Feature Extraction and Classification of Text Data by Combining Two-Stage Feature Selection Algorithm and Improved Machine Learning Algorithm

Hua Huang
School of Computer and Artificial Intelligence, Henan Finance University. Zhengzhou 450046, China
E-mail: huanghuahafu@163.com

Keywords: Two-stage feature selection, SVM, text data, classification, MRMR

Received: February 26, 2024

Efficient text classification is crucial for information processing due to the generation of massive text data. However, the uneven distribution and redundancy of text data often result in poor classification performance. To address this issue, a two-stage feature selection algorithm is proposed using the fusion of information gain and maximum correlation minimum redundancy algorithm. To improve SVM performance in text data classification, an improved SVM algorithm based on Fourier hybrid kernel function is proposed. The study found that the proposed improved algorithm achieved an accuracy of 0.82 on the IMDB dataset using only 40 feature subsets. Even when the number of features exceeded 390, the F1 value of the proposed algorithm remained 1% to 2% higher than that of other algorithms. The improved algorithm performed best when the feature dimension was around 400. The proposed algorithm, which combines the Fourier hybrid kernel function with a two-stage feature selection algorithm based on the information gain and maximum correlation minimum redundancy algorithm, achieved a 1%~3% higher F1 value and increased the number of correctly classified texts by 20 to 45. These results demonstrate the effectiveness of the algorithm as a classification tool for processing large-scale text data, which is significant for information retrieval and data mining.

Povzetek: Predstavljena sta dvodstopenjski algoritem za izbiro značilk in izboljšani algoritem strojnega učenja za povečanje točnosti klasifikacije besedilnih podatkov. Združujeta informacijski dobiček in metodo minimalne redundance ter maksimalne korelacije (MRMR) z izboljšano SVM.

1 Introduction

As information technology develop, especially in many fields such as medicine, finance, and journalism, the Internet has generated massive amounts of text data. These text data contain a wealth of information and knowledge, significant for improving business decision-making, market analysis, disease diagnosis, etc. However, due to the large and complex volume of these data, it has become a challenging problem to effectively extract useful information from them and perform accurate text classification [1-3]. The core of text classification lies in how to accurately and efficiently identify and classify a large number of unlabeled text data, which directly affects the quality and application effect of information extraction. Firstly, there is a large amount of redundant information in the text data, which is not only irrelevant to the classification task, but will interfere with the judgment of the classifier and reduce the accuracy of classification. Secondly, the feature distribution of text data is often uneven, which makes it difficult for traditional classification algorithms to maintain stable and efficient performance in the face of different types of datasets [4-5]. To this end, a Two-stage Feature Selection (TFS) Algorithm that fuses Information Gain (IG) and improved Minimum Redundancy Maximum Relevance (MRMR) is proposed, and a Fourier hybrid kernel function is introduced to enhance the Support Vector Machine (SVM) effect in text classification. Through these technological innovations, the research aims to process large-scale text data more efficiently and improve the accuracy and efficiency of classification. This has important practical value for information processing and decision support in the fields of medical diagnosis, news topic analysis, and market trend forecasting. The overall structure of the study consists of four parts. The first part summarizes the relevant research results and shortcomings of feature extraction at home and abroad. The second part proposes the fusion of TFS and improved machine learning algorithms. The third part analyzes the experimental results through the proposed algorithm and includes a discussion section related to the current research. The fourth part summarizes the experimental results, points out the shortcomings of the research, and proposes future research directions.

In the field of machine learning, SVM has become one of the core technologies of text classification due to its excellent classification performance. The performance of SVM depends largely on the quality of feature selection and extraction. Feature selection is important when dealing with large-scale text data, and effective feature extraction is essential to improve classification accuracy and efficiency [6]. Here are some of the relevant studies by scientists and scholars. Ahmed Y A et al. proposed a weighted MRMR algorithm for better estimating the feature significance of data captured by cyberattacks. This technique combined enhanced weighted MRMR with frequency inverse document frequency and further accommodates an improved approach to entropy. It was used to evaluate the weights of the features generated by the algorithm. Results showed a good performance of proposed algorithm [7]. Jiménez-Cordero et al. proposed an MRMR-based embedded feature selection method for the trade-off between complexity and classification accuracy. The algorithm used duality theory to reformulate the min-max problem and solved it using off-the-shelf nonlinear optimization software. Compared with public datasets, the proposed method proved its effectiveness and practicability [8]. Wang et al. proposed a SVM kernel function selection mechanism. First, the types of kernel function best suited for the given data were chosen. Then, these types were classified as SVMs. The results showed that the mechanism superiority was verified [9]. Sun et al. proposed a feature selection algorithm for multi-label data with missing labels. Firstly, a multi-label uncertainty measure based on fuzzy neighborhood entropy was proposed, and the MRMR algorithm was improved to evaluate the candidate features. Results showed that this algorithm selected important features with better classification performance [10].

Jia et al. proposed an improved barnacle pairing optimizer combined with an SVM algorithm. The Gaussian mutation and logic model were used to improve the performance of the improved algorithm from different perspectives, and results showed a better performance than other comparison methods. In addition, the model showed significant superiority over other classifiers [11]. Yin et al. proposed an SVM algorithm based on simulated annealing algorithm for the identification of different motion patterns. Firstly, the simulated annealing algorithm obtained the SVM optimal parameters. Then, the MRMR algorithm was used for feature extraction, and the five-layer cross-validation trained the classifier. Results showed that the accuracy of the algorithm was 98% [12]. Bansal et al. proposed a hybrid MRMR feature selection technique using a multi-objective method for automatic sign language recognition. Firstly, the MRMR algorithm was used as a preprocessor to remove redundant and irrelevant features. A multi-class SVM was used as a classifier. The results showed that a more accurate classification was achieved with a decrease in the size of the feature vector [13]. Zhou et al. proposed a feature selection method based on Mutual Information (MI) and correlation coefficients. In this method, the correlation coefficient was first introduced, and then combined with MI to measure features' relationship. To effectively select low redundancy features, minimization was also used in the evaluation criteria. Results showed that the proposed method had good feature classification ability [14].

Table 1: Research status and shortcomings of related works

Related Works		Research findings	Shortcomings
Reference number	Author		
[7]	Ahmed Y A et al	Selecting ransomware attack features through weighted MRMR algorithm	There is no involvement in the field of text classification and a lack of further research.
[8]	Jiménez-Cordero et al	Select features from the dataset using an embedded feature selection method based on MRMR.	There is no involvement in the field of text classification and a lack of further research.
[9]	Wang et al	A SVM kernel function selection mechanism was proposed for bearing fault diagnosis.	Using a single kernel function may not match the data distribution.
[10]	Sun et al	A fuzzy neighborhood entropy based MRMR algorithm was proposed for feature selection.	There is no involvement in the field of text classification and a lack of further research.
[11]	Jia et al	Using SVM algorithm based on improved rattan pot mating optimizer for high-dimensional data testing.	Lack of consideration for data redundancy issues.
[12]	Yin et al	Perform motion pattern recognition using SVM algorithm based on simulated annealing algorithm.	Using a single kernel function may not match the data distribution.

