Applied Artificial Intelligence and Computing Systems
https://www.scigate.org
Science Gate Academic
Vol. 1, No. 2, 2025

Article

A Data-Driven Intelligent Optimization Framework for Automated Text Classification

Carys Merton
Saint Mary's University, Halifax, Canada
cmerton909@smu.ca

Abstract: *This paper proposes an intelligent optimization framework for text classification based on data-driven machine learning models. To address the limitations of traditional parameter tuning in classification algorithms, a swarm intelligence-inspired optimization approach is employed to enhance model convergence and generalization. The framework integrates word vector representation, feature weighting, and adaptive optimization mechanisms to improve semantic understanding and classification precision. A large-scale Chinese text dataset is used to evaluate the performance of the model through metrics including precision, recall, and F1-score. Experimental results demonstrate that the proposed approach effectively improves parameter search efficiency, enhances classification accuracy across multiple categories, and achieves superior performance compared with conventional optimization and learning methods. The study provides a scalable methodology for intelligent text analysis and contributes to the broader application of data-driven optimization in natural language processing and automated information systems.*

Keywords: *Text classification; data-driven learning; intelligent optimization; swarm intelligence; machine learning; feature representation; natural language processing*

1. Introduction

In the digital age, the exponential increase in the volume of text data is a huge challenge that people face on a daily basis. The analysis of such large amounts of text data has become more important in a variety of scenarios, including social media opinion monitoring, product review analysis, and personalized news recommendations. As a result of this expanding requirement, text mining techniques have arisen that combine technologies such as machine learning, deep learning, statistical analysis, and computational linguistics to extract meaningful information from unstructured text. Among the applications of text mining, text classification stands out as a critical component of language processing with widespread utility in modern society.It is extensively employed in tasks such as spam filtering, legal text categorization, and sentiment analysis of user reviews[1].

In the research of Chinese text classification algorithm and model, various classification models have been utilized by domestic scholars. Ma et al. used the Naive Bayes model to address Chinese comment classification[2]; Ding et al. opted for the improved K-nearest neighbor algorithm (KNN) to explore short

text classification[3]; He utilized the maximum entropy model for text classification[4];Yuan delved into the application of the decision tree classifier in Chinese text classification[5]; Wang et al. implemented the support vector machine (SVM) for the text classification of police patrol information[6]. Numerous experiments have showcased the robust generalization ability of SVM in Chinese text classification. However, it is noteworthy that the model's sensitivity to parameter selection is a crucial aspect to consider. The size of the penalty factor and kernel function parameters directly influences the performance of the model classification[7]. Therefore, the optimization of model parameters is essential to achieve high classification performance for SVM.

Traditional parameter optimization methods include experimental method and grid search method, the former is too time-consuming and not easy to find the optimal solution, and the latter is too complicated to compute. Therefore, scholars have introduced the swarm intelligence optimization algorithm to find the optimal parameters of the model. Swarm intelligence optimization algorithm is a heuristic search strategy based on the natural process of biological evolution or group activities, which is carried out around a population. It shows high flexibility and robustness in solving complex problems, so it is widely used in optimization problems[8]. In recent years, more new swarm intelligence optimization algorithms have been proposed: Fruit Fly Optimization Algorithm (FOA), Brain Storm Optimization algorithm (BSO), Zebra Optimization Algorithm (ZOA) etc.The Water Wave Optimization (WWO) algorithm studied in this paper is a swarm intelligent optimization algorithm ,which was proposed by Chinese scholar Yujun Zheng in 2015[9]. By simulating water wave's motion,the algorithm balances the global search and local search and shows good parameter optimization ability.

In this paper, WWO algorithm based on chaos theory is proposed to optimize SVM parameters, and weighted Word2vec algorithm is used to realize text vectorization. Finally, it is applied to the research of Chinese text classification.

2. Relevant theories

2.1 Water Wave Optimization algorithm

The Water Wave Optimization(WWO) algorithm is a new heuristic algorithm,which is based on the shallow water wave theory, and simulates the natural behavior of water waves during flow in coastal waters.The algorithm is divided into the following three operations: propagation, refraction and breaking. The propagation operation makes the high fitness water waves search locally, and the low fitness water waves expand the search area;the refraction operation increases population diversity by updating water waves, and also avoids premature convergence caused by stagnation in the search process;the breaking operation divides the water waves that has the best fitness into more little independent water waves, so that further precise search can be made in the area of potential optimal solutions. Combined with the above three operations, the algorithm achieves a perfect balance between global search and local search.

