000 China
E-mail: tingzhu915@163.com
[6]*School of Modern Circulation, Guang Xi International Business Vocational College, Nanning, China*
School of Management and Marketing, Taylors University, Malaysia
[8]*Department of Public Relations*
Guangdong Only Network Science and Technology Co., Ltd., 523000 China
E-mail: 279149042@qq.com

With the popularity of the Internet, e-mail has gradually become one of the important communication tools for people's work and life with its fast and convenient advantages. However, the problem of spam has become increasingly serious. It not only spreads harmful information, but also consumes a lot of public resources and infringes the legitimate rights and interests of e-mail users and enterprises. Although there are many spam filtering methods at present, the situation that spam does not fall but rises shows that the existing spam filtering methods have not achieved ideal filtering effect. This paper uses naive Bayesian method and small sample to classify e-mail, and combines Chinese information processing technology to propose an efficient filtering system BETSY. The experimental results show that the method proposed in this paper has achieved good results and has direct application value.

Keywords: Naïve Bayes, classifier, E-mail classification, filtering system, small sample

1. INTRODUCTION

Text classification is useful in document classifications and many techniques have been developed already. In this paper, Naïve Bayes method is used for e-mail classification and proved to be effective based on the experiments.

1.1 Text Classification

Text classification [1, 2] is an important research content of text mining. It refers to determining a category for each document in the document collection according to the predefined topic category. Through text classification, people can store, retrieve and further process text by category, which can help people better find the information and knowledge

Received May 29, 2022; revised December 12, 2022; accepted January 24, 2023.
Communicated by Mu-Yen Chen.
+ Corresponding author: Ting Zhu (tingzhu915@163.com)

they need. Text classification has gone through several different stages of development. The first text classification is mainly manually identified. By 1964, Mosteller and Wallace had created a new stage of text classification in the work of identifying the author of the article. They considered the characteristics of words, sentence length, frequency of functional words and lexical differences in classification. At present, although the information carriers on the Internet are diversified, text is still the main source of information on the Internet, which makes the recent text classification have a wide range of applications, such as extracting symbolic knowledge, news distribution, e-mail classification, e-mail filtering, learning user interests, search engines, information retrieval, *etc.*

Classification is one of the most basic cognitive forms of information. The traditional literature classification [3-5] research has rich research results and considerable practical level. However, with the rapid growth of text information, especially the surge of online text information on the Internet, the traditional manual classification methods have been powerless. Using advanced computer technology to carry out automatic classification is not only convenient, fast and simple, but also can further carry out deeper information processing to improve the efficiency of information utilization. At the same time, with the increasing abundance of information, people's requirements for the accuracy and recall of content search will become higher and higher. Automatic classification using advanced computer technology is an effective means and inevitable trend to replace the traditional manual classification methods, which is of great significance to improve the efficiency and quality of information search.

1.2 E-mail Client

With the development of Internet technology and network office, e-mail has become one of the main means of communication. E-mail system includes client [6-8] and server. Among the current two popular operating systems, Windows operating system is known for its friendly interface and convenient operation. However, e-mail system based on Windows platform generally does not disclose the source code, and people can not complete or reduce its functions according to specific needs; Although the network service function of Unix operating system is powerful, it requires administrators to understand the mail system structure and have rich experience in Unix platform development, which is more suitable for installing large mail servers.

With the development of computer software and hardware and the improvement of communication service quality, the mail system has realized the transmission of various types of information such as electronic letters, document numbers, images and digital voice. At the same time, users can easily obtain a lot of free news information and various special mail information to achieve fast and efficient information retrieval. Its extensive use has changed the traditional way of communication. The number of e-mails sent every day continues to increase. Experts estimate that by 2021, more than 306 billion e-mails will be sent every day. By 2024, this figure is expected to increase by 20%. These figures highlight the convenience of e-mail communication, and also indicate that e-mail is beginning to consume our lives. Many people complain that they can hardly keep up with all the emails they receive every day. Spam is undoubtedly a problem, but even legitimate e-mail can cause its inbox to overflow. The e-mail client is helping us solve the problem of excessive e-mail by classifying e-mail content.