[13]	Bansal et al	Sign language feature selection is performed using a hybrid MRMR feature selection technique, and classification is performed using multi class SVM.	Using a single kernel function may not match the data distribution.
[14]	Zhou et al	Using feature selection method based on MI and correlation coefficient for feature selection.	There is still room for optimization in handling redundant feature problems.

In Table 1, recent research findings and shortcomings are presented. In summary, although many scholars have conducted research on SVM and feature selection in machine learning and applied them to many fields, the existing methods still face the problems of high redundancy, data sparsity and insufficient classification accuracy in processing large-scale text data. To solve the redundancy problem in feature selection, a TFS using the fusion of IG and improved maximum correlation and minimum redundancy is proposed. To further improve the text classification, an improved SVM algorithm based on Fourier hybrid kernel function is proposed. This study has a significant positive effect on improving the accuracy and processing efficiency of text classification [15-19].

Previous studies have addressed the issue of feature redundancy, but there is still room for optimization and improvement. Some studies have focused on feature redundancy but neglected the optimization of classification algorithms. Others have used a single kernel function in classification algorithms, which may result in a mismatch of data distribution. It is important to consider both feature redundancy and algorithm optimization to achieve accurate classification results. Compared to previous studies, this research considers not only the issue of high data redundancy but also the correlation between features and the semantic relationship of the context. This approach is beneficial for improving the accuracy of text feature selection through the TFS algorithm. The classification algorithm employs a hybrid kernel function based on the Fourier kernel function, which overcomes the limitations of a single kernel function. This study is better adapted than previous studies to facilitate classification.

2 Text data feature extraction and classification by integrating two-stage feature selection and machine learning algorithms

In order to improve the text classification and redundancy, a fusion TFS and an improved machine learning algorithm are proposed. Firstly, a TFS based on IG and MRMR algorithms is proposed. On this basis, an improved SVM algorithm is further proposed.

2.1 Text data feature extraction and classification based on two-stage feature selection algorithm

In text classification tasks, it is crucial to select the right features. This process mainly involves removing secondary words and retaining keywords with strong expressiveness to reduce the feature space complexity of text data and avoid the high complexity of dimensions affecting classification performance. In this study, a TFS for IG-MRMR is used to fuse IG and MRMR. Through the IG-MRMR algorithm, the selected feature words are vectorized by text and used by SVM for text classification processing, as shown in Figure 1.

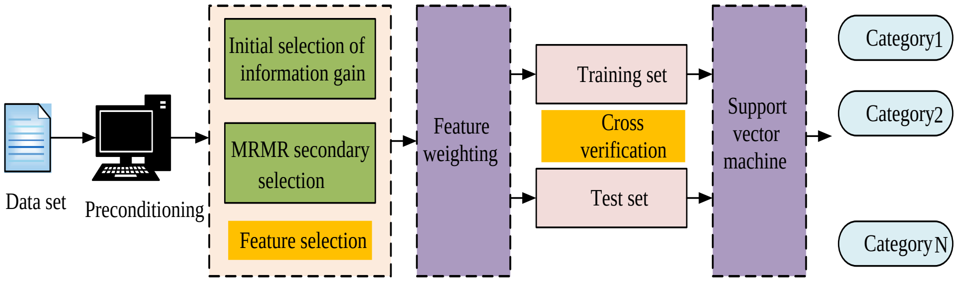


Figure 1: Text classification process based on two-stage feature selection

Figure 1 shows the steps involved in data preprocessing, feature selection, feature weighting, and feature classification. The IG algorithm relies on comparing the difference between the initial entropy of the whole dataset and the conditional entropy under the influence of specific features, so as to determine the effectiveness of the feature in classification, and select the main feature set suitable for text classification. When dealing with text classification, the algorithm involves evaluating occurrence frequency of a feature word t_j in a specific classification C, so as to estimate the IG rate of a feature word t_j, as shown in equation (1).

$$IG(t_i) = P(t_j)\sum_{i}^{m} P(C_i \mid t_j)\log\frac{P(C_i \mid t_j)}{P(C_i)} + P(\bar{t}_j)\sum_{i}^{m} P(C_i \mid \bar{t}_j)\log\frac{P(C_i \mid \bar{t}_j)}{P(C_i)} \quad (1)$$

In equation (1), m is the number of different categories in text data, C_i is the example of the i category in text data, $P(t_i)$ and $P(C_i)$ are the frequency of feature words in the sample text and total text data, and $P(C_i \mid t_j)$ is the probability that the text belongs to C_i under the condition that the feature words. $P(\bar{t}_j)$ refers to the probability that the text does not contain feature words, $P(C_i \mid t_j)$ is the probability that the text belongs to C_i under the condition that there are no feature words. In the process of feature screening, the IG algorithm focuses too much on the number of documents and ignores the importance of word frequency, which leads to the decline of the ability of selected features in prediction and representation. In addition, IG not only considers the existence of feature words, but also pays attention to their absence, mainly focusing on the role of features in classification, ignoring the distribution of features between and within categories. Therefore, the feature set selected by IG needs to be further optimized. The MRMR algorithm is a filtering method using spatial search, which calculates the relevance and redundancy of features through MI. Figure 2 illustrates this feature selection process.

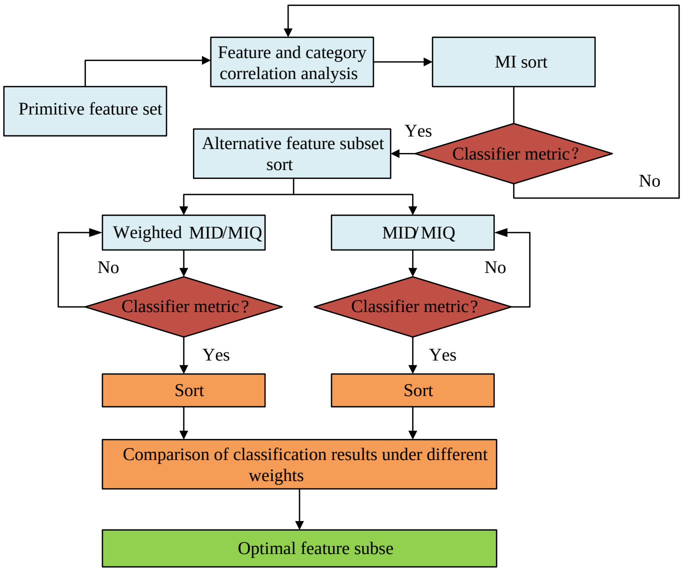


Figure 2: Block diagram of feature selection algorithm

In TFS, the research is based on the preliminary feature word set T_1 screened by the IG algorithm, which contains n features. After performing IG filtering, there is still redundancy among the feature words in the subset. Therefore, it is necessary to perform secondary feature extraction on the selected subset. The task at this stage is to apply the MRMR criterion to n feature words and select a more optimized feature subset S from T_1. This process is based on maximum correlation D and minimum redundancy R, as calculated in equation (2).

