The Email Author Identification System Based on Support Vector Machine (SVM) and Analytic Hierarchy Process (AHP)

Qinghe Zheng, Xinyu Tian, Mingqiang Yang and Huake Su

Abstract—**At present, more and more crimes are handled by e-mail. The offender's email often contains traces and evidence of the criminal process. Although it is usually very short, it contains obvious evidence of the criminal process. Therefore, how to use it to be reliable evidence and to identify authors is an urgent problem. In this paper, based on reasonable hypothesis, we try to establish a mathematical model to successfully solve this problem by using the combination of analytic hierarchy process (AHP), the SVM intelligent classification model, and the statistical analysis. According to the extracted feature of textual language, we filter out the message set and some representative samples through MySQL. By analyzing the text, we draw five representative features (*i.e.*, word frequency, syntax structure, sentence length, format, and punctuation), which can be used to make up the linear space vector set. We use the improved term frequency–inverse document frequency (TF-IDF) algorithm to calculate the weight of each word and use AHP to re-weight the five elements. Moreover, the space vector model is used to obtain the feature vector of each message. In order to solve the problem of classification model, we use the previously obtained vector set as experimental samples. Then, the multi-class support vector machine (SVM) is used as the final classification model, and the cross-validation is used to determine the model parameters. By randomly partitioning dataset, 80% is used as training set and 20% is used as test set. Finally, experimental results show that the accuracy is more than 95%.**

Index Terms—**e-mail author identification, support vector machine, term frequency–inverse document frequency, analytic hierarchy process**

I. Introduction

Nowdays, with the development of computer science and information technology, especially the popularity of the Internet, email has become one of the most indispensable and convenient means for the information exchange [1] [2] [3] [4]. Unfortunately, online email abuse occurs from time to time, such as spam, fraud, threat mail, reactionary mail and so on. In these emails, the senders always try to hide their real identity in order to avoid reconnaissance. Through anonymous mail servers, the senders can change or forge their addresses, their real names and so on. Therefore, it is difficult to find out the true identity of the author of the email itself [4] [5] [6] [7]. Studying the method of determining the true identity of the original author of e-mail provides the key basis for computer forensics [8] and criminal investigation [9] of the criminal responsibility of illegal authors of e-mail, which undoubtedly provides an effective way to control illegal e-mail behavior [10] [11].

The phenomenon of counter-propaganda, fraud, extortion, terrorist threats, pornography, viruses and spam sent by e-mail has become increasingly serious, which is causing more and more harm. Gartner's survey report [12] show that the direct losses caused by e-mail fraud in the United States reached $1.2 billion in 2003. According to the US Fraud Information Center [13], the number of e-mail fraud cases was as high as 22% between January and June in 2004. In China, the mail server of Dongfang.com was used by criminals as a vicious incident to promote Falun Gong movements in 2001, which caused extremely bad effects [14]. In 2004, cases of extorting banks and enterprises by email were frequently encountered. It endangers the security and stability of society. To this end, countries have introduced relevant laws. For example, in the United States [15] [16], e-mail scammers can be convicted of telecommunications fraud. Similarly, China [17] formulated the plan of "Measures for the Administration of International Network Security Protection for the Computer Information Networks", and the General Office of Ministry of Education [18] has also issued a series of related notices, such as the "Emergency Notice on Further Strengthening the Prevention of Spam E-mail Prevention" (Information Hall Letter 2004-3) [19] [20].

At present, China has become the hardest hit and victim of spam. According to a survey by the Internet Society of China [21], from August 2014 to April 2015, the average number of spammers received by Chinese netizens was 14.5 per week, and the number of spam messages accounted for 54.44% of the total number of emails received. The largest amount of Internet spam sent is in the United States and China, and the output of Internet spam in the United States is more than double the number of other countries in the world [22] [23]. According to the survey [24] released by Gartner, a US-based cybersecurity company and market research firm, the 20055 report stated that 57 million Americans received emails about phishing.

Nowadays, researchers at home and abroad mainly study the author identification of e-mail from two aspects. On the

Manuscript received October 1, 2018; revised March 1, 2019. This work is supported by the Fundamental Research Funds of Shandong University (Grant 2018JC040), National Natural Science Foundation of China (Grant 61571275), and Shandong Provincial Natural Science Foundation (Grant ZR2014FM010 and ZR2014FM030).

Qinghe Zheng and Mingqiang Yang (corresponding author) are with the School of Information Science and Engineering, Shandong University, Jimo, Qingdao 266237, China. (email: yangmq@sdu.edu.cn).

Xinyu Tian is with the College of Mechanical and Electrical Engineering, Shandong Management University, Changqing, Jinan 250357, Shandong, China.

Huake Su is with College of Microelectronics, Xidian University, Xian 710126, Shanxi, China.

(Advance online publication: 27 May 2019)

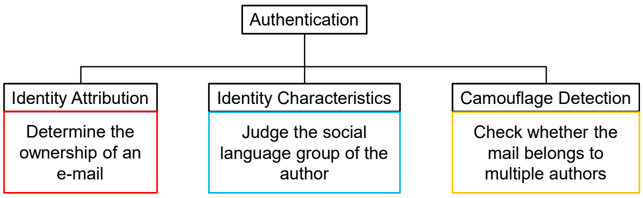

Authentication
Identity Attribution
Determine the ownership of an e-mail
Identity Characteristics
Judge the social language group of the author
Camouflage Detection
Check whether the mail belongs to multiple authors

Fig. 1. The relationship between identity attribution, characteristics, and plagiarism detection.

one hand, in the research of obtaining the author's identity from the physical information such as the header information and addresses of the e-mail, a set of schemes for tracking the author's identity is proposed, but the effect is not satisfactory [25] [26]. On the other hand, it attempts to identify the author by studying the mail content mail structure, language features, and others [27] [28] [29]. The latter has been accumulated for some time in foreign countries, but it has only just started in China. Although it is still far from the point of actual forensics, and many problems still need further study, it has taken a gratifying step. Computer forensics lays a good foundation and provides a good basis for the transformation of phased research results [30].

The application of machine learning to the identification of e-mail authors has become a hot topic in this field in China in recent years, and some preliminary research results have been obtained [31] [32] [33] [34]. However, on the one hand, the recognition algorithms used by predecessors are limited to support vector machines, and only little kinds of classification methods are used [35] [36]. At present, there is no application system for the identification of Chinese e-mail authors, which performs very bad in terms of the experimental conditions and the conversion rate of the research results.

In addition to the related research on email identity analysis, some researchers are also working on other aspects of email, such as mail sorting [37] and mail filtration [6] [38]. The main idea is to use some machine learning methods, such as Naive Bayes algorithm [38], rule-based method [39] [40], to extract various characteristics of the email. But its main purpose is to classify by mail subject to reduce the time of manual sorting of mail, therefore, these characteristics The extraction method is difficult to study in the classification of mail authors.

The study of email sender identity analysis can be divided into three distinct sub-problems: identity attribution, identity characteristics, and plagiarism detection [7]. The relationship between them and their research contents are shown in Fig. 1. Among them, the task of identity attribution is to determine the author of the email, and collect evidence from some mail samples written by the same author to prove that some mails are out of the author's hand. Identity features [10] are used to determine the author's sociolingual characteristics, including their gender, age, education, language, and so on. Plagiarism detection [8] is used to calculate the similarity of two or more messages without determining the attribution of the message. The main purpose is to determine whether a message has been deliberately disguised.

In this paper, based on the reasonable hypothesis, we try to establish a mathematical machine model to successfully solve the problem by using the combination of analytic hierarchy process (AHP), the SVM intelligent classification model, and the statistical analysis. According to the extracted feature of textual language, we first filter out the message set and some representative samples through MySQL. Then, by analyzing the content of text, we draw five representative features (*i.e.*, word frequency, syntax structure, sentence length, format, and punctuation), which can be used to make up the linear space vector set. Finally, we make use of the improved term frequency–inverse document frequency (TF-IDF) algorithm to calculate the weight of each word and use AHP to re-weight the five elements. Moreover, the space vector model is used to obtain the feature vector of each message. In order to solve the problem of classification model, we use the previously obtained vector set as the experimental samples. Then, the multi-class SVM is used as the final classification model, and the cross-validation is used to determine the parameters. By randomly partitioning dataset, 80% is used as training set and 20% is used as test set. Finally, experimental results show that the accuracy is more than 95%.

The rest of this paper is organized as the follows. We first introduce and analyze some related works of email author recognition in Section II. Then, the model assumption and system framework are presented in Section III. The model building and solving process are introduced in Section IV. Model evaluation method and experimental results are given in Section V. Finally, we discuss our conclusions, what we learned, and future works in Section VI.

II. Related Works

In this section, we introduce the related works of email author recognition method, including some basic concepts, the email characteristics, and related algorithms and machine leaning models.

A. Basic Concepts.

Text categorization is completed by creating classification model and then using that model to classify text documents of unknown categories into predefined corresponding categories. Its applications are broad, including the search engines [9], automatic grading of online resources [11], spam filtering [13], document style recognition [12], author discrimination [14], etc. Its history can be traced back to the 1960s, but it was not until the 1990s that the emergence of various electronic documents became a hot topic. The accuracy of the modern text classification system is comparable to that of trained professionals, thanks to the information retrieval and machine learning techniques. A machine learning-based approach was developed in the 1990s to construct classifiers [22]. There are many machine learning methods for constructing classifiers. And the main techniques include: probability and statistics methods [27], decision trees [41], decision rules, regression models [42], artificial neural networks [43], KNN [44], SVM [45], and so on.

The text classification is divided into two steps. The first step is to build a classification model using the training data set, whose key is to extract effective features and use these features to build the model. The second step is to classify the documents of the unknown category with that established machine model. In the whole classification process, the key technologies are divided into text preprocessing, effective feature extraction and representation.