The solution space of the problem in the WWO algorithm is like the seabed, a solution is analogous to a “water wave”, and the fitness value of each solution is analogous to the distance between the wave and the sea level[10]. The fitness value of the water wave is inversely proportional to its distance to the sea level: the lower fitness means the lower energy of the water wave, it also has the smaller wave height, the larger wavelength, and the closer the distance to the sea level, as shown in Figure 1.

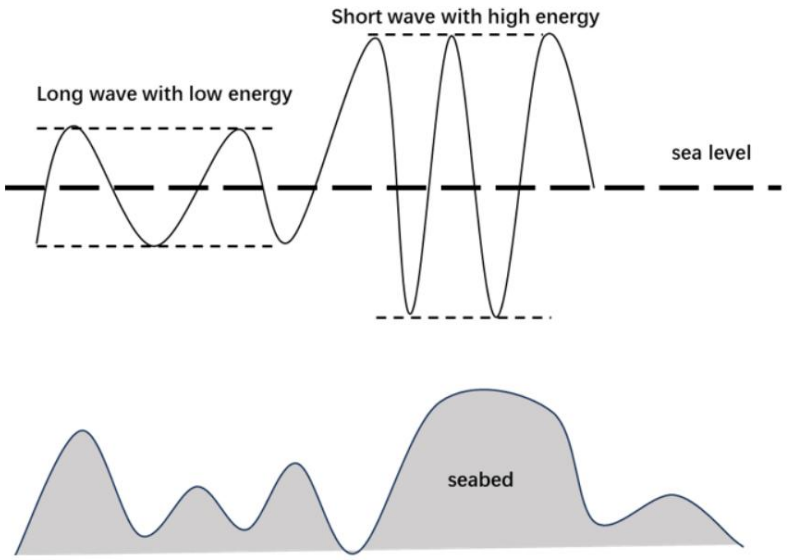

Figure 1. Schematic diagram of water wave optimization algorithm model.

Starting from solving the maximum problem, set the problem dimension to *D* ,the objective function to *f* ,the position of the water waves before propagation be *X* ,the wave height of each wave be *h* ,the wavelength be λ。

Propagation

At each iteration, each wave needs to be propagated once to generate a new one, which is calculated as:

$$x'(d) = x(d) + rand(-1,1)\cdot\lambda L(d)$$

where,$rand(-1,1)$ is a random number that follows a uniform distribution in $[-1,1]L(d)$. is the length of the dimensional search space. If the new position is out of the search range, it will be replaced by a random position within the valid range.

After updating the position of the wave, calculate the fitness $f(x')$ for the new wave. If $>f(x)$,update the position to and reset its wave height to X' . Otherwise, to retain the original wave position *x*, but the wave height h minus 1 due to energy loss during propagation.

After each iteration, the wavelength of each wave should be updated according to the corresponding equation.

$$\lambda = \lambda \cdot \alpha^{-(f(x)-f_{\min}+\varepsilon)/(f_{max}-f_{min}+\varepsilon)}$$

where $f_{\max}$ and $f_{\min}$ are respectively the maximum and minimum fitness values in the current population; α is the wavelength constraint coefficients; and ε is nonzero infinitesimal positive integers (to avoid a denominator of 0).

Refraction

Due to the energy loss of the water wave propagation, when the wave height h decreases to 0, it is necessary to simulate the refraction of the water wave to avoid stopping its propagation. Reset the wave height h to $h_{\max}$, and the position of the wave is updated according to the corresponding equation.

$$x'(d) = Gaussian\left(\frac{x^*(d) + x(d)}{2}, \frac{x^*(d) - x(d)}{2}\right)$$

where $x^*(d)$ is the current optimal wave's position.

$$\lambda' = \lambda \cdot \frac{f(x)}{f(x')}$$

During the propagation process, when the energy of the water wave exceeds the threshold value, its peak will gradually steepen and eventually break into a series of independent water waves. For the current optimal wave, it is necessary to simulate the breaking of the water wave to realize the fine search in its vicinity. At first, randomly select the k dimensions (*k* takes a random number between 1 and the preset parameter $k_{\max}$), then generate independent water

waves X' in each dimension, the position update formula as:

$$x'(d) = x(d) + N(0,1) \cdot \beta L(d)$$

where β is the breaking coefficient of water wave. If the independent water wave is not better than the optimal water wave, the optimal water wave is kept; otherwise, the most suitable water wave among the independent waves is used instead.