$$\begin{cases} \max D(S,C), D = \dfrac{1}{|S|}\sum_{t_i \in S} I(t_i;C) \\ \min R(S), R = \dfrac{1}{|S|^2}\sum_{t_i,t_j \in S} I(t_i;t_j) \end{cases} \quad (2)$$

In equation (2), $\max D$ and $\max R$ represent the maximum relevance and minimum redundancy, $|S|$ represents the amount of selected feature words, $I(t_i;C)$ represents the amount of MI between feature words t and text classification C, and $I(t_i;t_j)$ represents the MI between feature words t_i and t_j. These two criteria are combined to calculate MRMR value, as shown in equation (3).

$$\max \phi(D,R), \phi = D - R \quad (3)$$

In equation (3), D is correlation and R is redundancy. When processing text data, due to the large number of feature words, it is often time-consuming to calculate the MI between them. The MRMR strategy takes a step-by-step iterative approach to identify the ideal combination of features S. If it has already selected $k-1$ features to form a subset S_{k-1}, the next task is to extract the next feature from the pool of features $\{T_1 - S_{k-1}\}$ that have not yet been selected. The rules followed in the selection process are described in equation (4).

$$\max_{t_i \in S_{k-1}}[I(t_i;C) - \frac{1}{k-1}\sum_{t_j \in S_{k-1}} I(t_i;t_j)] \quad (4)$$

To optimize the selection of feature subsets, an improved MRMR TFS is further proposed, which mainly increases the weight of the relationship between features and categories. By introducing the class difference degree a, the improved algorithm can more accurately evaluate the distribution and influence of features in different categories. It combines inter-class dispersion AC and coupling degree DC to measure the distribution of feature words in different categories of documents and the uniformity within the same category of documents, respectively. The representation of features can be enhanced to increase their prominence in a particular category and ensure even distribution across documents within a class. equation (5) shows the relevant calculations.

$$\begin{cases} AC = \sqrt{\dfrac{1}{m-1}\left(\sum_{k-1}^{m}(f_k(t_i) - \overline{f(t_i)})^2\right)} \\ DC_k = \sqrt{\dfrac{1}{n}\sum_{p}^{n}(g_p(t_i) - \overline{g(t_i)})^2)} \end{cases} \quad (5)$$

In equation (5), n and m denote the total number, $f_k(t_i)$ and $f(t_i)$ are the number of documents and the average number of documents for the feature words. If the dispersion AC value is higher, the feature words t_i are more effective in distinguishing categories. $g_p(t_i)$ is the word frequency of the p document, $g(t_i)$ is the average word frequency of the feature word across all documents in the class C_k. A lower value for intra-class coupling DC indicates that it is more efficient on behalf of the class C. Next, the MRMR algorithm considers the MI of feature words in all categories, fine-tunes the weight of the MI by introducing the class difference degree β, and selects the two largest class difference degree values for processing, as detailed in equation (6).

$$a = \frac{1}{\lambda}\log_2(\beta_{\max 1} - \beta_{\max 2}) = \frac{1}{\lambda}\log_2\left(\frac{AC}{DC_{\min 1}} - \frac{AC}{DC_{\min 2}}\right) \quad (6)$$

In equation (6), λ is a constant, a represents the difference in the degree of difference of the class, and this difference is logarithmic. This calculation method is applied to the MRMR algorithm, as shown in equation (7).

$$\max_{t_i \in T_1 - S_{k-1}} [aI(t_i;C) - \frac{1}{k-1}\sum_{t_j \in S_{k-1}} I(t_i;t_j)] \quad (7)$$

A significant difference indicates that the feature words are primarily present in one category, making them highly identifiable to that category. Conversely, a small difference suggests that the feature words are common across multiple categories and are not enough to distinguish between categories with certainty. Logarithmic processing helps maintain data characteristics and the relationship between features and categories, while reducing data size and ensuring stability. In summary, the MRMR algorithm steps are shown in Figure 3.

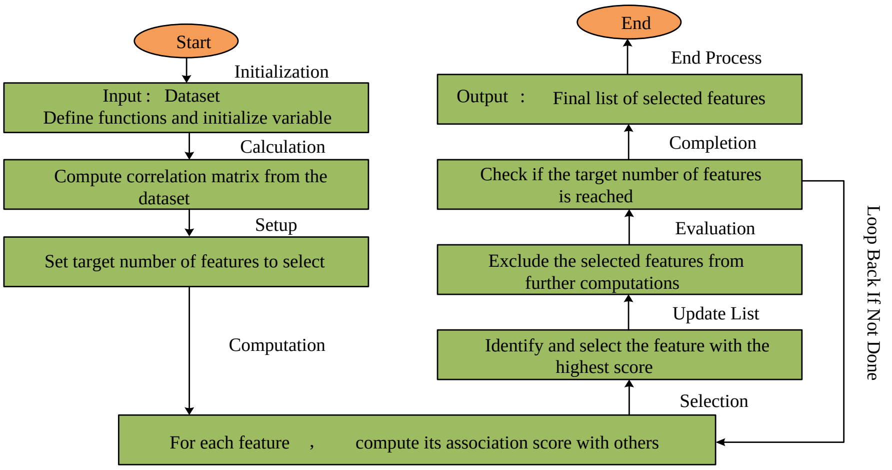

Start
Initialization
Input : Dataset
Define functions and initialize variable
Calculation
Compute correlation matrix from the dataset
Setup
Set target number of features to select
Computation
For each feature , compute its association score with others
Selection
Identify and select the feature with the highest score
Update List
Exclude the selected features from further computations
Evaluation
Check if the target number of features is reached
Loop Back If Not Done
Completion
Output : Final list of selected features
End Process
End

Figure 3: MRMR algorithm steps

2.2 Application of fourier mixed kernel function in SVM text classification algorithm

To enhance SVM's performance in text classification, SVM text data classification algorithm with Fourier hybrid kernel function is further introduced. In the text classification task, features are usually feature words or n-grams, forming a large number of text vectors. The SVM algorithm maps the input vectors to a higher-dimensional space, identifies a hyperplane that separates the data, and maximizes the margin between the hyperplane and the data points to enhance the classification accuracy. Linear SVMs includes linearly separable and indivisible, linear separable means that the data can be directly sliced by the hyperplane. Binary classification data on a 2D plane, if a line can divide the two classes, the line is a hyperplane. To simplify the calculation, the data labels on both sides are set to $y = +1$ and $y = -1$ as shown in Figure 4.