As people increasingly rely on e-mail technology to communicate with friends and colleagues, the number of contacts in the e-mail list will continue to grow. This will bring various complications to people. It may be difficult for them to track different people in the contact list, which may make them forget how the relationship was initially formed. Even if their e-mail addresses and names are still stored in the database, they may completely lose track of some contacts. The e-mail client manages the contact list through big data technology, and can also rely on new AI technology to help recommend new sorting options to help us solve the e-mail contact management problem. Big data is very useful for e-mail clients. They use this technology to solve many problems, including contact management and network threat prevention. Everyone who uses e-mail will benefit from the new development of big data technology.

In practical applications, it is often necessary to carry out secondary development of the mail system, and most of the existing mail client software does not disclose the source code, so people cannot enhance its performance. Therefore, this paper describes and implements a mail filter BETSY.

2. RELATED WORK

Email filtering [9, 10] is a tool for users to filter, classify and manage email. Email filtering can be defined for the address and subject of the letter, to control whether the mail meeting a certain condition is stored in the specified location, and to classify the user's related mail. For example, all messages from a user's uncle Joe may be placed in a folder named "Uncle Joe". Filters can also be used to block or receive e-mail from a specified source. With the rapid development of Internet-related applications, the progress of advertising technology and the popularity of e-mail, more and more spam is flooding our lives. Spam generally has the characteristics of mass sending. Its contents include moneymaking information, adult advertising, commercial or personal website advertising, electronic magazines, serial letters, *etc.* Spam can be divided into benign and malignant. Benign spam is a kind of information mail that has little impact on the recipient, such as various advertisements. Malicious spam refers to destructive email. In this section, we will focus on personal email filtering and discuss the relevant technologies of email filtering.

2.1 Text Representation Model

The biggest difference between text and database records is its unstructured characteristics [11-14]. Text is a one-dimensional linear character stream. From the perspective of modern linguistics, text has a three-dimensional recursive structure. However, in order to analyze this recursive generation structure, it is necessary to carry out deep natural language processing of large-scale real text. At present, this technology has not reached the practical level. When we say that the content of text information is unstructured, it mainly means that its surface layer has no vector structure like the records in the database. Therefore, if you want to use the mature classification and clustering technology of structured data to classify and cluster text, the first problem to be solved is the structure of unstructured data. In order to enable the computer to truly analyze the text features, it is necessary to effectively represent the text features, and express the text as a mathematical vector that

the computer can process. From the first day of the emergence of text classification technology, there have been many text representation models, such as Boolean model, vector space model, probability model, potential semantic index model, *etc.* These models use different methods to deal with feature weighting, category learning and similarity judgment from different perspectives. This paper adopts probability model.

The probability model considers the correlation between words and classifies the text in the text set into relevant text and irrelevant text. Based on the probability theory in mathematical theory, the probability of the occurrence of these words between the relevant text and the irrelevant text is expressed by giving some probability value to the characteristic words, and then the probability of the correlation between the texts is calculated, and the system makes decisions based on this probability.

There are many forms of probability models, one of which is Bayesian probability model. Bayesian probability model uses probability architecture to represent feature items, decomposes training examples into feature vectors and decision category variables. This model assumes that the components of feature vectors are relatively independent from the decision variables, that is, each component acts independently on the decision variables. Although this assumption limits the scope of application of Bayesian model to a certain extent, in practical applications, the complexity of Bayesian network construction is reduced exponentially. In many fields, even if this assumption is violated, Bayesian probability model also shows considerable robustness and efficiency, and has been successfully applied to classification.

2.2 Supervisions

When you define a classification category, the text will be supervised and classified [15-17]. Its working principle is training and testing. We provide tag data for machine learning algorithms. The algorithm trains on the marked data set and gives the required output (predefined category). In the test phase, the algorithm uses the unobserved data and classifies them according to the training phase.