2.2 Support Vector Machines

Support Vector Machine (SVM) is a machine learning method developed on the basis of statistical learning theory, which can more effectively solve the classification problems of small samples, nonlinearity and high dimensions[11], and then it is also widely used in the field of text classification. The learning strategy of SVM is to determine the classification hyperplane with the largest boundary: the hyperplane furthest from the edge observation points of the two categories. Generally, there are two cases: 1) When the samples are linearly separable, the maximum boundary hyperplane that can correctly separate the samples is solved in the original space;2) When the samples are nonlinear separable, nonlinear mapping is used to transform them into linear separable problems in high-dimensional space for solution, and kernel function technology is used to overcome dimensional disaster problems. The training of an SVM can be expressed as a convex quadratic programming problem, as shown below:

$$\begin{cases} \min \quad \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\varepsilon_i \\ s.t \quad y_i(\omega^\top x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0, i = 1, 1, \cdots n \end{cases}$$

Where, C is the penalty factor, ξ_i is the slack variable, and b is the classification threshold.

The quadratic programming problem with linear constraints is solved using the Lagrange multiplier method, and the corresponding classification decision function is finally obtained.

$$f(x) = sign\left(\sum_{i=1}^{n}\alpha_i y_i K(x_i, x) + b\right)$$

Where, $\alpha_i C$ is the Lagrange multiplier and $K(x_i,x)\xi_i$ is the kernel function. Common kernel functions include linear kernel, polynomial kernel, perceptron kernel (Sigmoid), and Gaussian kernel (RBF). Among them, the RBF can approximate any nonlinear function and has advantages such as wide convergence domain, few parameters, and strong universality. This paper adopts the RBF kernel function to train the SVM, as described by the corresponding equation.

$$K(x_i,x) = exp^{-\frac{\| x-x_i \|^2}{2\sigma^2}}$$

In the equation, $\frac{1}{2\sigma^2}$ is the width of RBF kernel function.

2.3 Word2Vec

Text representation is the key step in natural language processing , which can be categorized as One-Hot encoding and distributed representation. But, the former has dimension disaster problem and disadvantages such as semantic gap; the latter by training the corpus, each word from the high-dimensional word vector space is distributedly mapped to a low-dimensional space, which solves the disadvantages of missing semantics and sparse data. The models for distributed representation of word learning can be categorized into two types: matrix decomposition-based and neural network-based[12].The matrix decomposition model does not consider the relationship between words, such as latent index analysis (LSA) and Latent Dirichlet Allocation(LDA). The neural network model takes context into account and increases the plenty of semantic content.

As a training word vector tool released by Google in 2013, Word2vec is essentially a simplified neural language model [13]. The ultimate goal is not to achieve accurate prediction, but to obtain the weight matrix in the mapping relationship, that is, the word vector of all words.

As shown in Figure 2, Word2vec includes two models: continuous bag of words model(CBOW) and skip-gram model. The former predicts the head word from the words above and below, and the latter predicts the context word from the head word.