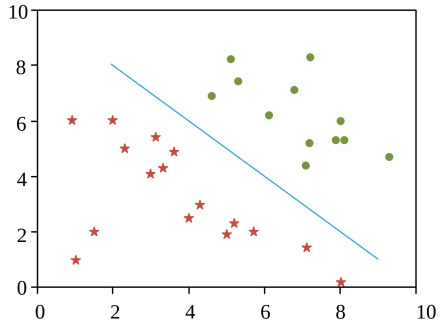
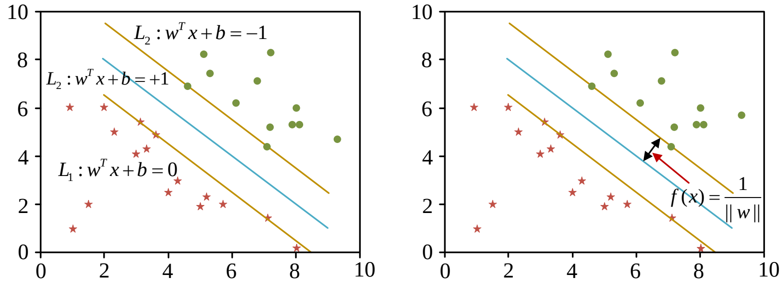


Figure 4: Schematic diagram of linear classification structure

In solving a linear separability problem, the SVM defines a hyperplane by determining a function $f(x) = w^T x + b = 0$. The two sides of this hyperplane $y = -1$ and $y = +1$ can be represented as $w^T x + b = -1$ and $w^T x + b = +1$, respectively, The SVM aims to find an optimal segmentation surface that maximizes the classification interval, i.e., the distance between the two sides L_1 and L_2 of the hyperplane, as shown in Fig. 5(a). While there are multiple possible segments, only one can segment the data perfectly. The optimal segmentation surface is represented as a hyperplane L_1, and the points on both sides of L_1 and L_2 are called sample points, which are also the key to SVM calculations, i.e., support vectors. The distance of these support vectors to the hyperplane L_1 determines the interval of the classification, as shown in Figure 5.



(a) Linear classification representation in SVM

(b) Hyperplane spacing in SVM

Figure 5: Linear classification and hyperplane interval description of SVM

In Figure. 5, the distances between the support vectors are equal to $\frac{2}{\|w\|}$. The goal of setting the training sample set D is to find a partition hyperplane with a maximum interval, which requires determining the parameters w and b that satisfy a particular constraint to maximize $\frac{2}{\|w\|}$. In fact, maximization $\frac{2}{\|w\|}$ is equivalent to minimization $\|w\|^2$, so the original problem becomes a minimized $\|w\|^2$ problem, as detailed in equation (8).

$$\begin{cases} \max\limits_{w,b} \frac{2}{\|w\|} s.t. y_i(w^T x_i + b) \geq 1, i = 1, 2, ..., n \\ \max\limits_{w,b} \frac{1}{2} \|w\|^2 \; s.t. y_i(w^T x_i + b) \geq 1, i = 1, 2, ..., n \end{cases} \tag{8}$$

Equation (8) is a convex quadratic programming problem with constraints. Considering their characteristics, in order to simplify the calculation, Lagrangian multiplier is applied to transform it into a dual problem. By setting the L partial derivative relative to w and b to zero, the calculation process can be transformed to obtain the expression of w and b. Substituting these into $L(w,b,a)$, equation (9) can be obtained to construct a classification model.

$$\begin{cases} L(w,b,a) = \frac{1}{2}\|w\|^2 - \sum\limits_{i=1}^{n} a_i(1 - y_i(w^T x_i + b)) \\ \max\limits_{a} \sum\limits_{i=1}^{n} a_i - \frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{i=1}^{n} a_i a_j y_i y_j x_i^T x_j \\ f(x) = sign(w^T x + b) = sign(\sum\limits_{i=1}^{n} a_i y_i x_i^T + b) \end{cases} \tag{9}$$

In reality, most data is non-linear and cannot be directly classified by linear methods. SVM solves it by mapping data to a high-dimensional space. The kernel function is used for inner product operations, which avoids complication and dimensional disaster. The kernel function must meet the Mercer condition. SVMs with kernel functions can also be solved using the Lagrangian multiplier method, as shown in equation (10).

$$\begin{cases} \max\limits_{a} \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n} a_i a_j y_i y_j K(x_i, x_j) \\ s.t. \sum_{i=1}^{n} a_i y_i = 0, 0 \le a_i \le C, i = 1, 2, ..., n \end{cases} \quad (10)$$

In equation (10), $K(x_i, x_j)$ is the kernel function, and the final classification model is shown in equation (11).

$$f(x) = sign(\sum_{i=1}^{n} a_i y_i K(x_i, x) + b) \quad (11)$$

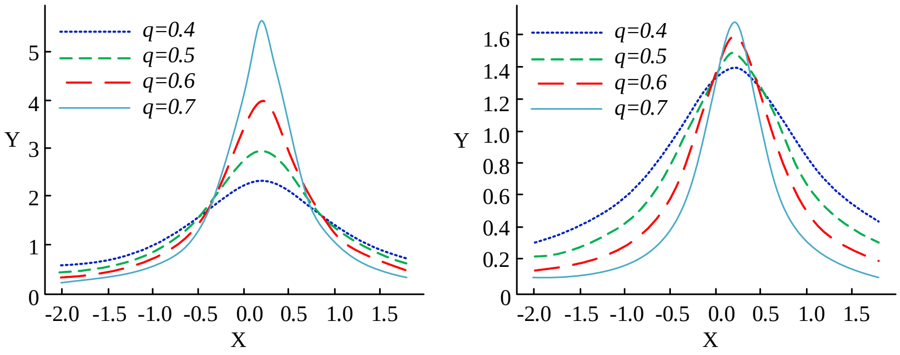
Next, the Fourier kernel function is proposed. In practical use, in addition to the universal Gaussian kernel and polynomial kernel, this function performs well in specific fields and has a high learning effect. There are two main forms of manifestation, and the one-dimensional Fourier kernel function corresponding to the two types is detailed in equation (12).

$$\begin{cases} K(x_i, x_j) = \dfrac{1 = q^2}{2(1 - 2q\cos(x_i, x_j) + q^2)} \\ K(x_i, x_j) = \dfrac{\pi}{2\gamma} \dfrac{ch(\dfrac{\pi - |x_i - x_j|}{\gamma})}{sh(\dfrac{\pi}{\gamma})} \end{cases} \quad (12)$$

In equation (12), q is $\in (0,1)$. The n dimensional expressions for the two kernel functions are defined in Eq. (13).

$$K(x_i, x_j) = \prod_{t=1}^{m} K([x_i]_t, [x_j]_t) \quad (13)$$

As above, the corresponding one-dimensional and n dimensional Fourier kernel functions are shown in Figure 6.