The earliest research on the authorship of email was by Professor Olivier of Australia [22]. He tried to use the support vector machine as a classification algorithm to identify and

TABLE I
NOTATION OF SYMBOLS IN THIS PAPER

Symbols	Notations
w_i	The weight of the i-th feature item in document d
α	The learning rate
L	The loss function
N	The total number of training documents
DF_i	The number of documents containing feature t_i in all training documents
n	The dimension of the feature vectors
$LTF_i(t_i,d)$	The weights sum of the feature t at various locations in document d
$T(L)$	The weight coefficient in the different places
$TF(t_i,d,L)$	The occurrences number of the characteristic term t_i at a certain position L
BY_{Ti}	The scale factor
$T_{i,max}$	The maximum value in the feature
$T_{i,min}$	The smallest of the features
LB_{Ti}	The lower limit of proportion
UB_{Ti}	The upper limit of the ratio
$T_i(i=1,\ldots,15)$	Feature vectors
A_{ij}	Unbiased components of the feature
B_1	Word frequency
B_2	Grammar structure
B_3	Sentence length
B_4	Format
B_5	Punctuation
w_i	Weights
λ_{max}	The largest characteristic root
$C.I.$	Consistency index
$R.I.$	Mean randomness consistent indicators
x_n	Data samples
y_n	The ground-truth label
R^d	d dimensional Euclidean space
H	Classification surface
w	Slope
μ	Intercept
b	Classification threshold

attribute the author of the mail by analyzing the language features and structural features of the mail. Weng *et al.* [46] verified that the e-mail text can be classified into the analysis author by the support vector machine method according to the style of the work, and the accuracy rate can reach 85%. The experiments showed that the functional words that are not related to the content of the mail topic are most suitable for inferring the author's identity. The two-gram feature has good resolution, but this feature set is more biased towards content differentiation. On the other hand, more than 200 words can be used to analyze the author's identity, and the subject has little influence on the classification. The structure and format features can be used to successfully identify the author. By analyzing and comparing the effects of various features on the test results, it can be seen that feature selection step is the most important basis of analysis. However, their research is limited to English mail. The research on emails in other languages has not been carried out, and the recognition accuracy is not satisfactory. Juan *et al.* [47] conducts an author identification study of Japanese heterogeneous text files and explored the sequence word patterns of the articles. He uses the support vector machine as the classification algorithm, and achieves the ideal effect on the experiment of mail, but there are still problems in the classification of Japanese features. And it is worth noting that the proposed method for Japanese feature extraction is difficult to use in other languages. Because Japanese and other languages are very different. For example, Japanese characters include Japanese kanji, hiragana and katakana, and their word order is completely different from that of Chinese, so the method is not suitable for the author analysis of Chinese text.

B. Email Characteristics.

In addition to the characteristics of the email headers, the content and structure of email documents have the following characteristics compared to ordinary text documents:

➢ There are unstructured data in emails such as pictures, sounds, etc.

➢ The content of the email text usually includes a large amount of irrelevant information, such as the respect words, modal particles, etc. The language text is not formal and the colloquialization is serious.

➢ The content of the e-mail content is small, the words are concise, the writing style is relatively free, the sentences and paragraphs are few, and there are a lot of grammatical errors compared with the ordinary text.

➢ Like regular letters, the email documents have a fixed writing format.

In fact, the Chinese emails have other characteristics over English emails:

● The encoding way of Chinese mail is different from the English mail. The plain text content of English mail is ASCII code, while the plain text content of simplified Chinese mail is generally gb2312 code.

● The words in each sentence of the Chinese emails are usually connected and not separated. Therefore, it is difficult to extract and recognize the direct words in the process of handling English mail. Therefore, it is necessary to separate the Chinese sentences before they can be better used to extract and handle words.

● The format of Chinese emails is also very different from that of English email. For example, the English mail is usually followed by a comma after the greeting, then the next line is signed, and the style is more casual. By contrast, the Chinese message generally ends with an exclamation mark or a period, and then writes a title on the other line, and then writes the author's name.

● Due to cultural differences, the content of Chinese mail is also very different from that of English mail. For example, the name of an English mail is generally an English name or a nickname, and the name of the Chinese email is generally a distinguished name with a tribute title.

C. Related Optimization Models and Algorithms.

The research on Chinese text classification has achieved certain results in recent years [42], but the classification core algorithm is mainly realized by referring to related methods of foreign text classification, such as KNN [23], neural network [24] or support vector machine [26]. In addition, combined with the characteristics of Chinese mail, the first thing needs to be solved is the Chinese text representation and the Chinese character encoding double-byte extraction problem, followed by the Chinese word-specific word segmentation steps. Since the Chinese characters in each sentence in Chinese text are connected together before and after, in order to better analyze Chinese text automatically, it is necessary to perform word segmentation in advance, and then mark the word segment after the word segmentation, and then automated extraction work is available. In the Chinese text classification, Harbin Institute of Technology [34], Chinese Academy of Sciences [34], Peking University [35], Shanghai Jiaotong University [35], and Xiamen University [38] have been opened earlier. Some of them have set up language research institutes and obtained certain academic research results, such as applying vector space models, rough sets and support vector array method for text classification.

In terms of the identification system of text classification, the first application to mention is the Construe system [48] that developed by Carnegie Group for Reuters. Although the method achieves a good classification effect, the system has the disadvantages of difficult classification rules and poor generalization, and it is difficult to promote the application on the large scale datasets. In recent years, several systems for text classification have appeared, such as application support vector machines and vector space model methods. However, these studies [49] [50] [52] are mainly limited to the field of text categorization. For the optimization methods [27] [51] of feature selection, extraction, and classification for Chinese email, the system needs further research and verification. At present, there are still no systems developed for text mining and author identification.

In terms of author identification of Chinese email, most of China's email research is oriented to the identification and the processing of the spam. For example, the Bayesian method is more commonly used to deal with spam research, and a small amount of mail feature representation is introduced, while the research work on other aspects of mail is currently less.

In the field of Chinese email author identification system, the research achievements have been made, including feature selection and extraction of email, experiment of email author identification, application-oriented improvement of algorithm and so on. In addition, there are extensive research results, such as the application of intelligent email filtering. Although some research results in this field can be used for reference, email classification has its own characteristics compared with traditional classification. On the other hand, there is no end to the research work. There are some left-over problems, new problems or problems that need to be studied and optimized, such as the design and development of research-based system specially used for the identification of Chinese email authors, the selection and attempt of various classification tools, and the optimization of classification algorithms, especially the multi-classification methods. There are many improvements and developments need to be further studied.

III. Model Assumption and System Framework

A. Problem Restatement and Analysis

More and more email crime cases have made us have to pay attention to how to identify the attribution of emails . Due to the short wording property of the email, the author's language characteristics will be obvious. However, how to reasonably obtain the language features of the emails and improve the accuracy of authentication is an urgent problem that needs to be solved.

The problem to be solved in this paper is to build an effective model that identifies the author by capturing the linguistic features of the email and uses the mail data table to train and test the constructed model.

When people are writing, they will invisibly reveal their own writing habits. because the e-mails are short, and thus they make the author's language features more pronounced. Therefore, one of the key of the problem is that how to extract the reasonable and effective features to identify the author's attribution. We need to consult the relevant literature to find out the reasonable features that can be used for quantification, and make a reasonable combination of these features. In order to make the identification results more accurate, we utilize the multi-class SVM based algorithm for data mining, which is very popular currently. After the email is sort out, the emails are studied as training samples to ensure that the test results can be accurately classified.

B. Model Assumption

Generally, language style refers to the different language materials and methods that people use when they interact, according to different communication occasions, purposes, tasks, and the nature and quality of the communicators. It can be divided into four categories: daily oral style, applied style, artistic style, and individual language style. Each style has its own set of vocabulary, grammar, phonetics, and rhetorical elements. Therefore, we assume that each person's language style maintains the original habit and does not change over a long period of time in this paper.

C. System Framework

The main contents of the overall system framework of this study are as follows.

(1) Analyze the writing style of Chinese mail, and study the corresponding feature extraction, representation, and utilization methods.

We study the characteristics of Chinese emails and analyze the characteristics that represent its writing style in detail, including formatting, structure, language, and statistics. Then we study the intrinsic connections of various features, analyze their interactions and their status in email identification, and establish new quantitative models. Finally, we study feature extraction techniques, feature representation and utilization methods comprehensively, and improve and develop feature selection and evaluation algorithms.

(2) Study the classification algorithm for email author identification.

We study the training methods for email samples. Since the samples of the email are not as much as the text, the suitable algorithm for small samples classification must be selected. According to the domestic and foreign research reports [37], the SVM classification algorithm is considered to be the most effective model, so this study uses multi-class SVM as the final classification algorithm for email attribution.

(3) Analyze the validity of the feature.

Through experiments, the analysis verifies whether SVM is suitable for Chinese email author classification, as well as its optimal kernel function and optimal parameter selection. By comparing the effectiveness of various characteristics to the identity of the email author, we can therefore be able to find the best combination way of multiple features.

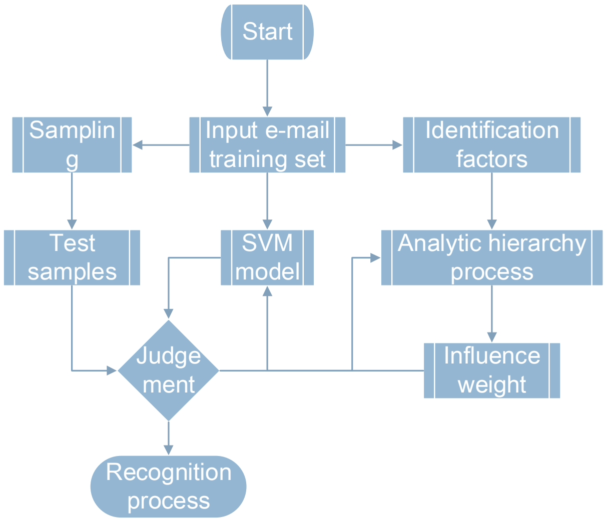


Fig. 2. The overall process of email author classification system.