Email filtering is an example of supervised classification. Received e-mails are automatically classified according to their contents. Language detection, intention, emotion and emotion analysis are all based on the monitoring system. It can operate on special use cases, such as identifying emergencies by analyzing millions of online information.

In the BETSY system we have described and implemented, a large number of supervision training will greatly improve the accuracy of BETSY's filtering email. For example, BETSY filtered 40 messages in 10 different folders by using training data sets, and 8 messages were filtered incorrectly, with an accuracy rate of 80%.

2.3 Naive Bayes

Bayesian classification [18-22] is a statistical classification method, which is a kind of classification algorithm using probability and statistical knowledge. In many cases, naive Bayesian classification algorithm can be compared with decision tree and neural network classification algorithm. This algorithm can be applied to large databases, and the method is simple, the classification accuracy is high, and the speed is fast. A Bayesian classifier assumes that a document is generated by a mixture model with parameters θ,

consisting of components $C = \{c_1, \ldots, c_n\}$ that correspond to the classes. A document is generated by first selecting a component $c_j \in C$ according to the prior distribution $P(c_j|\theta)$ and then choosing a document d_i according to the parameters of c_j with distribution $P(d_i|c_j; \theta)$. The likelihood of a document is given by the total probability.

$$P(d_i \mid \theta) = \sum_{j=1}^{n} P(c_j \mid \theta) P(d_i \mid c_j; \theta) \tag{1}$$

Bayes method requires that the subject words of the text are independent of each other, the parameter θ and parameter c shown in Eqs. (2)-(4):

$$\hat{\theta}_{c_j} = P(c_j \mid \hat{\theta}) = \frac{\sum_{i=1}^{|D|} P(c_j \mid d_i)}{|D|}, \tag{2}$$

$$P(c_j \mid d; \hat{\theta}) = \frac{P(c_j \mid \hat{\theta}) P(d \mid c_j; \hat{\theta})}{P(d \mid \hat{\theta})}, \tag{3}$$

$$c_d = \arg\max_{c_j \in c} P(c_j \mid \hat{\theta}) P(d \mid c_j; \hat{\theta}) \tag{4}$$

2.4 Small Sample

Data-driven evaluation methods mainly include fuzzy clustering, support vector machine, neural network, *etc.* These methods require large sample size, but in the actual evaluation process, due to the impact of cost, period and other factors, it is impossible to conduct a large number of tests to obtain large samples. In this case, the traditional statistical method under the condition of large sample is still used, which is difficult to ensure the reliability of the final evaluation results. Therefore, it is of practical significance to study the state detection and evaluation method under the condition of small sample. Small sample analysis methods refer to statistical analysis methods applicable to processing small sample data, such as gray model, Bayes method, *etc.* Yang J W *et al.* calculated the cumulative failure probability of the anti-skid valve according to Bayes reliability theory, solved the posterior distribution using MCMC algorithm, and then obtained the reliability information of the entire anti-skid system [23]. Fan X *et al.* reasonably combined the experimental data and historical information, and estimated the life of LED with higher confidence using Bayes method [24]. Ali S *et al.* modified the generalized exponential distribution model and used Bayes method to estimate the parameters of the modified model to study the life of electronic equipment at different voltage levels [25]. Zhou K *et al.* analyzed the reliability of transmission line interruption based on layered Bayes model with limited interruption data [26].

3. BETSY

BETSY is a windows-based program that classifies text based on trained material. It was designed for automated essay scoring and can be applied to any text classification task. Its features include following: Multinomial & Bernoulli Naive Bayes models, Optional

Porter stemming, popular database format, output on screen and into CSV files, re-entrant training, training file trimming, infrequent term purge, phrase search, web interface, diagnostic information and free. BETSY was originally proposed and developed by the University of Maryland. In recent years, more and more users have used BETSY for email filtering, and it has become a mature software package. This paper also carries out function test on this system.