(a) One-dimensional Fourier kernel function

(b) N-dimensional Fourier kernel function

Figure 6: Fourier kernel function graph

As a local kernel, the Fourier kernel function is characterized by adjusting its amplitude only by parameter q, which provides an effective learning mechanism for text classification. The Fourier nucleus provides buffer attenuation near the test point, which improves the sparse distribution in high-dimensional spaces. However, the right q value selection is critical, as inappropriate q value can lead to too rapid attenuation near the test point. In order to optimize the performance, the principle of linear weighted combination of kernel functions is adopted. This method combines the different kernel functions and aims to improve the accuracy and efficiency of text classification. The specific combination and parameter adjustment are shown in equation (14).

$$K_{mix} = aK_1 + (1 - a)K_2, 0 \le a \le 1 \quad (14)$$

In equation (14), K_{mix} represents the hybrid kernel function, which combines the respective characteristics of the two single-kernels K_1 and K_2 that satisfy the Mercer condition, and a denotes the influence of these two single-kernels. In order to construct a hybrid kernel with better performance, it is proposed to combine the polynomial kernel (as the global kernel) and the Fourier kernel (as the local kernel) to integrate the advantages of the two. At the same time, the combination of polynomial kernels and widely used Gaussian kernels is also considered to compare the classification effects of the two hybrid kernels, as shown in equation (15).

$$\begin{cases} K(x_i, x_j) = a \times (x_i \cdot x_j + c)^d + (1 - a) \times \dfrac{(1 - q^2)}{2(1 - 2q\cos(x_i, x_j) + q^2)} \\ K(x_i, x_j) = a \times (x_i \cdot x_j + c)^d + (1 - a) \times \exp(-\gamma \| x_i - x_j \|^2) \end{cases} \quad (15)$$

In equation (15), $a (0 \le a \le 1)$ is the weight coefficient, which balances the combined effect of the two kernel functions. Fourier nuclei are prioritized for their easy parameter adjustment q and buffer attenuation away from the test point. Based on the principle of combinatorial kernels, the proposed Fourier hybrid kernel function combines the linear weighting of the Fourier kernel and the polynomial kernel, which conforms to Mercer's theorem and is suitable for the kernel function of SVMs. Overall, the process of improving the SVM algorithm is shown in Figure 7.

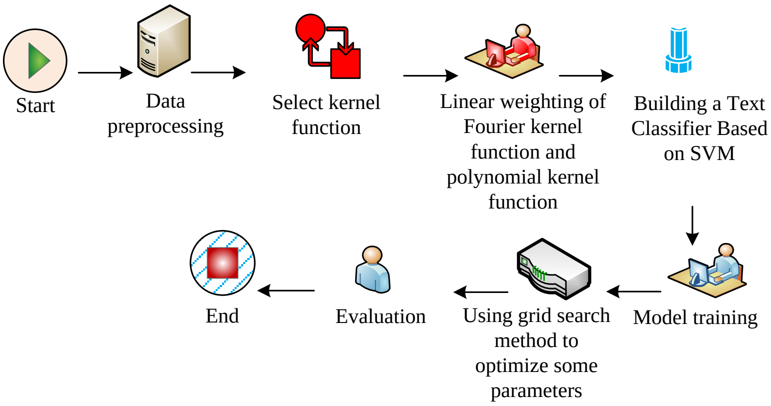

Start
Data preprocessing
Select kernel function
Linear weighting of Fourier kernel function and polynomial kernel function
Building a Text Classifier Based on SVM
Model training
Using grid search method to optimize some parameters
Evaluation
End

Figure 7: Improve the process of SVM algorithm

Figure 7 shows the preprocessed data being input into the SVM algorithm, followed by the selection of the kernel function. The selected Fourier and polynomial kernel functions are linearly weighted to construct a text classification model. The model is trained using a partitioned training dataset and parameter selection is done using the grid search method. Finally, the model is evaluated using the test set.

3 Text classification results analysis based on two-stage feature selection and improved machine learning

In this study, three datasets and their parameter configurations are first identified. Subsequently, feature selection and classification results are analyzed for these different datasets. Finally, a variety of kernel functions are analyzed in depth, and the proposed algorithm evaluates SVM classification performance of these kernel functions in detail.

3.1 Results analysis of IG-MRMR two-stage feature selection algorithm under different datasets

Experiments are conducted using the LING-SPAM, IMDB, and Cornell datasets. The text data is pre-processed by filtering out noisy feature items, reducing feature dimensions, alleviating classifier burden, and improving text classification accuracy through the removal of stop words, punctuation, and special characters. 70% of data are the training set and 30% the test set, the classifier is an SVM model using Gaussian kernels, and the experimental environment is Python. To evaluate the effect of IG-MRMR algorithm in extracting feature subsets, the accuracy and F1 value are used as evaluation indexes. The algorithm's performance improves as the accuracy of its feature selection increases. A higher F1 value indicates better accuracy and recall, resulting in a more effective feature selection. The IMDB dataset is applied to the Chi-Square (CHI), MI and TFS of IG, IG-MRMR and IG-MRMR, and the feature subsets from 10 to 100 dimensions are selected, respectively. The dimension interval for each feature subset is 10. After selecting the first 20-dimensional feature subset of each method, the number of extracted words ranges from 15 to 14, 16, 16, and 18, and the priority order of each feature subset is also not the same. The accuracy results of the algorithm are presented in Fig. 8. The number of feature subsets required to achieve an accuracy of 0.82 for each algorithm is 60, 63, 59, 46, and 40, respectively. This shows that the IG-MRMR two-stage feature algorithm has the best prediction effect while using fewer feature words, and has the highest classification accuracy with the same feature subsets.

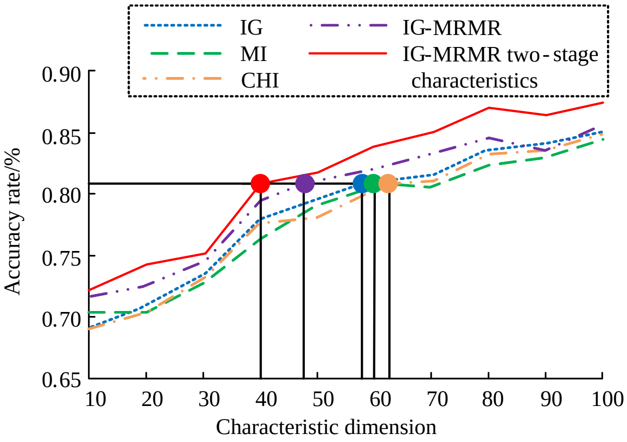
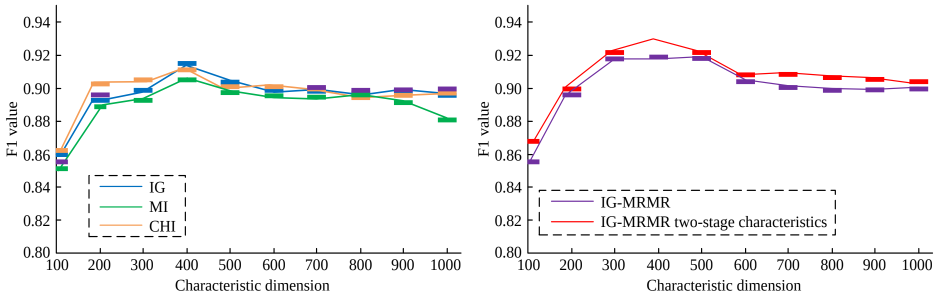