(4) Try to study the verification of new mail categories. For the test documents whose contents or categories are fundamentally different from the predefined text document, it should be a new category after classification. However, in the current text categorization, it is often forcibly assigned to a class similar to its characteristics according to the established algorithm, resulting in misclassification. In order to solve this problem in Chinese email author classification task, we try to use the F-test method to verify the new mail author category problem.

Finally, all the symbols used in the paper are organized in Table I for the reader's easy understanding.

IV. Model Building and Solving

A. Algorithm Flow: General Steps to Classify the E-mail Author

In this section, we first show the following overall steps established by the algorithm, and the overall process is shown in Fig. 2.

● Step one: the standard classified email documents are prepared as training samples.

● Step two: after the preprocessing of the email documents, we need to extract the useful information and store them in a text file in a fixed format. Then, through the feature extraction process, we can obtain the feature vector of the mail document. And we should make the feature vector convert into the fixed format of the space vector form.

● Step three: according to AHP, the weight of each feature item can be obtained and the space vector form corresponding to all feature items is integrated into a vector by weight.

● Step four: through the SVM algorithm, we can find the optimal hyperplane constructed by the feature vector of the email documents in the high-dimensional space, and therefore the classification model is formed to construct the classifier.

● Step five: as dealing with the training documents, we can extract the feature vector of email document to be categorized, and convert them into the form of a fixed format of space vector.

● Step six: through the above mentioned SVM classifier, the mails that need to be classified are automatically classified into the predefined author categories.

B. Feature Extraction Based on E-mail Features

E-mail text contents are described in natural language, with unstructured features, so the computer is difficult to deal with its semantics. How to convert the contents of the text to a computer can handle the format is the difficult work which must be faced by e-mail identification.

1) Language Feature Extraction.

According to the research of related literature, we found that words in the email cannot be ignored as just a linguistic feature. After preprocessing the email, a statistical method is used to construct the evaluation function. The effect of feature extraction depends mainly on the evaluation function, where we choose the improved term frequency-inverse document frequency (TF-IDF). TF-IDF uses the ratio of times that the feature word appears in the document *d* with the number of documents containing the feature word as the weight of the term. In addition, in order to eliminate the influence of the document length on the weights of feature items, normalized formulas are generally used to represent the weights of feature items. For email, the text content of the beginning of the text reflects the authors' personal characteristics, which is better than the ending content of the characteristic words. Therefore, the calculation method of the weights of feature words in different locations should be different. By giving different coefficients to the feature words in different locations, the effect of expressing the feature words can be further improved. Therefore, in order to improve the TF-IDF algorithm, we define the following formula to obtain the weight:

$$w_i = \frac{LTF_i(t_i,d)\times\log(\frac{N}{DF_i})}{\sum_{i=1}^{n}[LTF_i(t_i,d)\times\log(\frac{N}{DF_i})]^2} \tag{1}$$

and

$$LTF_i(t_i,d)=\sum_{L} TF(t_i,d,L)\times T(L) \tag{2}$$

where w_i is the weight of the *i*-th feature item in document *d*. *N* is the total number of training documents. DF_i indicates the number of documents containing feature item t_i in all training documents. *n* represents the dimension of the feature vectors. LTF_i (t_i, *d*) represents the sum of the weight of the feature item t_i at each position in the document *d*. *L* means that the text appears in different boundaries in the same location. *T*(*L*) is the weight coefficient assigned to the different positions, and $T(L) \geq 1$, The size of *T*(*L*) can be adjusted experimentally, *TF* (t_i, *d, L*) indicates the number of occurrences of the feature word t_i at a certain position *L*.

2) Structural Feature Extraction.

When we look up the relevant information, we find that the classification results only based on the word frequency are not strong enough. Therefore, we propose a feature extraction method, which is to study the relevant structural features of e-mail and mainly refers to syntactic structure and sentence length.

Based on literature review and related research work [32] [33] [36] [47], this paper aims to provide five commonly used statistical syntax structures, as shown in Table II. In order to guarantee the same processing of all features, the scale factor

TABLE II
FIVE COMMONLY USED STATISTICAL SYNTAX STRUCTURES

Project	Sentence pattern
T_1	Be doing
T_2	Be done
T_3	Have/has done
T_4	Can/be able to
T_5	It is that

TABLE III
THE PUNCTUATION MARKS OF COMMAS, PERIODS, COLONS, EXCLAMATION MARKS, QUESTION MARKS, SEMICOLONS AND OTHER SIX COMMON PUNCTUATION MARKS

Project	Punctuation pattern
T_6	Comma
T_7	Period
T_8	Colon
T_9	Exclamation mark
T_{10}	Question mark
T_{11}	semicolon

TABLE IV
THE FEATURE ITEMS ACCORDING TO THE ENGLISH FILE FORMAT

Project	Format
T_{12}	Date
T_{13}	Call
T_{14}	End honorific
T_{15}	Sender signature

BY_{Ti} is introduced according to

$$BY_{Ti} = \frac{UB_{Ti} - LB_{Ti}}{T_{i,\max} - T_{i,\min}} \tag{3}$$

where $T_{i,max}$ and $T_{i,min}$ are the maximum and minimum value in the features, respectively. LB_{Ti} and UB_{Ti} are defined as the lower and upper limits of the ratio, respectively. And UB_{Ti}=1, LB_{Ti}=0. Then the corresponding eigenvalue belongs between 0 and 1, *i.e.*,

$$T_i^B = (T_i - T_{i,\min})BY_{Ti} + LB_{Ti} \tag{4}$$

Through email programming, we can count the length of each sentence. Here we use the average sentence length as one of the features, as given by

$$Average\ sentence\ length = \frac{The\ total\ number\ of\ words\ in\ the\ body\ of\ e\text{-}mail}{Number\ of\ sentences} \tag{5}$$

$$Punctuation\ ratio = \frac{The\ number\ of\ specific\ punctuation\ marks}{The\ total\ number\ of\ punctuation\ marks} \tag{6}$$

Then we calculate the punctuation marks of commas, periods, colons, exclamation marks, question marks, semicolons and other six common punctuation marks, as shown in Table III. The normalization of the eigenvalues also completed by the usage of equations (3) and (4).

3) Format Feature Extraction.

According to the English file format, we further extract the following feature items in Table IV. if there is a feature item, the flag is 1, otherwise it is set to 0. The normalization of the eigenvalues also completed by the usage of equations (3) and (4).

C. Hierarchical Analysis to Determine the Weight of Each Part

At first, B_1, B_2, B_3, B_4, and B_5 are defined as the word frequency, the grammar structure, the sentence length, the format, and the punctuation. Then the degree of importance scale and contrast matrix are given in Table V and Table VI, respectively.

Each column is normalized according to

$$B_{ij} = \frac{A_{ij}}{\sum_{n=1}^{5} A_{nj}} \tag{7}$$

Then we can get the normalized features, as shown in Table VII. Finally, we calculate the weight of indicator: normalized eigenvector *w*, as given by

$$w_{ij} = \frac{B_i}{\sum_{n=1}^{5} B_n} \tag{8}$$

The obtained results are shown in Table VIII. On the next step, we calculate the largest eigenvalue of the matrix according to matrix consistency test:

$$\lambda_{\max} = \frac{\sum (AW)_i}{nW_i} \tag{9}$$

where *AW* indicates that matrix *A* is multiplied by *W*. The meaning is that the result of multiplying two matrices is the column vector, then each element is divided by the product of the order and the corresponding weight. Finally, we can get λ = 8.09428 through calculating the consistency indicator *C.I.* = 0.7357 according to

$$C.I. = \frac{\lambda_{\max} - n}{n-1} \tag{10}$$

where *n* represents the order of the matrix.

$$C.R. = \frac{C.I.}{R.I.} \tag{11}$$

Among them, *R.I.* is the average randomness consistent index, which is 1.12 by the randomized consensus index table, and C.R. is 0.0817 < 0.1. That is, to maintain a significant level, the contrast matrix is used to maintain consistency, its validity and the reliability. Finally, word frequency, syntax structure, sentence length, format, and punctuation weight are 0.4663, 0.1676, 0.1014, 0.1432, and 0.1214, respectively, as shown in Fig. 3.

D. SVM-Based Classification Algorithm

SVM algorithm was proposed by Vapnik *et al.* in 1995, which is one of the most famous machine learning algorithm

TABLE V
THE DEGREE OF IMPORTANCE SCALE

	B_1	B_2	B_3	B_4	B_5
B_1	1	5	3	7	9
B_2	0.2	1	1/3	3	5
B_3	1/3	3	1	0.2	1/7
B_4	1/7	1/3	5	1	3
B_5	1/9	1/5	7	1/3	1

TABLE VI
CONTRAST MATRIX

	B_1	B_2	B_3	B_4	B_5	Sum
B_1	1	0.5	0.3	0.02	0.08	2
B_2	0.5	1	0.15	0.1	0.25	2
B_3	0.3	0.15	1	0.3	0.25	2
B_4	0.02	0.1	0.3	1	0.58	2
B_5	0.08	0.25	0.25	0.58	1	2
Sum	2	2	2	2	2	10

TABLE VII
THE NORMALIZED FEATURES

	B_1	B_2	B_3	B_4	B_5	Sum
B_1	0.5595	0.5245	0.1837	0.6069	0.4961	2.3707
B_2	0.1119	0.1049	0.0204	0.2601	0.2756	0.7729
B_3	0.1865	0.3147	0.0612	0.0173	0.0079	0.5876
B_4	0.0799	0.0350	0.3061	0.0867	0.1654	0.6731
B_5	0.0622	0.0210	0.4286	0.0289	0.0551	0.5959
Sum	1	1	1	1	1	5

TABLE VIII
THE NORMALIZED EIGENVECTOR

Value	B_1	B_2	B_3	B_4	B_5	Sum
w	0.4663	0.1676	0.1014	0.1432	0.1214	1
w/%	46.63%	46.76%	10.14%	14.32%	12.14%	100%

based on the statistical learning theory. Through automatic learning algorithms, SVM can automatically find those who have better classification ability of the support vector. The classifier thus constructed can maximizes the class-to-class spacing and therefore has the better adaptability and higher resolution. At present, SVM algorithm has already become a new research hotspot in the field of machine learning at home and abroad, and has been applied to face recognition [22] [23], document recognition [17] [19], document classification [27] [28], handwriting recognition [33], etc.