BETSY uses Naive Bayes algorithm, which can quickly classify and filter junk e-mail. The premise of Naive Bayesian algorithm is that each attribute is independent of each other. When the data set meets this assumption of independence, the accuracy of classification is higher, otherwise it may be lower.

Table 1. E-mail corpus information.

User	Number of Email Messages	Number of Folders
1	2715	27
2	373	16
3	655	13
4	4447	33

The weakness of Bayes method is that the probability distribution of the class population and the probability distribution function (or density function) of various samples are often unknown. In order to obtain them, the sample is required to be large enough. In addition, the Bayes method requires that the subject words of the expression text are independent of each other, which is generally difficult to meet in the actual text, so the method is often difficult to achieve the theoretical maximum in effect.

4. EXPERIMENTS

From the perspective of system implementation, mail classification generally includes three modules: establishing feature library, training, and testing.

First, the BETSY system extracts the feature items from the text of the training email, and establishes a feature database about the feature information; Then use the training set email to represent the text according to the information of the feature items in the feature library for the training of the classifier; Finally, the test set email is used to evaluate the effect of the system classification based on the text representation of the feature items in the feature library.

The purpose of establishing feature library is to reduce the dimension of text feature vector in email, remove redundant feature items, and retain distinguishing feature items. At the same time, distinguishing features can improve the accuracy of the system. The process of building feature database is to select an evaluation function to evaluate all feature items according to the statistics of a large number of samples. Select the feature items with high evaluation value to form the feature library. We believe that the feature items with high evaluation score are the backbone of the article and play a key role in the understanding and expression of the article, while the feature items with low evaluation score make little contribution to the content of the article. Removing these feature items will not affect the expression and classification of the meaning of the article.

The essence of training is to obtain the classification model through the training samp-

les of known categories, which can be divided into three processes: text preprocessing, text representation and training classifier. The purpose of text preprocessing is to remove stop words from the text. The purpose of text representation is to retain the feature items with strong distinguishing ability and calculate their weights, and replace the text with the form of feature vectors. The purpose of training classifier is to establish a functional mapping relationship of text classification by using machine learning method for the training text and its category.

The purpose of the test is to measure the effect of the text classification system by comparing the classification of some articles made by experts and the text classification system. The testing process includes four steps: text preprocessing, text representation, classification and evaluation of classification results. Text preprocessing and presentation are exactly the same as the work done in the training module. Classification is to judge which category the mail belongs to through the trained classifier. Finally, the classification results are evaluated using evaluation indicators. We collected 150 e-mails as training samples, and 5 e-mails were used to test the dataset.

We start the experiment with 50 training samples for each group. After finishing training step, then we proceed to classify the testing samples. Next, we add another 50 training samples (100 samples) in the training step and classify testing samples again. Finally, we use all the training samples (150 samples) to precede training and classify the testing samples. For testing samples, we collect 5 mails for each group. The result of classification is shown in Table 2. The classification effect is very good. It shows that BETSY mail filter has good filtering effect, and the detailed results are as follows.

Table 2. The classification accuracies in different number of training samples.

Number of Training Samples	Classification Accuracy
50	90%
100	100%
150	100%

In Tables 3-5, the rows are testing samples and each cell means the probability of each testing sample in each group (business, entertainment, science and sports). Bold font means the highest probability of the testing sample.

From the results in Tables 3-5, we can see that when I use 50 training samples, the probability of the testing sample e3 is 0.6131942, and this means after calculation by BETSY, e3 is classified as entertainment with about 60%. When I use 100 training samples, the probability becomes 0.9991125. After using all training samples, the probability becomes 0.9999985, and it is close to 1. It is obviously that when we use more training samples, we will get higher precision.

Table 3. The classification result with 50 training samples.