Figure 8: 10 to 100 dimensional results for different algorithms on IMDB datasets

To evaluate the influence of feature dimension improvement on different feature selection algorithms, the experimental set of feature subset dimension range is increased from 100 to 1000, with each 100 as an interval. A comparison of the five methods is shown in Figure 9. The F1 values of all algorithms begin to decrease when the number of features exceeds 390, indicating that the key features have been extracted and the additional features have reduced the classification effect. In Figure 9(b), the IG-MRMR TFS algorithm shows an advantage, with an average F1 value of about 1% to 2% higher than that of other algorithms, which means that more text can be correctly classified, about 18 more articles, showing its efficient and accurate feature selection ability.

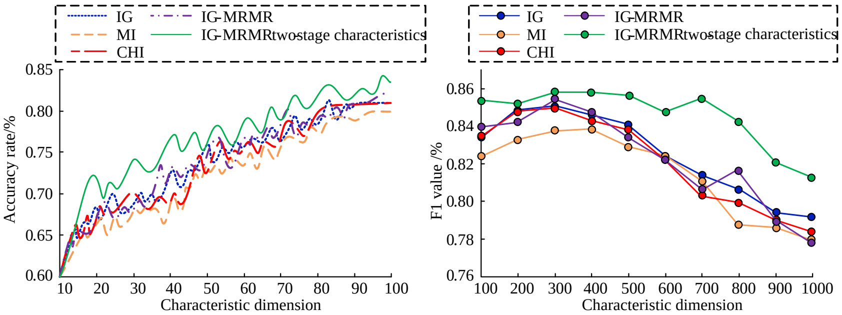


(a) Feature selection algorithm

(b) Improved algorithm

Figure 9: Comparison of F1 values of different algorithms on IMDB datasets

Five different algorithms are applied to the Cornell dataset for experiments, the same as the IMDB dataset, with feature dimensions set between 10 and 100. The analysis focuses on the first 20-dimensional feature subsets extracted by each algorithm. It is found that the number of extracted evaluation words ranges from 15 to 17, as shown in Figure 10(a). To further explore the effect of feature dimension increase on the classification effect, the experimental range is extended to 100 to 1000 dimensions, with 100 intervals, as shown in Figure 10(b).

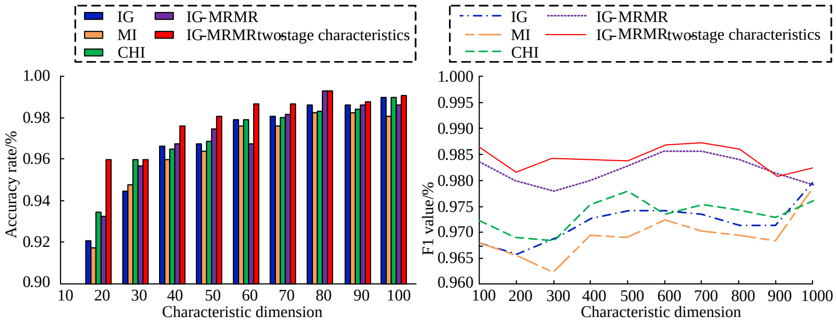


(a)Accuracy changes in the Cornell data set

(b) Changes in F1 values in the Cornell dataset

Figure 10: Different algorithms in Cornell data set

Figure 10(a) shows that at an accuracy of 0.76, the number of feature subsets required for the five algorithms is about 57, 60, 59, 55, and 40, respectively. The IG-MRMR TFS requires the least number of feature subsets, and its accuracy is higher than that of the same number of feature subsets. Figure 10(b) shows that the classification effect is best when the number of features is close to 285. As the number of features increased, the classification effectiveness of all algorithms gradually decreased. This suggests that the additional features contain more words with weak representation abilities. IG-MRMR TFS algorithm only shows a significant decrease after the feature exceeded 700, and its F1 value is 2% higher than that of other methods on average, and the number of correctly classified texts are increased by about 18. Next, experiments of five algorithms are carried out on the LING-SPAM dataset, and the feature words of this dataset mainly focuses on advertising-related words. In this study, 10-dimensional to 100-dimensional feature words are selected for comparison of classification effects, and the detailed results are shown in Figure 11(a). In order to have a more comprehensive understanding of the classification performance of feature subsets, the feature dimension is further extended to 100 to 1000, and the classification results of five feature selection algorithms are compared in Figure 11(b).



(a)Accuracy changes in the LING - SPAM data set (b) Changes in F1 values in the LING - SPAM dataset

Figure 11: Different algorithms in LING-SPAM data set

Figure 11(a) shows the number of feature subsets required to achieve an accuracy of 0.95 for the five algorithms, which are 39, 40, 29, 35, and 22. The IG-MRMR TFS algorithm requires significantly less feature subsets than other methods while maintaining accuracy. At the same time, in the same number of feature subsets, the accuracy of IG-MRMR TFS is generally higher than that of other feature selection algorithms. Figure 11(b) shows that most of the algorithms have reached 0.96 for 100-dimensional features, which means that the words with strong representational ability in the dataset are mainly concentrated in the first 100 dimensions. The F1 value of IG-MRMR TFS peaks when the feature dimension is about 680, and its average accuracy is 1% higher than that of IG-MRMR, and the classification of about 6 articles is correctly increased, which is 2% higher than that of the IG and CHI single-stage algorithms, and about 14 articles are correctly added, showing its accurate feature selection advantage.

3.2 Text classification results analysis based on two-stage feature selection algorithm

To ensure data standardization, a preprocessing is performed to remove stop words, punctuation, and special characters, and the processed corpus words are vectorized using the term frequency-inverse document frequency method. And the weight of the words in the text is calculated and normalized. 60% of dataset is training set and 40% is test set. Parameter selection includes the use of a grid search method to determine the penalty parameters C in the SVM (ranging from 1 to 100, adjusted every 10) and the exponent d of the polynomial kernel, set to 3. The kernel weight range a of the hybrid kernel function is set to 0.1 and the step size is 0.1. The experimental platform uses Python 3.6. A 5-fold cross-validation is adopted, and F1 is the evaluation index. IMDB dataset is selected to compare the performance of the proposed algorithm with other kernel functions. The dataset comprises 2000 reviews of films and television programs, with an equal number of positive and negative reviews. The document frequency algorithm is used as the feature selection algorithm to process the dataset. Considering the excellent performance of the Fourier a kernel function, the weight coefficient in the hybrid kernel function is set to 0.25, and the results are shown in Figure 12.