The essence of SVM is to solve the quadratic programming problem (the objective function is a quadratic function and the constraint is a linear constrained optimization problem), and the global optimal solution can be therefore obtained. This gives unparalleled benefits to some other statistical learning techniques. SVM is developed from the optimal classification under linear separability. The basic concept can be illustrated by the two types shown in Fig. 4. Suppose the training set is defined as $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, and $x_i \in R^d$, $y_i \in \{-1, 1\}$, i=1, 2, …, n. R^d represents the D-dimensional Euclidean space and y_i denotes the class designation for the two types of samples. In the figure, solid dots and hollow dots represent two types of samples, respectively, and H is a classification surface. H_1 and H_2 are called the classification interval. The so-called best classification line is that the classification of the text author is required to not only correctly separate the two categories, but also can be used to get largest classification interval.

The classification line is defined as

$$g(x) = wx + \mu \tag{12}$$

Then the equation for the constraint solving is given by

$$wx + \mu = 0 \tag{13}$$

The discriminant function is normalized so that all two types of samples satisfy: $|g(x)| \geq 1$. Let $|g(x)| = 1$, the classification interval is $2/|W|$, so the largest classification interval is making $|W|^2$ smaller. If the classification line is required to correctly classify all samples, it is required to meet

$$y_i[wx_i + \mu] - 1 \geq 0, i = 1, 2, ..., n \tag{14}$$

If above conditions is met and make $|W|^2$ smaller, the smallest classification surface is the best classification surface. The training samples on the hyperplane of the multi types of data that are closest to the classification surface and the parallel optimal classification surfaces are those samples that have the equality numbers in equation (16) called support vectors. In Fig.4, the support vector is between H_1 and H_2. The optimal interface can be seen as a quadratic optimization problem, the Lagrange multiplier method can be turned into a dual problem [38], *i.e.*, the constraint is transformed into

$$\sum_{i=1}^{n} a_i y_i = 0, a_i \geq 0 \; and \; i = 1, 2, ..., n \tag{15}$$

And the maximum value of the function is given by

$$W(a) = \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i,j=1}^{n} a_i a_j y_i y_j (x_i * x_j) \tag{16}$$

which is a quadratic function optimization problem under the inequality constraints where includes a unique solution. The non-zero solution a_i is the optimal solution, which is denoted as a^*. The corresponding sample is the support vector, the optimal classification function is

$$f(x) = \text{sgn}\{wx + \mu\} = \text{sgn}\{\sum_{i=1}^{n} a_i * y_i (x_i \mu) + \overline{\mu}\} \tag{17}$$

where μ is the classification threshold, which can be obtained by taking the median of any pair of support vectors in both categories. sgn () is a symbol. For a given unknown sample x, we can determine the category that x belongs through simply entering the calculation,.

In this paper, the e-mail file is segmented, and multiple feature are extracted, then the SVM classification algorithm is used to construct the classifier for writing feature. It exhibits many unique advantages in solving small sample, nonlinear and high dimensional pattern recognition problems, which is very suitable for fewer email samples and has many types of features.

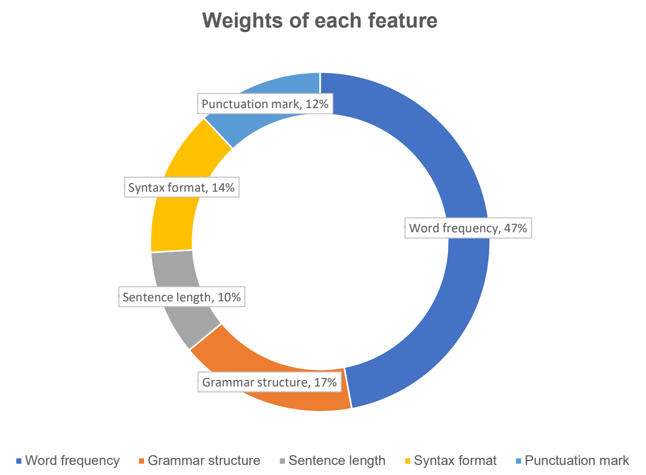


Fig. 3. The proportion of word frequency, syntax structure, sentence length, format, and punctuation weigh.

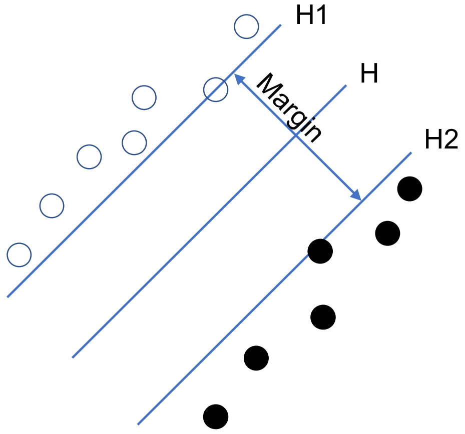


Fig. 4. The description of SVM classification hyperplane.

E. Solving Process of SVM

The task of the SVM is to find such a hyperplane H to split the sample into two parts without errors and to maximize the distance between H1 and H2. To find such a hyperplane, we can simply maximize the interval margin, which is equivalent to minimizing $\|w\|^2$. Then we can construct the following conditional extreme value problem:

$$\min \frac{\|w\|}{2}, s.t.\ y_i(wx_i+\mu)-1\ge 0 \tag{18}$$

For the conditional extremum of inequality constraints, the Lagrangian method can be used to solve the problem. The Lagrange equation is constructed according to multiplying the constraint equation by a non-negative Lagrangian coefficient and then subtracting it from the objective function. So the Lagrangian equation is as follows:

$$\begin{aligned} L(w,b,\alpha_i) &= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l}\alpha_i(y_i(wx_i+\mu)-1) \\ &= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l}\alpha_i y_i(wx_i+\mu)+\sum_{i=1}^{l}\alpha_i \end{aligned} \tag{19}$$

where

$$\alpha_i \ge 0 \tag{20}$$

Then the planning problem we have to deal with becomes:

$$\min_{w,\mu}\max_{\alpha_i\ge 0} L(w,\mu,\alpha_i) \tag{21}$$

The equation (21) is an expression of Lagrangian conditional extremum of strict inequality constraints, which is a convex programming problem. The significance is to first derive the partial derivative of α, let it equal to 0 to eliminate α, and then find the minimum value of L for w and μ. It is difficult to solve (21) directly, but this problem can be solved by Lagrangian dual problem. Then, we can make equivalent transformation of (21) as

$$\min_{w,\mu}\max_{\alpha_i\ge 0} L(w,\mu,\alpha_i) = \max_{\alpha_i\ge 0}\min_{w,\mu} L(w,\mu,\alpha_i) \tag{22}$$

The above equation is dual transformation, so that this convex programming problem is transformed into the dual problem solution:

$$\max_{\alpha_i\ge 0}\min_{w,\mu} L(w,\mu,\alpha_i) \tag{23}$$

The significance of this equation is that the original convex programming problem can be transformed into the first partial derivation of w and b, so that it equals to 0 to eliminate w and b, and then find the maximum value of L for α. Then we solve the (23). For this we first calculate the partial derivatives of w and b. By (19) we can get

$$\begin{cases} \dfrac{\partial L(w,\mu,\alpha_i)}{\partial w} = w - \displaystyle\sum_{i=1}^{l}\alpha_i y_i x_i \\ \dfrac{\partial L(w,\mu,\alpha_i)}{\partial \mu} = -\displaystyle\sum_{i=1}^{l}\alpha_i y_i \end{cases} \tag{24}$$

In order for L to take the minimum value on w and b, the two partial derivatives of (24) are 0, so that we can get

$$\begin{cases} w = \displaystyle\sum_{i=1}^{l}\alpha_i y_i x_i \\ \displaystyle\sum_{i=1}^{l}\alpha_i y_i = 0 \end{cases} \tag{25}$$

Substituting (25) back to (29), we can get:

$$\begin{aligned} \min_{w,\mu} L(w,\mu,\alpha_i) &= \frac{1}{2}\|w\|^2 - w\sum_{i=1}^{l}\alpha_i y_i x_i - \mu\sum_{i=1}^{l}\alpha_i y_i + \sum_{i=1}^{l}\alpha_i \\ &= \frac{1}{2}\|w\|^2 - w\cdot w - \mu\cdot 0 + \sum_{i=1}^{l}\alpha_i \\ &= \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\|w\|^2 \\ &= \sum_{i=1}^{l}\alpha_i - \sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j(x_i x_j) \end{aligned} \tag{26}$$

And through substituting (26) into (23), we can get:

TABLE IX

THE CLASSIFICATION RESULTS OF MULTIPLE FEATURES

Author	Precision (%)					Recall (%)					F1-value (%)				
	L	LF	LS	FS	LFS	L	LF	LS	FS	LFS	L	LF	LS	FS	LFS
1	94.85	97.50	**97.36**	96.69	95.89	85.39	87.39	86.38	87.36	**89.65**	**97.31**	94.90	96.10	95.99	96.27
2	95.16	94.65	**96.85**	95.07	94.82	85.22	85.14	85.51	86.40	**88.39**	97.17	96.28	**97.47**	96.78	96.78
3	95.71	**96.97**	95.67	95.31	95.48	84.60	86.73	**87.42**	85.57	85.15	95.24	95.31	95.01	**96.56**	95.45
4	96.72	97.00	**97.09**	95.33	96.87	84.68	86.85	86.28	86.07	**87.42**	96.90	95.55	95.35	95.93	**97.37**
5	**96.60**	95.24	96.21	96.39	94.86	86.85	86.84	84.92	**87.39**	85.49	**95.98**	94.75	95.30	95.92	94.90
6	**96.58**	96.23	94.6	96.53	96.34	85.19	86.3	86.69	86.57	**88.00**	96.08	96.20	95.85	95.36	**96.57**
mAP	95.94	96.27	**96.30**	95.89	95.71	85.32	86.54	86.20	86.56	**87.35**	**96.45**	95.50	95.85	96.09	96.22