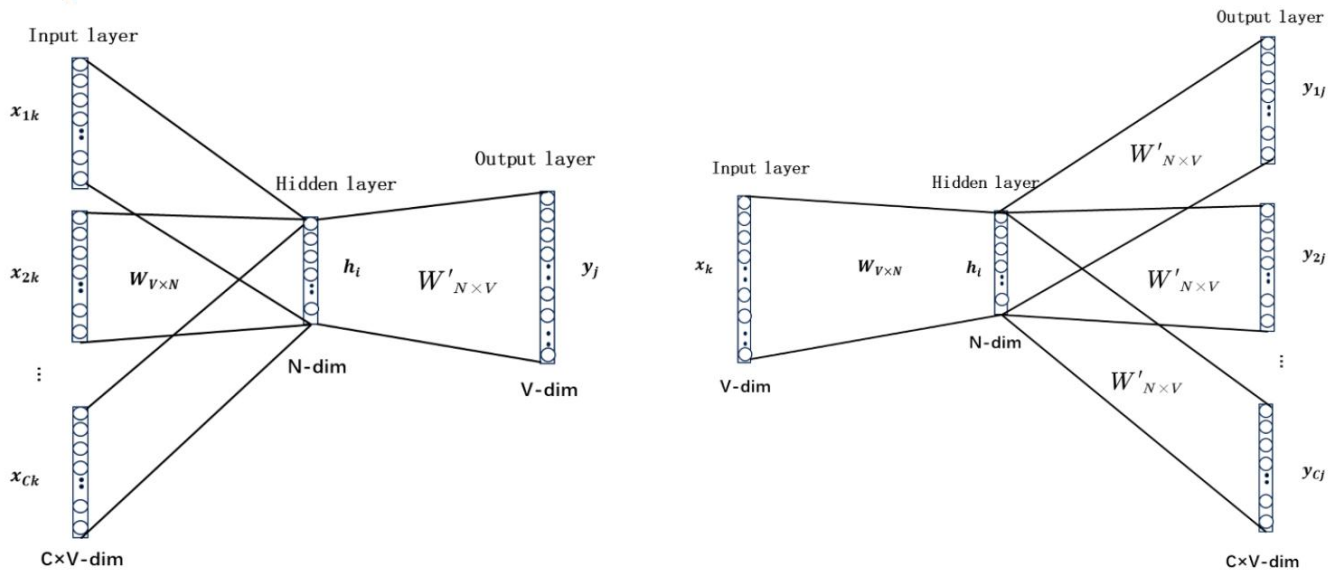

Figure 2. CBOW model and skip-gram model.

2.4 TF-IDF Algorithm

Feature selection is particularly vital in text classification, and the value of feature word weight greatly affects the performance of text classification algorithms. As a common feature weight calculation method, TF-IDF has relatively simple algorithm and better classification effect, so it is widely used in text classification calculation. The definition is shown as:

$$W_{ij} = TF_{ij} \times IDF_j = \frac{n_w}{\sum_{k=1}^{N} n_k} \times \log\left(\frac{N}{df_j + 1}\right)$$

Where TF_{ij} represents the number of occurrences of feature words w_j in the document D_i , and IDF_j represents the probability of inverse text; n_w is the number of times the characteristic word appears in the document, n_k is the number of times the k-th word appears in the document, N is the total number of documents, and df_j is the number of documents in which the characteristic word appears.

3. Improved Algorithm

3.1 WWO-SVM algorithm based on chaos theory

3.1.1 Chaos Initialization

In the WWO algorithm, the random generation of the initial position of the wave leads to the uneven distribution of the solution space, which affects the optimization efficiency of the algorithm. Therefore, this paper proposes an improved algorithm (Chaotic Water Wave Optimization, CWWO) based on chaos theory to initialize the position of wave. Embedded chaotic sequence is a process of obtaining uncertain random motion state from deterministic equation, which has the characteristics of pseudo randomness, regularity and ergodicity[14]. The initial positions generated by chaotic mapping can be evenly distributed in the solution space, and easily jump out of the local optimal solution, which is more conducive to seraching the global optimal solution and improving the convergence speed of the algorithm. Many scholars have proposed various algorithms for generating chaotic sequences: Logistic mapping, Henon mapping, etc. Among them, Cubic chaotic map performs better[15]. Therefore, this paper calculates the sequence based on the corresponding equation and then maps the generated chaotic sequence back to the solution space to obtain the initial position of the water wave.

$$y_{n+1} = \rho y_n (1 - y_n^2)$$

$$x_{id} = L_d + y_{id}(U_d - L_d)$$

Where x_n is a chaotic sequence, $x_n \in (0,1)$ is a control parameter, and cubic chaotic map has good ergodicity when $\rho = 2.595$. U_d and L_d respectively represent the upper and lower limits of the dimension of the solution space, y_{id} is the dimension of the i-th wave, and x_{id} is the coordinate of the i-th wave in the search space.

3.2 Algorithmic flowchart

The specific process of the text classification algorithm based on weighted Word2Vec and CWWO-SVM proposed in this paper is shown in the Figure 3.

The operating system information of this experiment is as follows: the CPU is Intel i5-12500H, memory 16GB, the running environment is Pycharm, and the programming language is Python 3.11.1.