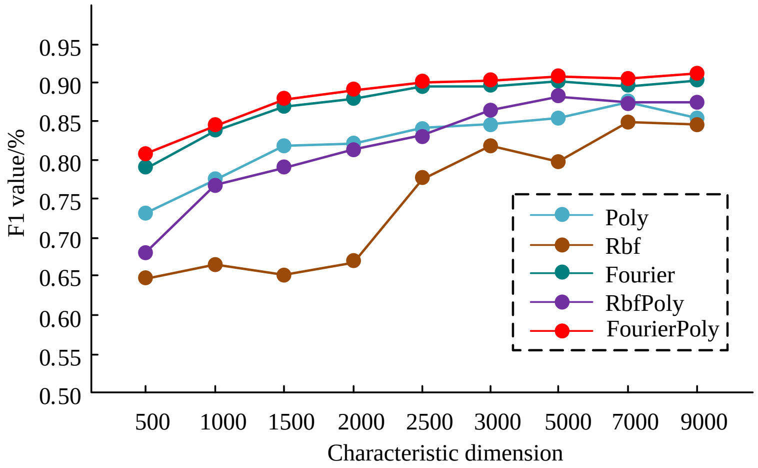
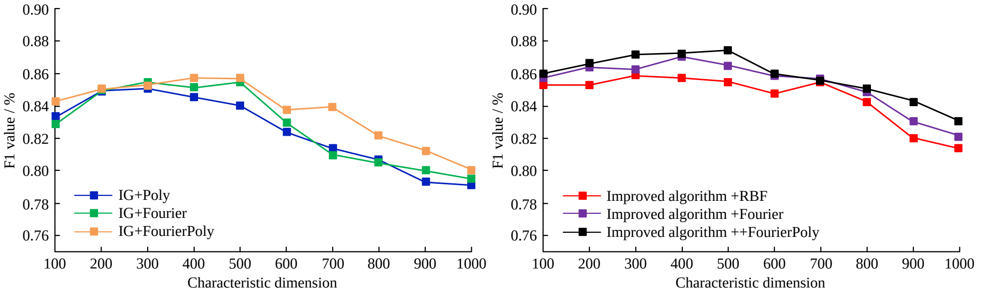


Figure 12: Comparison of multiple kernel functions

In Figure 12, the classification effect is improved with the expansion of the feature dimension. The proposed Fourier hybrid kernel function surpasses the combination of single kernel and Gaussian kernel and polynomial kernel in terms of performance, which verifies the effectiveness of the concept of combinatorial kernel function, highlights the advantages of Fourier kernel function, and provides important value for improving the effect of text classification. In this study, the IG-MRMR TFS algorithm will be used to analyze the SVM classification performance on the Cornell dataset, as shown in Figure 13. IG and IG-MRMR TFS algorithms are used to select features, and the SVMs of Gaussian kernel, Fourier kernel and Fourier hybrid kernel functions are compared with these two feature selection methods.



(a) Classification function

(b) Improved algorithm combined with classification function

Figure 13: Multiple kernel function analysis of IG-MRMR two-stage feature selection algorithm

As can be observed in Figure 13, the classification effect first increases and then decreases. When the feature dimension is about 400, the IG-MRMR two-stage algorithm shows excellent classification performance. As features increase, the effect of TFS decreases significantly, which shows that the increase of weaker feature words in the selected feature subset interferes with the classification effect. Figure 13(a) and 13(b) show that SVM using IG-MRMR TFS using any kernel function is generally better than IG method in terms of F1 value compared to the IG method, confirming the effectiveness of IG-MRMR. The combination of Fourier hybrid kernel function and IG-MRMR two-stage algorithm is 1~3% higher than other combinations on average in F1 value, and the number of correctly classified texts increases by 20 to 45. The experimental corpus selected for analysis is the Cornell Film and Television Review. The comparative method chosen to analyze its classification effect is the SVM algorithm, as shown in Table 2.

Table 2: Comparison of classification effects

Method	Accuracy/%	F1 value
SVM	73.46	0.617
Research method	96.57	0.813

Table 2 shows that the research method has higher accuracy and larger F1 values (P<0.05) compared to the benchmark method. Specifically, the accuracy of the research method is 96.57%, which is 23.11% higher than the SVM algorithm. These results demonstrate the high performance of the research method, which is further improved through optimization.

3.3 Discussion

In text data feature classification, achieving higher accuracy in text feature selection involves considering data redundancy, correlation between features, and semantic relationships in context. Fan Y et al. conducted research on relevant selection algorithms based on label correlation and feature redundancy to improve the effectiveness of text feature selection. The results indicated that the proposed method has a relatively high selection accuracy [20]. The literature acknowledges the issue of data redundancy and correlation, but there are still shortcomings, such as a lack of research on contextual semantic relationships. This necessitates further optimization of feature selection. However, this study can address these gaps. During the feature selection process, Zhou H et al. analyzed the weight of MI redundancy terms through correlation coefficients and selected the principle of minimization. The proposed method was found to have good feature classification performance in experiments [21]. This reference is comparable with the proposed method. However, there has been no research conducted on the contextual semantic relationships involved in the feature selection process. This study explores this aspect, resulting in a more effective feature selection process. The accuracy and F1 value of feature selection are both high.

4 Conclusion

To enhance text classification redundancy and SVM performance, a TFS algorithm based on IG and improved MRMR is proposed. Additionally, to further improve the effect of SVM in text classification, an SVM text classification algorithm based on Fourier mixed kernel function is introduced. The study found that the IG-MRMR TFS algorithm had the best prediction accuracy with fewer feature words used on the LING-SPAT, IMDB, and Cornell datasets. The algorithm achieved the highest classification accuracy with the same feature subsets. On the IMDB dataset, the algorithm required only 40 feature subsets to achieve an accuracy of 0.82, which was fewer than other algorithms. On the LING-SPAM dataset, the single-stage algorithms IG and CHI were outperformed by 2%. The addition of about 14 articles was correctly classified. Furthermore, when the number of features exceeded 390, the F1 value of all algorithms began to decrease, indicating that the key features had been extracted and additional features were reducing the classification effect. In this case, the IG-MRMR algorithm maintained its advantage, with an average F1 value 1% to 2% higher than other algorithms, and correctly classified 18 more texts. In comparison to benchmark methods, research methods exhibit higher accuracy rates. Specifically, the research method boasts an accuracy rate of 96.57%, which is 23.11% higher than that of the SVM algorithm. However, the study has some shortcomings. The second-stage feature selection of the current IG algorithm may need improvement, and the IG algorithm can be further optimized in the future. Additionally, while the Fourier kernel function shows superiority, future studies can consider more efficient local kernel functions to enhance classification performance. In addition, when dealing with complex real-world problems, such as uneven data distribution, research methods may have limited generalization ability and certain shortcomings. Future work can focus on optimizing the algorithm through feature learning and multi-level feature learning to improve its performance.