TABLE X

THE CLASSIFICATION RESULTS OF MULTIPLE KERNEL FUNCTIONS: K_1 TO K_5

Author	Precision (%)					Recall (%)					F1-value (%)				
	K_1	K_2	K_3	K_4	K_5	K_1	K_2	K_3	K_4	K_5	K_1	K_2	K_3	K_4	K_5
1	95.00	96.80	**97.72**	96.54	96.86	85.80	87.01	85.64	**87.78**	85.49	96.83	96.43	**98.35**	94.86	96.76
2	94.52	94.70	**97.00**	95.85	94.09	85.79	**87.75**	86.02	86.64	87.74	93.65	**95.41**	95.37	94.44	94.04
3	**96.21**	96.07	96.13	94.35	95.18	84.10	**87.11**	**87.11**	85.83	85.08	95.69	**97.65**	94.09	94.12	95.25
4	96.91	96.21	**97.47**	94.43	96.75	84.77	85.29	85.29	**85.90**	87.96	97.67	98.06	**99.05**	96.78	98.01
5	**97.42**	95.12	96.04	95.67	95.72	86.34	85.67	**88.67**	88.13	85.03	94.65	96.38	**97.80**	96.70	95.72
6	96.63	96.24	93.63	96.54	**96.68**	85.43	85.89	85.89	86.31	88.87	96.46	95.58	96.03	**97.37**	97.11
mAP	96.11	95.86	**96.33**	95.56	95.88	85.37	86.33	86.94	**87.76**	87.53	95.83	96.58	**96.78**	95.71	96.15

TABLE XI

THE CLASSIFICATION RESULTS OF MULTIPLE KERNEL FUNCTIONS: K_6 TO K_{10}

Author	Precision (%)					Recall (%)					F1-value (%)				
	K_6	K_7	K_8	K_9	K_{10}	K_6	K_7	K_8	K_9	K_{10}	K_6	K_7	K_8	K_9	K_{10}
1	94.81	97.38	**97.62**	97.60	96.01	85.71	86.89	85.90	88.70	**90.61**	94.77	97.26	97.87	**98.51**	96.13
2	95.55	95.59	95.87	**96.06**	95.53	86.18	85.69	85.04	87.63	**88.45**	95.93	96.53	94.89	**97.04**	96.24
3	95.29	**96.13**	96.10	95.10	95.37	83.68	86.27	**87.54**	85.62	84.97	94.87	95.29	**96.52**	94.88	95.27
4	96.63	**97.36**	96.76	97.30	94.01	84.69	85.65	84.96	86.85	**87.15**	96.55	**97.72**	96.44	97.23	95.25
5	96.72	94.32	96.05	**97.30**	94.01	86.46	84.75	85.51	**89.04**	84.18	96.84	93.40	95.88	**98.20**	93.16
6	96.42	**96.91**	93.74	95.92	95.67	85.27	86.57	85.03	85.70	**88.20**	96.26	**97.60**	92.89	95.32	95.00
mAP	95.90	96.28	96.02	**96.38**	95.44	85.34	85.97	85.66	87.25	**87.26**	95.87	96.30	95.75	**96.86**	95.17

TABLE XI

THE CLASSIFICATION RESULTS OF MULTIPLE KERNEL FUNCTIONS: K_{11} AND K_{12}

Author	Precision (%)		Recall (%)		F1-value (%)	
	K_{11}	K_{12}	K_{11}	K_{12}	K_{11}	K_{12}
1	93.47	**96.85**	86.20	**86.27**	93.09	**97.34**
2	**93.55**	92.82	**86.72**	84.21	**94.63**	93.28
3	95.53	**96.00**	84.95	**87.47**	93.75	**95.42**
4	95.53	**96.76**	**85.29**	84.93	95.37	**96.89**
5	**95.04**	94.83	**85.37**	**85.37**	**95.93**	95.13
6	96.44	**95.79**	85.27	**86.50**	**95.61**	94.72
mAP	94.92	**95.51**	85.63	**85.79**	94.73	**95.46**

$$\max_{\alpha_i\ge 0}\min_{w,\mu} L(w,b,\alpha_i)=\max_{\alpha_i\ge 0}\left[\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j (x_i x_j)\right] \tag{27}$$

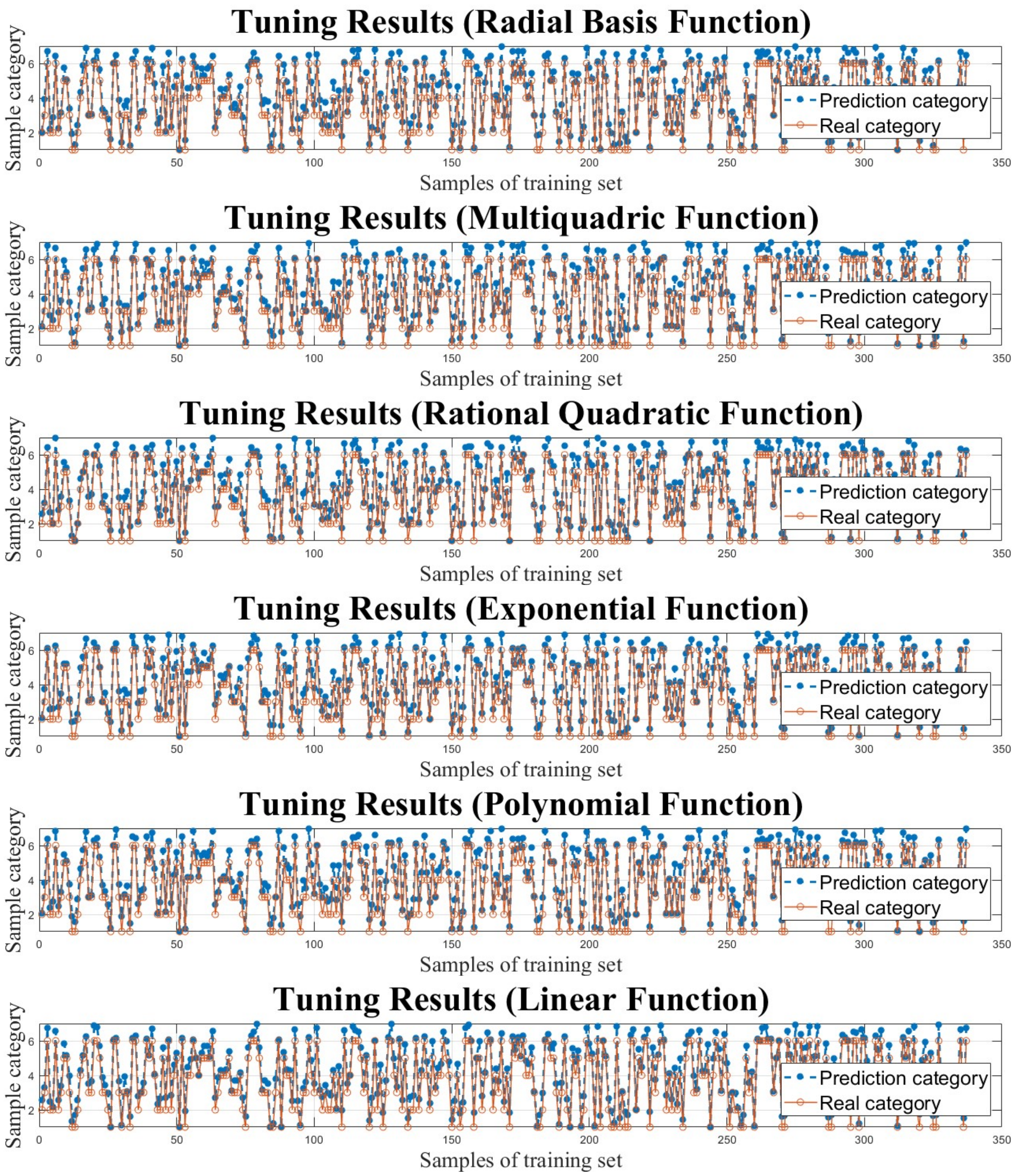
Taking into account (25), our dual problem becomes:

$$\max_{\alpha_i\ge 0}\left[\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j (x_i x_j)\right], s.t.\begin{cases}\sum_{i=1}^{l}\alpha_i y_i=0\\ \alpha_i\ge 0\end{cases} \tag{28}$$

The planning problem of (28) can be solved directly from the numerical method. And it is worth noting that the conditional extreme value problem of (18) can be transformed into the convex programming problem , *i.e.*, problem in (25), which implies a constraint, namely:

$$\alpha_i(y_i(wx_i+\mu)-1)=0 \tag{29}$$

This constraint is obtained in the following way. If (19) and (23) are equivalent, there must be:


Tuning Results (Radial Basis Function)

Tuning Results (Multiquadric Function)

Tuning Results (Rational Quadratic Function)

Tuning Results (Exponential Function)

Tuning Results (Polynomial Function)

Tuning Results (Linear Function)

Fig. 5. The tuning results of training samples under SVM with different kernel functions.

$$\max_{\alpha_i \geq 0} L(w,\mu,\alpha_i) = \frac{1}{2}\| w \|^2 \tag{30}$$

Substituting (3) into the above equation yields:

$$\begin{aligned}\frac{1}{2}\| w \|^2 &= \max_{\alpha_i \geq 0}\left\{\frac{1}{2}\| w \|^2 - \sum_{i=1}^{l}\alpha_i \left[y_i(wx_i+\mu)\right] - 1)\right\} \\ &= \frac{1}{2}\| w \|^2 - \min_{\alpha_i \geq 0}\left\{\sum_{i=1}^{l}\alpha_i \left[y_i(wx_i+\mu)\right] - 1)\right\}\end{aligned} \tag{31}$$

It can be simplified to

$$\min_{\alpha_i \geq 0}\left\{\sum_{i=1}^{l}\alpha_i \left[y_i(wx_i+\mu)\right] - 1)\right\} = 0 \tag{32}$$