Testing Sample	business_prob	entertainment_prob	science_prob	sports_prob
b1	1.0000000	0.0000000	0.0000000	0.0000000
b2	1.0000000	0.0000000	0.0000000	0.0000000
b3	1.0000000	0.0000000	0.0000000	0.0000000
b4	1.0000000	0.0000000	0.0000000	0.0000000
b5	1.0000000	0.0000000	0.0000000	0.0000000

e1	0.0070509	0.9926809	0.0002682	0.0000000
e2	0.0000000	0.9804164	0.0000113	0.0195723
e3	0.0067691	0.6131942	0.0022366	0.3778000
e4	0.0000125	0.9986121	0.0013539	0.0000215
e5	0.0044899	0.9949812	0.0004243	0.0001045
sc1	0.0543090	0.5653368	0.3803514	0.0000028
sc2	0.0000000	0.0000000	1.0000000	0.0000000
sc3	0.3266367	0.0004327	0.6729306	0.0000000
sc4	0.0000001	0.0105692	0.2497508	0.7396799
sc5	0.0000000	0.0000882	0.9999118	0.0000000
sp1	0.0000000	0.0000000	0.0000000	0.9999999
sp2	0.0000001	0.0000000	0.0000000	0.9999999
sp3	0.0000000	0.0000000	0.0000000	1.0000000
sp4	0.0000000	0.0000000	0.0000000	1.0000000
sp5	0.0000001	0.0000002	0.0000001	0.9999996

Table 4. The classification result with 100 training samples.

Testing Sample	business_prob	entertainment_prob	science_prob	sports_prob
b1	1.0000000	0.0000000	0.0000000	0.0000000
b2	1.0000000	0.0000000	0.0000000	0.0000000
b3	1.0000000	0.0000000	0.0000000	0.0000000
b4	1.0000000	0.0000000	0.0000000	0.0000000
b5	1.0000000	0.0000000	0.0000000	0.0000000
e1	0.0003478	0.9976545	0.0019978	0.0000000
e2	0.0000000	1.0000000	0.0000000	0.0000000
e3	0.0001408	0.9991125	0.0002817	0.0004650
e4	0.0000000	1.0000000	0.0000000	0.0000000
e5	0.0000000	0.9999997	0.0000002	0.0000000
sc1	0.0000468	0.0073511	0.9926016	0.0000006
sc2	0.0000000	0.0000000	1.0000000	0.0000000
sc3	0.0000000	0.0000000	1.0000000	0.0000000
sc4	0.0000000	0.0001454	0.9969147	0.0029399
sc5	0.0000000	0.0000000	1.0000000	0.0000000
sp1	0.0000000	0.0000000	0.0000000	1.0000000
sp2	0.0000000	0.0000000	0.0000000	1.0000000
sp3	0.0000000	0.0000000	0.0000000	1.0000000
sp4	0.0000000	0.0000000	0.0000000	1.0000000
sp5	0.0000000	0.0000000	0.0000000	1.0000000

Table 5. The classification result with 150 training samples.

Testing Sample	business_prob	entertainment_prob	science_prob	sports_prob
b1	1.0000000	0.0000000	0.0000000	0.0000000
b2	1.0000000	0.0000000	0.0000000	0.0000000
b3	1.0000000	0.0000000	0.0000000	0.0000000
b4	1.0000000	0.0000000	0.0000000	0.0000000
b5	1.0000000	0.0000000	0.0000000	0.0000000
e1	0.0000000	1.0000000	0.0000000	0.0000000
e2	0.0000000	1.0000000	0.0000000	0.0000000
e3	0.0000000	0.9999985	0.0000013	0.0000002

e4	0.0000000	1.0000000	0.0000000	0.0000000
e5	0.0000000	1.0000000	0.0000000	0.0000000
sc1	0.0000000	0.0000701	0.9999299	0.0000000
sc2	0.0000000	0.0000000	1.0000000	0.0000000
sc3	0.0000000	0.0000000	1.0000000	0.0000000
sc4	0.0000000	0.0000267	0.9999717	0.0000016
sc5	0.0000000	0.0000000	1.0000000	0.0000000
sp1	0.0000000	0.0000000	0.0000000	1.0000000
sp2	0.0000000	0.0000000	0.0000000	1.0000000
sp3	0.0000000	0.0000000	0.0000000	1.0000000
sp4	0.0000000	0.0000000	0.0000000	1.0000000
sp5	0.0000000	0.0000000	0.0000000	1.0000000