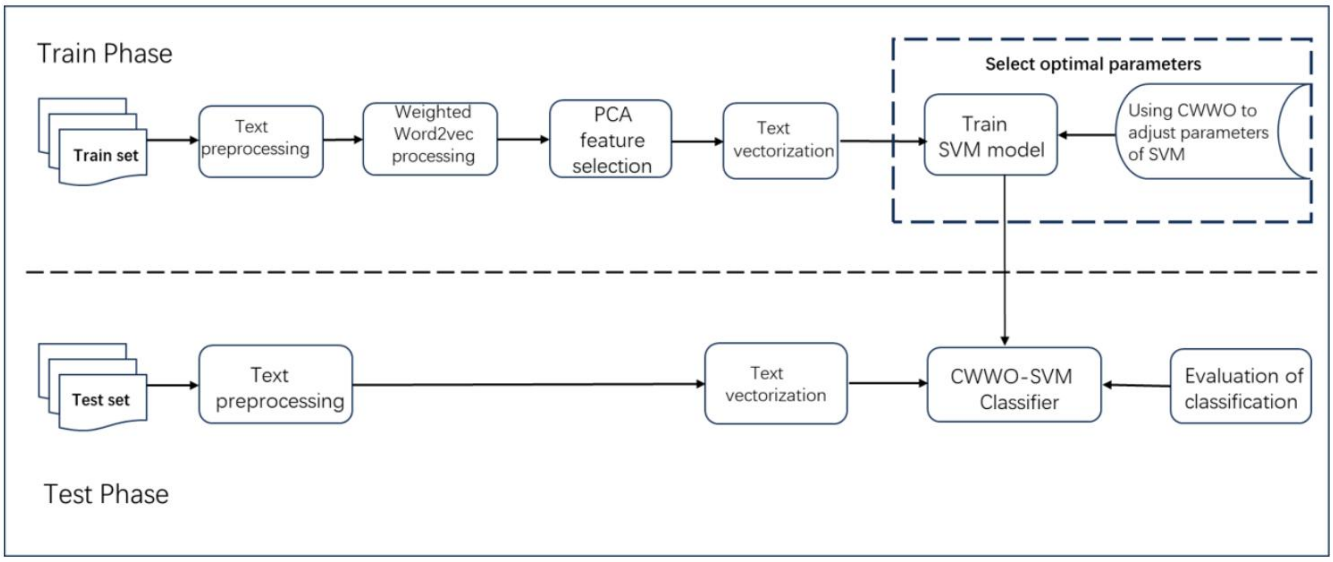

Figure 3. Experimental design process.

4. Experiments and analysis of results

In order to verify the applicability and effectiveness of CWWO-SVM algorithm based on weighted Word2vec in Chinese text classification, this paper selects Fudan Chinese Text Classification Corpus for experiment [18].The corpus is an open-source dataset provided by the Natural Language Processing Laboratory of Fudan University, including Chinese texts from news, forums, etc., covering 20 categories such as economy, agriculture, education, art, etc.

4.1 Data preprocess

(1) Text pre-processing

SVM is not good at classifying unbalanced sample sets, and the sample size of some categories in the experimental data set is far smaller than that of other categories, so this paper only selects five categories from the data set as the experimental corpus. In order to improve the accuracy of model classification, after eliminating duplicate text, missing text and minority categories, there are 5000 training sets and 1673 test sets. The information is shown in Table 1.

Use Python's built-in Chinese word segmentation component Jieba for word segmentation, summarize common Chinese stop words list, and filter the data after word segmentation with stop words list. Then manually mark the category of the document, and set the label as: $y_1=1$ represent aviation, $y_2=2$ represent computer, $y_3=3$ represent the art, $y_4=4$ represent the environment, $y_5=5$ represent agriculture.

Table 1. Experimental data information.

Category	Aviation	Computer	Art	Environment	Agriculture
Train	800	1000	900	1300	1000
Test	340	351	342	318	322
Total	1140	1351	1242	1618	1322

Text vectorization

In this paper, the weighted Word2vec algorithm is used to train the model to realize text vectorization. Firstly, use the Wikipedia Chinese corpus training Word2vec model [19], use the Word2Vec function of the gensim library in Python to get the word vector, and set dimension to 250. Figure 4 shows the specific process of training Word2vec model. Then the model is called to get the word vector matrix of the experimental corpus, and the weighted Word2Vec algorithm is used to get the sentence vector.