And based on the constraints (18) and (21), we can get

$$\alpha_i \left[y_i(wx_i+\mu)\right] - 1) \geq 0 \tag{33}$$

So, in order to make (30) true, only to satisfy

$$\alpha_i \left[y_i(wx_i+\mu)\right] - 1) = 0 \tag{34}$$

This gives the constraint of the equation (29). The meaning of constraint is that if sample is support vector, its corresponding Lagrangian coefficient is non-zero; if sample is not a support vector, its corresponding Lagrangian coefficient must be zero. This shows that most Lagrangian coefficients are zero. Once we have solved all Lagrangian coefficients from (28), we can get it by:

$$w' = \sum_{i=1}^{l}\alpha_i y_i x_i \tag{35}$$

The normal vector w' of the optimal segmentation plane H can be obtained by the iterative calculation. The segmentation threshold μ' can also be calculated by the constraint vector of (29), such as

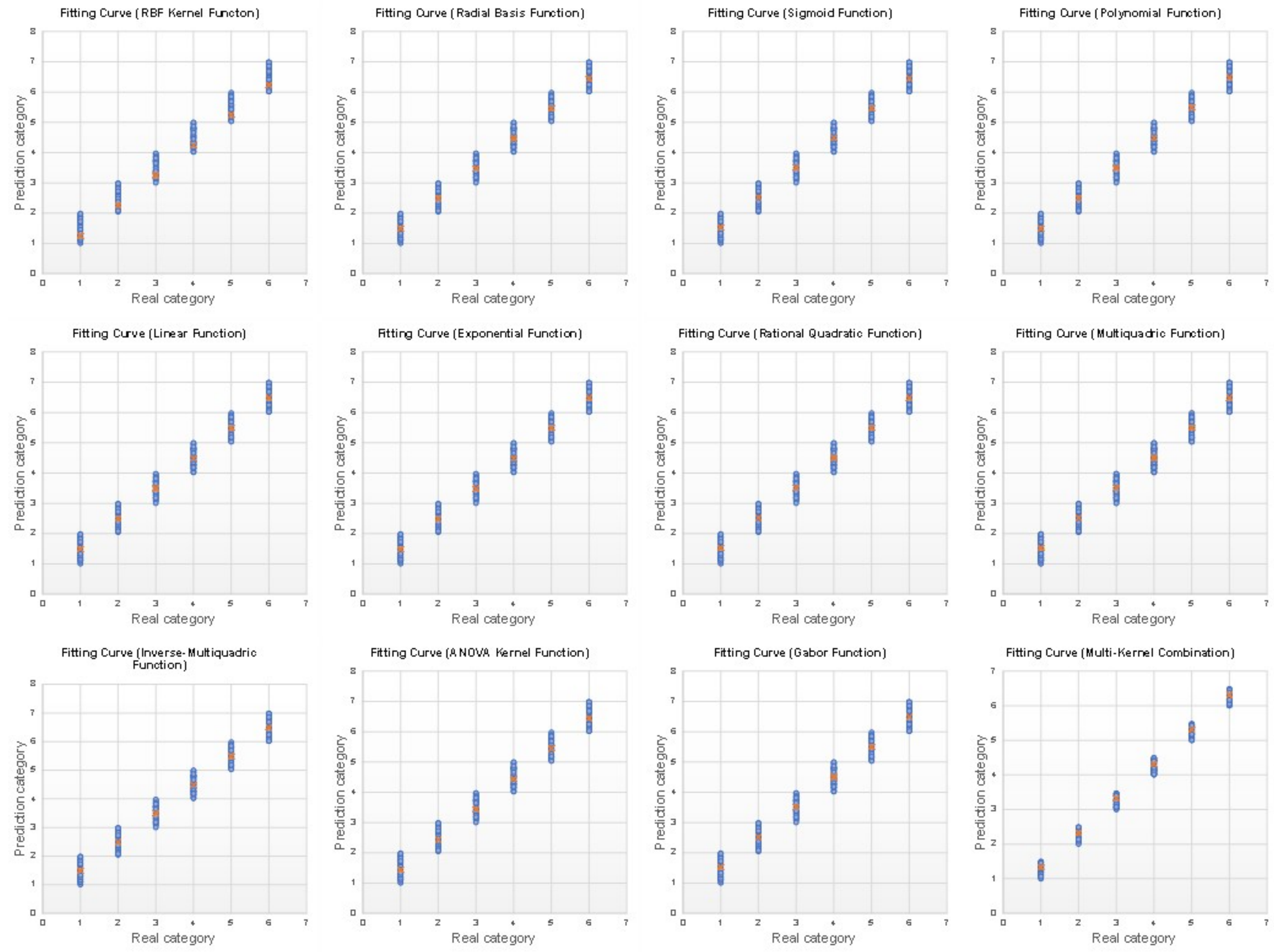


Fig. 6. The fitting cures of training samples under SVM with different kernel functions.

$$\mu' = -\frac{\max\limits_{i:y_i=-1} w'^T x_i + \min\limits_{i:y_i=1} w'^T x_i}{2} \qquad (36)$$

Finally, we found the best H1 and H2, which is the SVM we trained.

V. Experiments and Model Evaluation

A. Experimental Setup

We design an SVM between any two types of samples, so the six categories of samples need to design 15 SVMs. When an unknown sample is classified, the category with the most votes last is the belonging category of the unknown sample. Therefore, we do not need to retrain all SVMs, just retrain and add sample-related classifiers. When training a single model, the relative speed is often faster. In the experimental process, hyper-parameters such as learning rate are searched by grid method [46]. The grid search method gives the adjustment range and adjustment step size of each parameter, calculates the possible values of each parameter, and then traverses all the combinations to return the best parameter values.

Take the SVM with RBF kernel as the example. The hyper parameters we need to adjust at this time are the regularization parameter and the kernel function parameter gamma, which have a scope of 10^{-8}-10^{8} and 10^{-6}-10^{6}. And the regularization parameter represents the penalty coefficient of the model for the error, gamma reflects the distribution of the sample after mapping to the high-dimensional feature space; the larger the regularization parameter, the easier the model is over-fitting; the smaller the regularization parameter, the easier the model is to fit. The larger the gamma, the more support vectors, the smaller the gamma value, and the less the support vector. The smaller the gamma, the better the generalization of the model, but if it is too small, the model will actually degenerate into a linear model; the larger the gamma, the theoretical SVM can fit any nonlinear data.

By sorting the samples of the email, the authors' statistical information and corresponding mail information are obtained, representative samples and the corresponding features are extracted. We use fixed vector formats and weights to form training sets and label sets.

B. Experimental Results

1) Comparison of Multiple Features.

We first compare the classification results of the SVM algorithm under different features, including precision, recall and F1 score. The experimental results are shown in Table IX, where L represents language feature, LF is the combination of language feature and format feature, LS is the combination of language feature and structural feature, FS is the combination of format feature and structural feature, and LFS stands for the combination of language feature, format feature and structural feature. The values with bolded black in the table represent

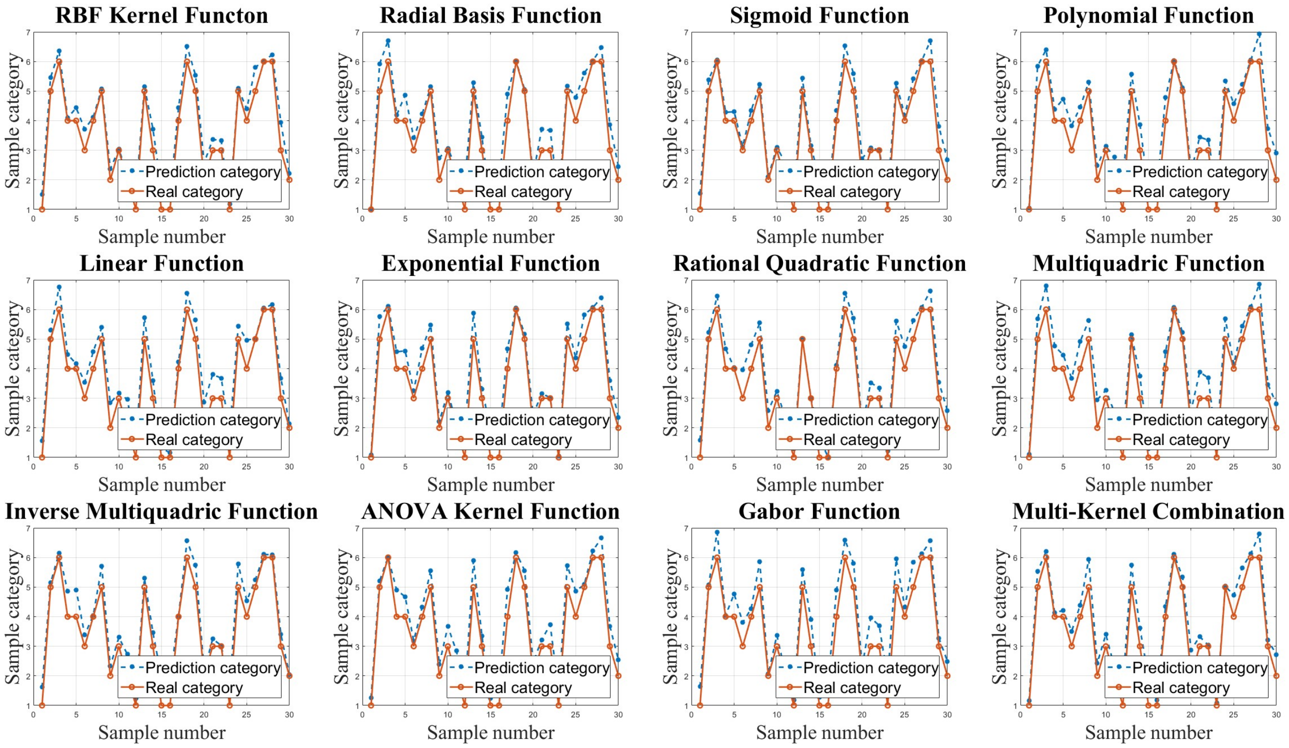


Fig. 7. The tuning results of validation samples under SVM with different kernel functions.