5. CONCLUSION AND FUTURE WORK

The essence of email filtering is text classification. Whether it is spam filtering or network public opinion analysis, it can be regarded as the two-classification problem of short text. In short text classification, most Chinese texts have the problem of sparse text and high-dimensional features; At the same time, the Bayesian classification model has the problems of feature limitation and the assumption of conditional independence between attributes does not exist. The high dimension of features, the limitation of features and the non-existence of the assumption of conditional independence of classification model have become important factors restricting short text classification. In order to reduce the adverse impact of the above defects on short text classification, combining with the actual situation of spam filtering, the naive Bayesian classification algorithm was improved, and achieved good results.

In this study, because BETSY is only able to process text content, so we focus on text-based e-mail classification. In other words, if e-mails contain graphic or other object, that would make incorrect result of classification. On the other side, because we take e-paper (news) as training samples, it means that the training samples have organized well, so that I can get high precision results with a few samples. In other words, if the training samples do not process first, BETSY may use more samples for training in order to get satisfied outcome.

Funding details

This paper is supported by the Special fund for electronic information engineering technology specialty group of national double high program of Dongguang Polytechnic in 2022(ZXD202208, ZXD202204), Dongguan Science and Technology of Social Development Program in 2020 (2020507156694), Special Fund for Science and Technology Innovation Strategy of Guangdong Province in 2021 (Special Fund for Climbing Plan) (pdjh 2021a0944), Various Scientific Research Projects Carried Out in 2020 in Colleges and Universities of the Education Department of Guangdong Province (No. 2020KTSCX 320), 2020 School-level Research Fund Key Project of Dongguan Polytechnic (2020a19), Special Projects in Key Fields of Colleges and Universities in Guangdong Province in 2021 (2021ZDZX1093), Dongguan Science and Technology Commissioner Project (202018005 00362), 2021 Special projects in Key Fields of Colleges and Universities in Guangdong

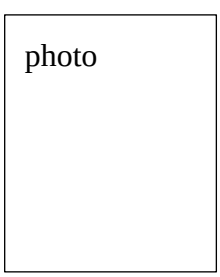
Province (2021ZDZX1092), 2021 Engineering Technology Center of Colleges and Universities in Guangdong Province (2021GCZX016), Special Projects in Key Fields of Colleges and Universities in Guangdong Province in 2021(2021ZDZX1146), 2021 Special projects in Key Fields of Colleges and Universities in Guangdong Province (2021ZDZX 1119), Dongguan Science and Technology of Social Development Program in 2021 (20211 800900252), Special fund for electronic information engineering technology specialty group of national double high program of Dongguang Polytechnic in 2021 (ZXYYD001), Special fund for electronic information engineering technology specialty group of national double high program of Dongguang Polytechnic in 2021 (ZXF002), Special fund for electronic information engineering technology specialty group of national double high program of Dongguang Polytechnic in 2022 (ZXB202203), Special fund for electronic information engineering technology specialty group of national double high program of Dongguang Polytechnic in 2022 (ZXC202201). Dongguan Science and Technology Commissioner Project (20201800500362), Special fund for electronic information engineering technology specialty group of national double high program of Dongguang Polytechnic in 2022-Violence Detection Method Based on Convolution Neural Network and Trajectory (2022H21). Dongguan Social science and Technology Development Project under Grant (2020507156696), 2022 Special innovation Projects of Universities in Guangdong Province, Research on key technologies of Automatic Guided Vehicle (AGV) Based on Deep learning and machine vision (2022KTSCX327), Dongguan Science and Technology of Social Development Program (20221800905842), Dongguan Science and Technology Ombudsman Project (20221800500452), Special Fund for Science and Technology Innovation Strategy of Guangdong Province (Special Fund for Climbing Plan) (pdjh2023b10 20), Guangdong Higher Vocational Education Teaching Reform Research and Practice Project (GDJG2021007), Dongguan Science and Technology Ombudsman Project in 2022 (20221800500812), 2023 The annual government, school, industry and enterprise project of the electronic information engineering technology specialty group of the national double high program of Dongguan Polytechnic (zxd202302). This paper is the mid-term research result of the 2022 Guangxi University Teachers' Basic Research Ability Improvement Constructionist's of multimodal transport smart logistics system based on the new western land-sea corridor under the influence of RCEP (code: 2022KY1252), Dongguan Science and Technology of Social Development Program (20231800900011).