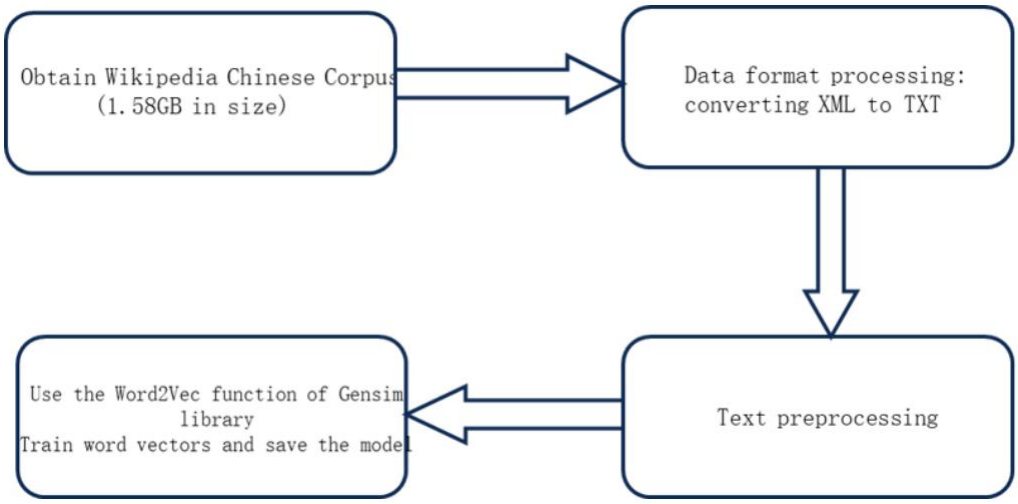

Figure 4. Training Word2Vec model based on Wikipedia corpus.

(3) Text feature selection

When the dimension of text feature vector is high, it will lead to problems such as noise features, increased computational complexity, and too long model training time. Therefore, this paper uses principal component analysis (PCA) to eliminate common words that are not easy to classify, so as to achieve the purpose of dimension reduction [20]. According to the maximum variance theory, extract the principal components whose cumulative contribution rate is greater than 85% and finally determine the first 60 dimensions as the model input, realize the dimensionality reduction of high-dimensional sentence vector matrices.

4.2 Analysis of experimental results

In this paper, the Accuracy, Recall and F1-score are selected as the model performance evaluation indicators.

Comparative experimental analysis of different word embedding representation methods

In order to verify the effectiveness of the improved weighted Word2vec algorithm, mean weighted Word2vec and the method proposed in this paper are respectively used for text vectorization, and experiments are carried out through SVM model. The comparison results of classification accuracy of the two methods under different categories are shown in Table 2.

Table 2. The impact of different word vector on classification results.

Category	Precision		Recall		F1	
	Word2Vec	Proposed method	Word2Vec	Proposed method	Word2Vec	Proposed method

Aviation	0.9	0.92	0.78	0.81	0.84	0.86
Computer	0.86	0.87	0.96	0.97	0.91	0.92
Art	1	1	1	1	1	1
Environment	0.91	0.93	0.93	0.94	0.92	0.94
Agriculture	0.99	1	1	0.99	0.99	0.99

It can be seen that, compared with the traditional Word2Vec method, the accuracy of texts classification in various categories has been improved by adding weight factors. Among them, the precision of the environmental category increased by 2%, and the recall rate increased by 1%. This is because the word embedding is weighted considering the importance of different categories of words, which increases the accuracy of subsequent classification.

Comparative experimental analysis of different parameter optimization methods

The CWWO-SVM model proposed in this paper is compared with WWO-SVM and PSO-SVM models. Among them, the inertia weight of the PSO algorithm is set to 0.5, and the learning factor is 2; The maximum wave height of the WWO algorithm is set to 6, the wavelength constraint coefficient is 1.0026, the maximum dimension of water wave splitting is 12, and the water wave breaking coefficient decreases linearly from 0.25 to 0.001.

According to the results in Table 3, compared with the PSO algorithm, the SVM classification accuracy combined with the WWO algorithm has improved in all aspects. Among them,the accuracy of traditional WWO-SVM is improved by 0.40%, and the accuracy of improved CWWO- SVM is improved by 0.82%. At the same time, the training time of the improved algorithm is shorter in the process of model parameter optimization. Therefore, it is confirmed that the WWO algorithm based on chaos theory has higher convergence accuracy and faster convergence speed,and shows better classification performance on the whole.

Table 3. Performance comparison of various classification models.

Classification Model	Precision	Recall	F1
PSO-SVM	93.92	93.45	93.3
WWO-SVM	94.32	93.45	93.24
CWWO-SVM	94.74	94.04	93.88

Comparative experimental analysis of CWWO-SVM with traditional classification algorithms Use CWWO algorithm to adjust SVM parameters, and determine the optimal parameters of SVM as follows: $C=100$，$\sigma=2.3731$ .So on the test set, the classification algorithm proposed in this paper is compared with SVM, K-Nearest Neighbors(KNN), Naive Bayes (NB),Decision Tree(DT) and Logistic Regression(LR).