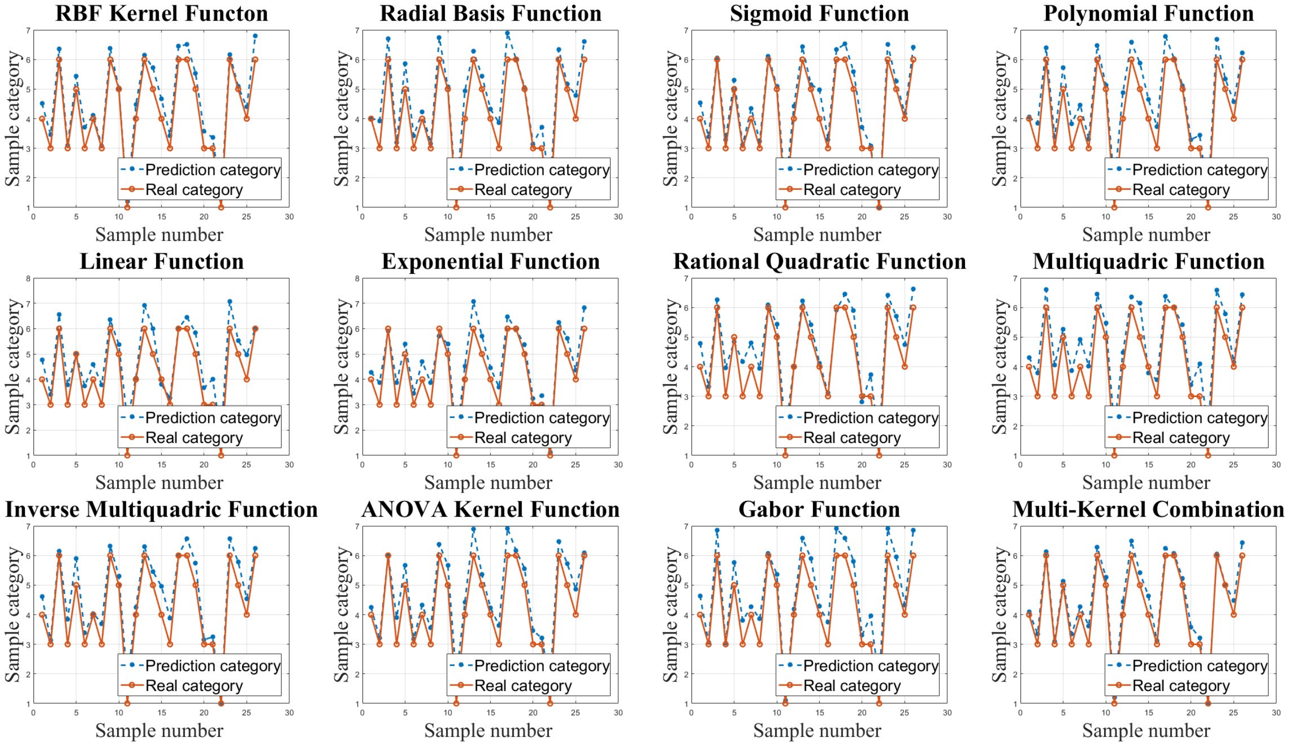


Fig. 8. The prediction results of test samples under SVM with different kernel functions.

the best experimental results. We can see that LS shows the best results in terms of precision, and LFS performs best in terms of recall. For the F1 score, there is almost no difference in the combination of different features. The best performance of the five feature combinations are 97.31, 97.47, 96.51, 97.37, and 98.58, respectively.

2) Comparison of Multiple Kernel Functions.

In this part, we compare the classification results of various kernel functions, including the RBF kernel function, Radial basis function, Sigmoid, Polynomial function, Linear function, Exponential function, ANOVA kernel function, Rational quadratic function, Gabor function, Multiquadric function, Inverse-multiquadric function, and Multi-kernel combination, as shown in Tables X, XI, and XII. The precision values of the 12 kernel functions are 96.11%, 95.86%, 96.33%, 95.56%, 95.88%, 95.90%, 96.28%, 96.02%, 96.38%, 95.44%, 94.92%, and 95.51%, respectively. The recall values of the 12 kernel functions are 85.37%, 86.33%, 86.94%, 87.76%, 87.53%, 85.34%, 85.97%, 85.66%, 87.25%, 87.26%, 85.63%, and 85.79%, respectively. The F1 score of the 12 kernel functions are 95.83%, 96.58%, 96.78%, 95.71%, 96.15%, 95.87%, 96.30%, 95.75%, 96.86%, 95.17%, 94.73%, and 95.46%,

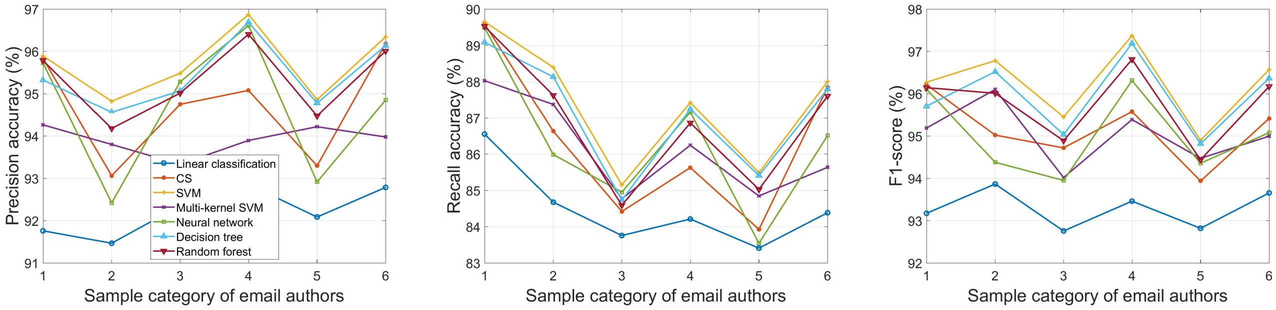


Fig. 9. The comparison of precision accuracy, recall accuracy, and F1 score of various classification algorithms.

respectively. Different kernel functions show complex results on precision, recall and F1 score, and it is difficult to observe the rules that can be summarized. However, it is worth noting that the results of combined kernel functions tend to be more robust and stable.

3) Tuning Results of Training Set and Validation Set.

In this part, we present the tuning results of samples in the training set and the corresponding fitting curves, as shown in Fig. 6 and Fig. 7, respectively. It can be seen that the SVMs with different kernel functions have almost the same fitting results on the training set, that is, they can all exhibit very good fitting ability. However, they show some differences in the validation set, as shown in Fig. 8. In fact, there are some kernel functions that perform well, such as ANOVA kernel function, Rational quadratic function, and the Gabor function. There are also some kernel functions that are disappointing, such as Polynomial function and Linear function. It is worth noting that multi-kernel combination has the robust and stable performance.

4) Prediction Results of Testing Set.

In this part, we present the prediction results of samples in the test set, as shown in Fig. 8. As can be seen from the figure, most of the classification results are accurate and the correct rate is over 90%.

5) Comparison of Multiple Classification Methods.

In this part, we compare the precision accuracy of various classification algorithms, including linear classification, CS, SVM, multi-kernel SVM, neural network, decision tree, and random forest, as shown in Fig. 9. It can be seen that the linear classification shows the worst performance with the lowest precision accuracy, recall accuracy and F1 score. By contrast, The multi-kernel SVM we used in the experiments performs the best, with highest precision accuracy, recall accuracy and F1 score. The performance of neural networks and random forest can also be accepted.

C. Advantages and Improvements

The classification algorithm in this paper uses the simple SVM model, and the input data after special processing will inevitably produce a certain distortion. In practice, we should consider a variety of factors, such as delivery time, sub forms, email fonts, and more. Based on the model, an appropriate data correction algorithm can be introduced to improve the extracted features. At the same time, more precise intelligent classification algorithms, such as deep learning and genetic algorithms, can be used to improve the classification accuracy of the machine learning model. In addition, this algorithm has the following advantages: it uses kernel functions to map the features to high dimensional spaces; it can be used to solve the nonlinear classification problem by using kernel functions; the idea of the classification is simple, that is, to maximize the interval between the sample and the decision surface.

VI. Conclusion

At present, more and more crimes are handled by e-mail. The offender's email often contains traces and evidence of the criminal process. Although it is usually very short, it contains obvious evidence of the criminal process. Therefore, how to make it to be reliable evidence and to identify authors is an urgent problem. In this paper, based on reasonable hypothesis, we try to establish a mathematical model to successfully solve this problem by using the combination of analytic hierarchy process (AHP), the SVM intelligent classification model, and the statistical analysis. By analyzing the text, we draw five representative features (*i.e.*, word frequency, syntax structure, sentence length, format, and punctuation), which can be used to make up the linear space vector set. We use the improved TF-IDF algorithm to calculate the weight of each word and use AHP to re-weight the five elements. Moreover, the space vector model is used to obtain the feature vector of each message. In order to solve the problem of classification model, we use the previously obtained vector set as experimental samples. Then, the multi-class multi-kernel SVM is used as the final classification model, and the cross-validation is used to determine the model parameters. By randomly partitioning dataset, 80% is used as training set and 20% is used as test set. Finally, experimental results show that the accuracy is more than 95%. Finally, we compare various algorithms and prove the effectiveness and superiority of our algorithm.

The algorithm proposed in this paper still has the following problems:

- ● The linear non-separable mapping to high-dimensional space may lead to the super large dimensions, resulting in huge computational complexity.
- ● It is difficult to train large-scale datasets by using SVM algorithm.
- ● The SVM algorithm cannot directly implement multi-classification directly, but it can be done indirectly.
- ● It is sensitive to missing data.

In the future, we will further explore and improve these questions.

References

[1] A. Shi, S. Lim, and K. Lee, "Surface roughness classification using pattern recognition theory," Optical Engineering, vol. 34, no. 6, pp. 1756-1760, 1995.