REFERENCES

1. L. Chen, H. Zhang, J. M. Jose, *et al.*, "Topic detection and tracking on heterogeneous information," *Journal of Intelligent Information Systems*, Vol. --, 2017, pp. 1-23.
2. S. Xu, "Bayesian naïve bayes classifiers to text classification," *Information Science*, Vol. 44, 2018, pp. 48-59.
3. S. K. Card, T. P. Moran, and A. Newell, *The Psychology of Human Computer Interaction*, Lawrence Erlbaum Associates, Hillsdale, NJ.
4. W. Cohen, "Learning rules that classify e-mail," in *Proceedings of AAAI Spring Symposium on Machine Learning in Information Access*, 1996, pp. -- --.
5. H. Drucker, D. Wu, and V. N. Vapnik, "Support vector machines for spam categorization," *IEEE Transactions on Neural Networks*, Vol. 10, 1999, pp. -- --.

6. D. Harris and H. Clark, "Worldtalk releases first Internet e-mail corporate usage report; concludes e-mail abuse at epidemic levels," Technical Report???, ----, ----.
7. http://www.worldtalk.com/Corporate%20Information/press%20releases/iecur.htm, 1999.
8. J. Helfman and C. Isbell, "Ishmail: Immediate identification of important information," Technical Report???, ----, ----.
9. http://www.research.att.com/~jon/ishmail, 1995.
10. T. Jaakkola, M. Meila, and T. Jebara, "Maximum entropy discrimination," Technical Report AITR-1668, MIT, 1999.
11. T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proceedings of European Conference on Machine Learning*, 1998, pp. -- --.
12. K. Lang, "Newsweeder: Learning to filter netnews," in *Proceedings of 12th International Conference on Machine Learning*, 1995, pp. -----.
13. D. D. Lewis and K. A. Knowles, "Threading electronic mail: A preliminary study. Information Processing and Management, 33(2):209–217, 1997.
14. A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification. In AAAI-98 Workshop on Learning for Text Categorization, 1998.
15. Howard J, Ruder S. Universal Language Model Fine-tuning for Text Classification [J]. 2018.
16. M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A bayesian approach to filtering junk e-mail. In Proceedings of Workshop on Learning for Text Categorization, 1998.
17. S. Scott and S. Matwin. Feature engineering for text classification. In Proceedings of Sixteenth International Conference on Machine Learning (ICML-99), 1999.
18. R. B. Segal and J. O. Kephart. Mailcat: An intelligent assistant for organizing e-mail. In Proceedings of the Third International Conference on Autonomous Agents, 1999.
19. R. B. Segal and J. O. Kephart. Incremental learning in swiftfile. In Proceedings of the Seventeenth International Conference on Machine Learning (ICML-00), 2000.
20. Y. Yang, T. Ault, T. Pierce, and C. W. Lattimer. Improving text categorization methods for event tracking. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2000.
21. Y. Yang and J. Pedersen. Feature selection in statistical learning of text categorization. In Fourteenth International Conference on Machine Learning (ICML-97), 1997.
22. Jason D. M. Rennie. ifile: An Application of Machine Learning to E-Mail Filtering. KDD-2000 Text Mining Workshop Boston, MA USA.
23. Rodrigues A S, Pereira C A D B, Polpo A. Estimation of component reliability in coherent systems with masked data [J]. IEEE Access, 2019, 7: 57476-57487.
24. Fan X, Guo W L, Sun J. Reliability of high-voltage Ga N-based light-emitting diodes [J]. IEEE Transactions on Device and Materials Reliability, 2019, 19(02): 402-408.
25. Ali S, Ali S, Shah I, *et al.* Reliability analysis for electronic devices using generalized exponential distribution [J]. IEEE Access, 2020, 8:108629-108644.
26. Zhou K，Cruise J R, Dent C J, *et al.* Bayesian estimates of transmission line outage rates that consider line dependencies [J]. IEEE Transactions on Power Systems. 2021, 360(02):1095-1106.