From the classification results in Figure 6, it is easy to see that, compared with the traditional SVM, the F1-score of each category of the model is significantly improved after using the CWWO algorithm to determine the optimal parameters of SVM; the overall effect of logistic regression classification in traditional classification algorithms is worst. To sum up, compared with other classifiers, the SVM

classification model based on weighted Word2Vec and CWWO algorithm has better classification performance.

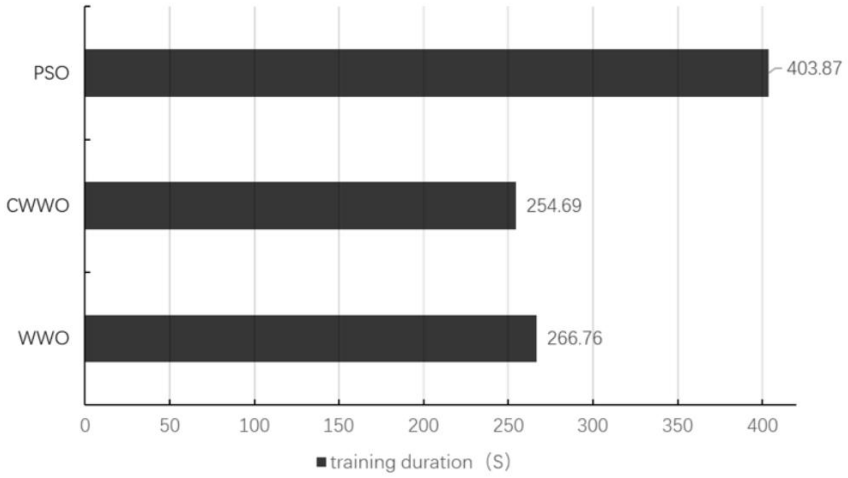

Figure 5. Training duration of each classification model.

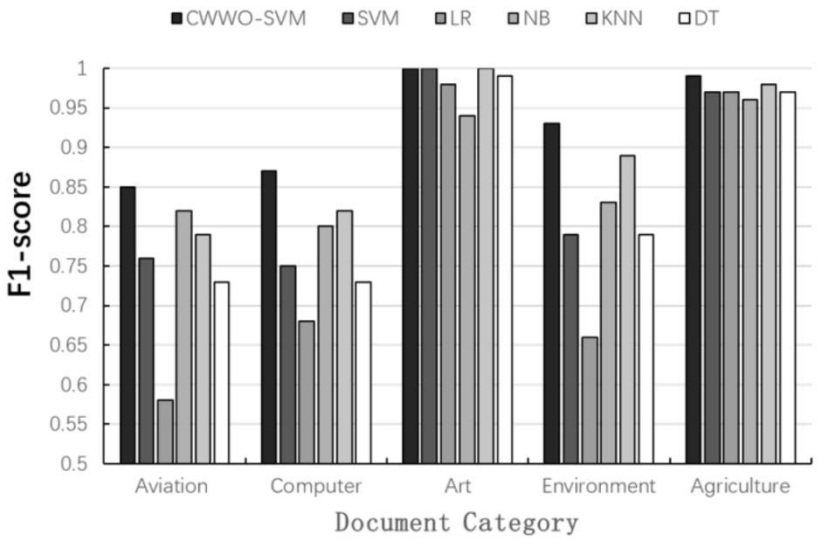

Figure 6. Classification accuracy of different algorithms under different categories.

5. Conclusion

The parameter optimization of SVM has always been an important problem to improve the generalization ability of the model, and various swarm intelligence optimization algorithms are also committed to solving this problem better. In this paper, a new intelligence optimization algorithm is studied, and an improved WWO algorithm based on chaos theory is proposed to solve the optimal parameter selection of SVM. Compared with other optimization algorithms (WWO, PSO), the feasibility and effectiveness of CWWO-SVM are verified. Finally, CWWO-SVM classifier is applied to Chinese text classification. At the same time, in order to further improve the efficiency of the model, this paper uses the weighted Word2vec method to process the text feature vector, and the overall classification effect of the model is also significantly improved. Therefore, this paper aims to solve the optimization problem of SVM parameters with a new algorithm, and provide an idea for the modeling of Chinese text classification.