Fundings

The research is supported by: Graduate Education Reform Project of Henan Province, Achievements of the Henan Province Higher Education Teaching Reform Research and Practice Project (Graduate Education), (No. 2023SJGLX365Y).

Reference

[1] T. Tuncer, S. Dogan, M. Baygin, and U. R. Acharya, “Tetromino pattern based accurate EEG emotion classification model,” Artificial Intelligence in Medicine, vol. 123, no. 4, pp. 102210-102211, 2022. https://doi.org/10.1016/j.artmed.2021.102210

[2] E. H. Houssein, D. S. Abdelminaam, and H. N. Hassan, “A hybrid barnacles mating optimizer algorithm with support vector machines for gene selection of microarray cancer classification,” IEEE Access, vol. 9, no. 1, pp. 64895-64905, 2021. https://doi.org/10.1109/ACCESS.2021.3075942

[3] C. Dang, Y. Liu, and H. Yue, “Autumn crop yield prediction using data-driven approaches: -support vector machines, random forest, and deep neural network methods,” Canadian Journal of Remote Sensing, vol. 47, no. 2, pp. 162-181, 2021. https://doi.org/10.1080/07038992.2020.1833186

[4] W. Al-Salman, Y. Li, and P. Wen, “Detection of k-complexes in EEG signals using a multi-domain feature extraction coupled with a least square support vector machine classifier,” Neuroscience Research, vol. 172, no. 2, pp. 26-40, 2021. https://doi.org/10.1016/j.neures.2021.03.012

[5] A. Siddiqa, R. Islam, and M. I. Afjal, “Spectral segmentation-based dimension reduction for hyperspectral image classification,” Journal of Spatial Science, vol. 68, no. 4, pp. 543-562, 2023. https://doi.org/10.1080/14498596.2022.2074902

[6] C. Hebbi, and H. Mamatha, “Comprehensive dataset building and recognition of isolated handwritten kannada characters using machine learning models,” Artificial Intelligence and Applications, vol. 1, no. 3, pp. 179-190, 2023. https://doi.org/10.47852/bonviewAIA3202624

[7] Y. A. Ahmed, S. Huda, and B. A. S. Al-rimy, “A weighted minimum redundancy maximum relevance technique for ransomware early detection in industrial IoT,” Sustainability, vol. 14, no. 3, pp.

1231-1235, 2022. https://doi.org/10.3390/su14031231

[8] A. Jiménez-Cordero, J. M. Morales, and S. Pineda, “A novel embedded min-max approach for feature selection in nonlinear support vector machine classification,” European Journal of Operational Research, vol. 293, no. 1, pp. 24-35, 2021. https://doi.org/10.1016/j.ejor.2020.12.009

[9] B. Wang, X. Zhang, C. Sun, and X Chen, “Sparse representation theory for support vector machine kernel function selection and its application in high-speed bearing fault diagnosis,” ISA transactions, vol. 118, no. 1, pp. 207-218, 2021. https://doi.org/10.1016/j.isatra.2021.01.060

[10] L. Sun, T. Yin, W. Ding, Y. Qian, and J. Xu, “Feature selection with missing labels using multilabel fuzzy neighborhood rough sets and maximum relevance minimum redundancy,” IEEE Transactions on Fuzzy Systems, vol. 30, no. 5, pp. 1197-1211, 2021. https://doi.org/10.1109/TFUZZ.2021.3053844

[11] H. Jia, and K. Sun, “Improved barnacles mating optimizer algorithm for feature selection and support vector machine optimization,” Pattern Analysis and Applications, vol. 24, no. 3, pp. 1249-1274, 2021. https://doi.org/10.1007/s10044-021-00985-x

[12] Z. Yin, J. Zheng, L. Huang, Y. Gao, H. Peng, and L. Liu, “SA-SVM-based locomotion pattern recognition for exoskeleton robot,” Applied Sciences, vol. 11, no. 12, pp. 5573-5575, 2021. https://doi.org/10.3390/app11125573

[13] S. R. Bansal, S. Wadhawan, and R. Goel, “mrmr-pso: A hybrid feature selection technique with a multiobjective approach for sign language recognition,” Arabian Journal for Science and Engineering, vol. 47, no. 8, pp. 10365-10380, 2022. https://doi.org/10.1007/s13369-021-06456-z

[14] H. Zhou, X. Wang, and R. Zhu, “Feature selection based on mutual information with correlation coefficient,” Applied Intelligence, vol. 1, no. 1, pp. 1-18, 2022. https://doi.org/10.1007/s10489-021-02524-x

[15] M. Yildirim, A. Çinar, and E. Cengil, “Classification of the weather images with the proposed hybrid model using deep learning, SVM classifier, and mRMR feature selection methods,” Geocarto International, vol. 37, no. 9, pp. 2735-2745, 2022. https://doi.org/10.1080/10106049.2022.2034989

[16] D. Wang, and G. Xu, “Research on the detection of network intrusion prevention with svm based optimization algorithm,” Informatica, vol. 44, no. 2, pp. 269-273, 2020. https://doi.org/10.31449/inf.v44i2.3195

[17] M. G. Lanjewar, J. S. Parab, and A. Y. Shaikh, “CNN with machine learning approaches using ExtraTreesClassifier and MRMR feature selection techniques to detect liver diseases on cloud,” Cluster Computing, vol. 26, no. 6, pp. 3657-3672, 2023. https://doi.org/10.1007/s10586-022-03752-7

[18] H. Azadi, M. R. Akbarzadeh-T, H. R. Kobravi, and A. Shoeibi, “Robust voice feature selection using interval type-2 fuzzy AHP for automated diagnosis of parkinson's disease,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, no. 3, pp. 2792-2802, 2021. https://doi.org/10.1109/TASLP.2021.3097215

[19] M. Mahapatra, S. K. Majhi, and S. K. Dhal, “Mrmr-ssa: a hybrid approach for optimal feature selection,” Evolutionary Intelligence, vol. 15, no. 3, pp. 2017-2036, 2022. https://doi.org/10.1007/s12065-021-00608-8

[20] Y. Fan, B. Chen, W. Huang, J. Liu, W. Weng, and W. Lan, “Multi-label feature selection based on label correlations and feature redundancy,” Knowledge-Based Systems, vol. 241, no. 6, pp. 1-15, 2022. https://doi.org/10.1016/j.knosys.2022.108256

[21] H. Zhou, X. Wang, and R. Zhu, “Feature selection based on mutual information with correlation coefficient,” Applied Intelligence, vol. 52, no. 5, pp. 5457-5474, 2022. https://doi.org/10.1007/s10489-021-02524-x