[2] Q. Zheng *et al.*, "An end-to-end image retrieval system Based on gravitational field deep learning," in *IEEE International Conference*

on Computer Systems, Electronics and Control (ICCSEC), Dalian, China, pp. 936-940, 2017.

[3] Q. Zheng *et al.*, "A bilinear multi-scale convolutional neural network for fine-grained object classification," IAENG International Journal of Computer Science, vol. 45, no. 2, pp. 340-352, 2018.

[4] Q. Zheng *et al.*, "Improvement of generalization ability of deep CNN via implicit regularization in two-stage training process," IEEE Access, vol. 6, pp. 15844-15869, 2018.

[5] A. Zhang, Y. Cai, and X. Xu, "Maximum margin sparse representation discriminative mapping with application to face recognition," Optical Engineering, vol. 52, no. 2, pp. 7202, 2013.

[6] A. Ebers *et al.*, "Study on three-dimensional face recognition with continuous-wave time-of-flight range cameras," Optical Engineering, vol. 50, no. 6, pp. 179-86, 2011.

[7] H. Lodhi *et al.*, "Text classification using string kernels," Journal of Machine Learning Research, vol. 2, no. 3, pp. 419-444, 2002.

[8] P. Zongo, R. Dorville, and E. Gouba, "Method for identifying spatial reservoirs of malaria infection and control strategies," IAENG International Journal of Applied Mathematics, vol. 48, no. 1, pp. 33-39, 2018.

[9] K. Nigam *et al.*, "Text Classification from Labeled and Unlabeled Documents using EM," Machine Learning, vol. 39, no. 2-3, pp. 103-134, 2000.

[10] T. Simon and K. Daphne, "Support vector machine active learning with applications to text classification," Machine Learning Research, vol. 2, no. 1, pp. 999-1006, 2002.

[11] Q. Zhang *et al.*, "Segmentation of hand posture against complex backgrounds based on saliency and skin colour detection," IAENG International Journal of Computer Science, vol. 45, no. 3, pp. 435-444, 2018.

[12] G. Forman, "An extensive empirical study of feature selection metrics for text classification," Journal of Machine Learning Research, vol. 3, no. 2, pp. 1289-1305, 2003.

[13] Z. Tong and F. Oles, "Text Categorization Based on Regularized Linear Classification Methods," Information Retrieval, vol. 4, no. 1, pp. 5-31, 2001.

[14] H. Zhuang *et al.*, "A method for static hand gesture recognition based on non-negative matrix factorization and compressive sensing," IAENG International Journal of Computer Science, vol. 44, no. 1, pp. 52-59, 2017.

[15] S. Kim *et al.*, "Some Effective Techniques for Naive Bayes Text Classification," IEEE Transactions on Knowledge & Data Engineering, vol. 18, no. 11, pp. 1457-1466, 2006.

[16] H. Kim *et al.*, "Dimension Reduction in Text Classification with Support Vector Machines," Journal of Machine Learning Research, , vol. 6, no. 1, pp. 37-53, 2005.

[17] D. Urynbassarova, B. Li, and Z. Zhang, "A Convolution Theorem for the Polynomial Fourier Transform," IAENG International Journal of Applied Mathematics, vol. 47, no. 4, pp. 381-387, 2017.

[18] G. Fung *et al.*, "Text classification without negative examples revisit," IEEE Transactions on Knowledge & Data Engineering, vol. 18, no. 1, pp. 6-20, 2005.

[19] P. Wang *et al.*, "Using Wikipedia knowledge to improve text classification," Knowledge & Information Systems, vol. 19, no. 3, pp. 265-281, 2009.

[20] Q. Zheng *et al.*, "Understanding and boosting of deep convolutional neural network based on sample distribution," in *Proc. IEEE. ITENC*, Chengdu, China, pp. 823-827, 2017.

[21] Q. Zheng *et al.*, "Fine-grained image classification based on the combination of artificial features and deep convolutional activation features," in *IEEE/CIC ICCC*, Qingdao, China, pp. 1-6, 2017.

[22] Q. Zheng *et al.*, "Static hand gesture recognition based on Gaussian mixture model and partial differential equation," IAENG International Journal of Computer Science, vol. 45, no. 4, pp. 569-583, 2018.

[23] A. Sai and N. Kong, "Sparse Grid Interpolation of Ito Stochastic Models in Epidemiology and Systems Biology," IAENG International Journal of Applied Mathematics, vol. 48, no. 1, pp. 45-52, 2018.

[24] S. Argamon *et al.*, "Stylistic text classification using functional lexical features," Journal of the American Society for Information Science & Technology, vol. 58, no. 6, pp. 802-822, 2010.

[25] C. Jiang *et al.*, "Text Classification using Graph Mining-based Feature Extraction," Knowledge-Based Systems, vol. 23, no. 4, pp. 302-308, 2010.

[26] J. Huh, M. Yetisgen-Yildiz, and W. Pratt, "Text classification for assisting moderators in online health communities," Journal of Biomedical Informatics, vol. 46, no. 6, pp. 998-1005, 2013.

[27] A. Sarker and G. Gonzalez, "Portable Automatic Text Classification for Adverse Drug Reaction Detection via Multi-corpus Training," Journal of Biomedical Informatics, vol. 53, pp. 196-207, 2015.

[28] Y. Lin, J. Jiang, and S. Lee, "A Similarity Measure for Text Classification and Clustering," IEEE Transactions on Knowledge & Data Engineering, vol. 26, no. 7, pp. 1575-1590, 2014.

[29] D. Merkl, "Text classification with self-organizing maps: Some lessons learned," Neurocomputing, vol. 21, no. 1–3, pp. 61-77, 1998.

[30] J. Jiang, R. Liou, and S. Lee, "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification," IEEE Transactions on Knowledge & Data Engineering, vol. 23, no. 3, pp. 335-349, 2011.

[31] P. Wang *et al.*, "Improving Text Classification by Using Encyclopedia Knowledge," in *IEEE International Conference on Data Mining*, pp. 332-341, 2007.

[32] A. Uysal and S. Gunal, "The impact of preprocessing on text classification," Information Processing & Management, vol. 50, no. 1, pp. 104-112, 2014.

[33] J. Huh, M. Yetisgen-Yildiz, and W. Pratt, "Text classification for assisting moderators in online health communities," Journal of Biomedical Informatics, vol. 46, no. 6, pp. 998-1005, 2013.

[34] F. Figueiredo *et al.*, "Word co-occurrence features for text classification," Information Systems, vol. 36, no. 5, pp. 843-858, 2011.

[35] C. Wan *et al.*, "A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine," Expert Systems with Applications, vol. 39, no. 15, pp. 11880-11888, 2012.

[36] L. Khreisat, "A machine learning approach for Arabic text classification using -gram frequency statistics," Journal of Informetrics, vol. 3, no. 1, pp. 72-77, 2009.

[37] S. Edeki, O. Ugbebor, and A. Owoloko, "Analytical Solution of the Time-fractional Order Black-Scholes Model for Stock Option Valuation on No Dividend Yield Basis," IAENG International Journal of Applied Mathematics, vol. 47, no. 4, pp. 407-416, 2017.

[38] R. Forsyth and D. Holmes, "Feature-finding for text classification," Literary & Linguistic Computing, vol. 11, no. 4, pp. 163-174, 1996.

[39] H. Li and K. Yamanishi, "Text classification using ESC-based stochastic decision lists," Information Processing & Management, vol. 38, no. 3, pp. 343-361, 2002.

[40] Y. Co and J. Seo, "Text classification from unlabeled documents with bootstrapping and feature projection techniques," Information Processing & Management, vol. 45, no. 1, pp. 70-83, 2009.

[41] A. Kananthai and T. Kraiwiradechachai, "On the White Noise of the Price of Stocks related to the Option Prices from the Black-Scholes Equation," IAENG International Journal of Applied Mathematics, vol. 48, no. 2, pp. 128-133, 2018.

[42] A. Sun *et al.*, "Blocking reduction strategies in hierarchical text classification," IEEE Transactions on Knowledge & Data Engineering, vol. 16, no. 10, pp. 1305-1308, 2004.

[43] Q. Kuang and X. Xu, "Improvement and Application of TFIDF Method Based on Text Classification," Computer Engineering, vol. 32, no. 19, pp. 1-4, 2006.

[44] S. Weng and C. Liu, "Using text classification and multiple concepts to answer e-mails," Expert Systems with Applications, vol. 26, no. 4, pp. 529-543, 2004.

[45] A. Juan and E. Vidal, "On the use of Bernoulli mixture models for text classification," Pattern Recognition, vol. 35, no. 12, pp. 2705-2710, 2002.

[46] S. Wermter, "Neural Network Agents for Learning Semantic Text Classification," Information Retrieval, vol. 3, no. 2, pp. 87-103, 2000.

[47] F. Ren and M. Sohrab, "Class-indexing-based term weighting for automatic text classification," Information Sciences, vol. 236, no. 1, pp. 109-125, 2013.

[48] A. Esuli and F. Sebastiani, "Active Learning Strategies for Multi-Label Text Classification," in *European Conference on Ir Research on Advances in Information Retrieval*, pp. 102-113, 2009.

[49] Q. Zheng and M. Yang, "A Video Stabilization Method based on Inter-Frame Image Matching Score," Global Journal of Computer Science and Technology, vol. 17, no. 1, pp. 35-40, 2017.

[50] Q. Zhang *et al.*, "Segmentation of hand gesture based on dark channel prior in projector-camera system," in *IEEE/CIC ICCC*, Qingdao, China, pp. 1-6, 2017.

[51] Q. Zheng, X. Tian, M. Yang, and H. Wang, "Differential Learning: A Powerful Tool for Interactive Content-Based Image Retrieval," Engineering Letters, vol. 27, no. 1, pp. 202-215, 2019.

[52] Q. Zheng *et al.*, "A multi-resolution mosaic method used for unmanned aerial vehicle (UAV) remote sensing image," Journal of Xi'an University of Posts and Telecommunications, vol. 22, no. 2, pp. 53-59, 2017.