References

[1] Chen Y.D., Qin Y.P. Review on Text Categorization Methods Based on Machine Learning. Bohai University (Natural Sciences), Vol.31 (2010) No.02, p.201 – 205. (In Chinese)

[2] Ma W., Chen G., Li X.J., et al. Chinese Comment Classification Based on Naive Bayesian Algorithm. Journal of Computer Applications, Vol.41 (2021) No.02, p.31 – 35. (In Chinese)

[3] Ding Zhengsheng, Ma Chunjie. Chinese Text Classification Algorithm Based on Improved Word Vectors and kNN. Modern Electronic Technology, Vol.45 (2022) No.01, p.100 – 103. (In Chinese)

[4] He Ming. A Maximum Entropy Model Text Classification Method Based on Improved Information Gain Feature Selection. Journal of Southwest Normal University (Natural Science Edition), Vol.44 (2019) No.03, p.113 – 118. (In Chinese)

[5] Yuan Q.Y. The Research and Implementation of Chinese Text Classification Technology Based on Decision Tree. Master's Thesis, Northeastern University, China, 2012, p.14. (In Chinese)

[6] Wang Yun, Li Cong. Application Research of Support Vector Machine Based on Adaptive Gravitational Search in Public Security Patrol and Police Situation Classification. Computer Applications and Software, Vol.37 (2020) No.07, p.56 – 60. (In Chinese)

[7] Feng Guohe. Comparison of SVM Classification Kernel Functions and Parameter Selection. Computer Engineering and Applications, Vol.47 (2011) No.03, p.123 – 124+128. (In Chinese)

[8] Lin Shijie, Dong Chen, Chen Mingzhi, Zhang Fan, Chen Jinghui. A Review of New Swarm Intelligence Optimization Algorithms. Computer Engineering and Applications, Vol.54 (2018) No.12, p.1 – 9. (In Chinese)

[9] Zheng Y.J. Water Wave Optimization: A New Nature-Inspired Metaheuristic. Computers and Operations Research, 2015, p.551 – 11. (In Chinese)

[10] Wang Wanliang, Chen Chao, Li Li, Li Weikun. Adaptive Water Wave Optimization Algorithm Based on Simulated Annealing. Computer Science, Vol.44 (2017) No.10, p.216 – 221+227. (In Chinese)

[11] Xue W. Python Machine Learning. China Machine Press, China, 2021, p.229. (In Chinese)

[12] Sun Fei, Guo Jiafeng, Lan Yanyan, Xu Jun, Cheng Xueqi. A Review of Distributed Word Representation. Journal of Computer Science, Vol.42 (2019) No.07, p.1605 – 1625. (In Chinese)

[13] Mikolov T., Chen K., Corrado G., et al. Efficient Estimation of Word Representations in Vector Space. Computer Science, 2013.

[14] Liu Qian, Feng Yanhong, Chen Yiying. Optimization Algorithm for Moth Flames Based on Chaos Initialization and Gaussian Mutation. Journal of Zhengzhou University (Engineering Edition), Vol.42 (2021) No.03, p.53 – 58. (In Chinese)

[15] Feng Yanhong, Liu Jianqin, He Yichao. Chaos-Based Dynamic Population Firefly Algorithm. Journal of Computer Application, Vol.33 (2013) No.03, p.796 – 799+805. (In Chinese)

[16] Dan Yuhao, Huang Jifeng, Yang Lin, et al. Research on Line Text Classification Based on TF-IDF and Word2Vec. Journal of Shanghai Normal University (Natural Sciences), Vol.49 (2020) No.01, p.89 – 95. (In Chinese)

[17] Yao Yanzhi, Li Jianliang. Feature Weight Analysis and Improvement of TF-IDF Based on Category Information. Computer System and Applications, Vol.30 (2021) No.09, p.237 – 241. (In Chinese)

[18] Guo Chaolei, Chen Junhua. Chinese Text Categorization Based on SA-SVM. Computer Applications and Software, Vol.36 (2019) No.03, p.277 – 281. (In Chinese)

[19] Deng Jun, Sun Shaodan, Wang Ruan, Song Xianzhi, Li He. Evolution Analysis of Weibo Public Opinion Based on Word2Vec and SVM. Study of Practical Experience, Vol.43 (2020) No.08, p.112 – 119. (In Chinese)

[20] Wen Wu, Wan Yuhui, Zhang Xuhong, Wen Zhiyun. Text Feature Selection Based on Improved CHI and PCA. Computer Engineering and Science, Vol.43 (2021) No.09, p.1645 – 1652. (In Chinese)