27. Pan, W., Zhao, Z., Huang, W., Zhang, Z., Fu, L., Pan, Z., Yu, J., and Wu, F. (2022). Video moment retrieval with noisy labels. IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2022.3212900.
28. Ma, L., Zheng, Y., Zhang, Z., Yao, Y., Fan, X. and Ye, Q. (2022) Motion Stimulation for Compositional Action Recognition, IEEE Transactions on Circuits Systems and Video Technology, 2022, Early Access.
29. Fu, L., Zhang, D. and Ye, Q, (2021) Recurrent Thrifty Attention Network for Remote Sensing Scene Recognition, IEEE Transactions on Geoscience and Remote Sensing, vol.59, no.10, pp. 8257-8268.
30. Ye, Q., Huang, P., Zhang, Z., *et al.*, (2022) Multi-view Learning with Robust Double-sided Twin SVM with Applications to Image Recognition, IEEE Transactions on Cybernetics, vol.52, no.12, pp.12745 - 12758.
31. Fu, L., Li, Z. and Ye, Q., *et al.*, (2022) Learning Robust Discriminant Subspace Based on Joint L2,p- and L2,s-Norm Distance Metrics, IEEE Transactions on Neural Networks and Learning Systems, vol.33, no.1,pp.130 -144.
32. Chen, X., Li, M., Zhong, H., Ma, Y. and Hsu, C. (2022) DNNOff: Offloading DNN-based Intelligent IoT Applications in Mobile Edge Computing. IEEE Transactions on Industrial Informatics, 18(4): 2820-2829.
33. Chen, X., Zhang, J., Lin, B. and Zheyi Chen, Z (2022) Katinka Wolter, Geyong Min. Energy-Efficient Offloading for DNN-based Smart IoT Systems in Cloud-Edge Environments. IEEE Transactions on Parallel and Distributed Systems, 33(3): 683-697.
34. Chen, X., Hu, J., Chen, Z. and Lin, B. (2022) Naixue Xiong, Geyong Min. A Reinforcement Learning Empowered Feedback Control System for Industrial Internet of Things. IEEE Transactions on Industrial Informatics, 18(4): 2724-2733.
35. Chen, X., Yang, L., Chen, Z., Min, G., Zheng, X. and Rong, C. (2022) Resource Allocation with Workload-Time Windows for Cloud-based Software Services: A Deep Reinforcement Learning Approach. IEEE Transactions on Cloud Computing, Publish Online, DOI: 10.1109/TCC.2022.3169157.
36. Huang, G., Luo, C., Wu, K., Ma, Y., Zhang, Y, and Liu, X. (2019) Software-Defined Infrastructure for Decentralized Data Lifecycle Governance: Principled Design and Open Challenges. IEEE International Conference on Distributed Computing Systems, 2019.



Yanjun Zhu

Ting Zhu

photo

photo

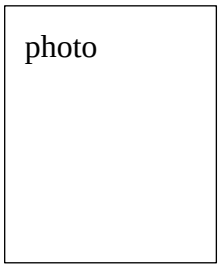
Jianxin Li

photo

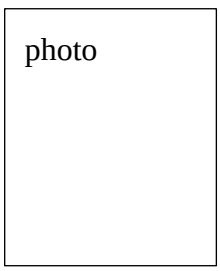
WenLiang Cao

photo

Peng Yong



Fei Jiang



Jie Liu