EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

Design of New Media Event Warning Method Based on K-means and Seasonal Optimization Algorithm

Zhenghan Gao[1,*] and Anzhu Zheng[2]

[1] Qingdao University of Technology, Qingdao 266000, Shandong, China
[2] Shandong University, Ji'nan 250000, Shandong, China

Abstract

INTRODUCTION: Timely and effective early warning of new media events not only provides academic value to the study of new media events, but also can play a positive role in promoting the resolution of public opinion.
OBJECTIVES: Aiming at the current research on early warning of new media events, there are problems such as the theoretical research is not in-depth and the early warning model is not comprehensive.
METHOD: In this paper, K-means and seasonal optimization algorithm are used to construct new media event early warning method. Firstly, by analyzing the construction process of new media event early warning system, extracting text feature vector and carrying out text feature dimensionality reduction; then, combining with the random forest algorithm, the new media event early warning method based on intelligent optimization algorithm optimizing K-means clustering algorithm is proposed; finally, the validity and superiority of the proposed method is verified through the analysis of simulation experiments.
RESULTS: The method developed in this paper improves the accuracy, time performance of new media event warning techniques.
CONCLUSION:Addresses the lack of comprehensiveness of current approaches to early warning of new media events.

Keywords: new media event warning methods, K-means clustering algorithm, seasonal optimization algorithm, random forests

Received on 7 September 2023, accepted on 25 January 2024, published on 2 January 2024

doi: 10.4108/eetsis.4873

1 Introduction

In recent years, with the continuous development of network communication technology, the awakening of netizens' consciousness and the gradual opening of the democratic system, a series of new media events have appeared with the New Media (The New Media) as the carrier, and the power of netizens from all walks of life is widely involved in and disseminated, which has caused a significant social impact [1]. At present, the world is entering an era of frequent new media events, and China is in the stage of socio-economic transformation, the continuous development of new media events can easily lead to mass incidents [2]. Mass events not only jeopardize the credibility and survival of enterprises, but also concern the development and stability of the country, causing obstacles to the positive and negative handling of mass events [3], therefore, timely and effective early warning of new media events has become a hot spot of research in the current academic community [4]. Timely and effective early

[*] Corresponding author. Email: gaozhenghan199311@163.com

warning of new media events not only provides academic value to the study of new media events, but also can play a positive role in promoting the resolution of public opinion [5]. Therefore, the research on new media events based on network technology is extremely necessary [6].

Constructing an effective and timely early warning system for new media events requires not only analyzing the process of network public opinion and crisis communication, but also studying the method of constructing early warning for new media events [7]. Therefore, the study of new media events mainly includes the study of network public opinion and crisis communication and the study of new media event early warning [8]. The new media event early warning construction method is mainly studied from the aspects of semantic analysis and opinion mining of network information text [9]. Literature [10] achieved better research results in the construction of sentiment corpus and single-grained viewpoint mining algorithm, which is of guiding significance to the construction of new media early warning model; Literature [11] analyzes the eight contents affecting network public opinion security for the network public opinion security monitoring and early warning of colleges and universities; Literature [12] analyzes eight contents affecting network public opinion security for the imperfection of public opinion early warning method in the establishment of network public opinion crisis early warning model In the establishment of network public opinion crisis early warning model, three sub-models of flexible public opinion mining, viewpoint evolution and network public opinion crisis early warning are established; Literature [13] utilizes the hierarchical analysis method, and constructs the early warning index system of network public opinion emergencies with three types of factors, namely, warning sources, warning signs and warning situations, by distributing questionnaires to the experts; Literature [14] starts from analyzing the current situation of network public opinion crisis early warning research findings, grasps the network public opinion early warning research and practice as a whole, and gives the corresponding strategies. and give the corresponding strategies. For the above literature analysis, the existing event early warning methods have the following defects [15]: 1) in-depth theoretical research, lack of methodological practice; 2) early warning methods are only analyzed from the perspective of text clustering, without forming the construction of the early warning system; 3) there are fewer algorithms for the new media early warning.

Aiming at the problems existing in the current new media event early warning algorithm method, this paper proposes a new media event early warning system design method based on clustering algorithm and random forest algorithm. The main contributions of this paper are: (1) By analyzing the construction process of new media event early warning system, it focuses on describing the steps of pre-processing new media event early warning information, partition mining, feature representation, dimensionality reduction, etc.; (2) By combining the K-means algorithm, seasonal heuristic optimization algorithm and Random Forest algorithm, it puts forward the methodology approach for new media event early warning methodology based on the combination of K-means and seasonal optimization algorithm; (3) The effectiveness of the proposed method in this paper is verified by experimental analysis.

2 Analysis of the construction of new media event early warning system

The complete flow of the early warning system is shown in Fig. 1, including information acquisition, preprocessing, text segmentation and mining, text clustering, and model evaluation, in which the timely acquisition of information and the correct selection of clustering algorithms are key [16].

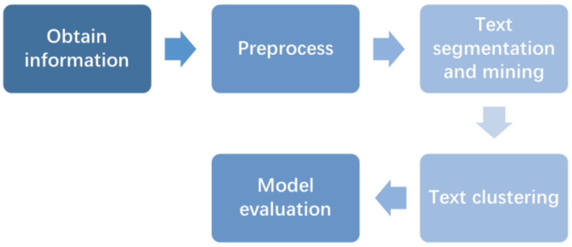


Fig. 1 Early warning system construction

(1) Early warning system information acquisition and pre-processing

Information acquisition, as the first step of network public opinion detection and analysis, generally starts from the homepage of a website, extracts all information in the current webpage, including webpage content and hyperlink addresses, and then discovers new webpages through hyperlink addresses, and keeps cycling until all the information of the website is collected [17]. Through the literature survey, it is found that the webpage post click rate, reply rate, and reprint rate directly reflect the attention of the post content, therefore, when the click rate, reply rate, and reprint rate reaches a certain threshold, the post information is alerted by the warning system, so as to collect the information content in a timely manner. At the same time, the acquired information is preprocessed to propose the content of the information that is not related to public opinion analysis and extract all the content that is related to public opinion analysis [18].

(2) Text Segmentation and Mining

Text segmentation is the basis for text analysis and processing, and is also a key step that affects the effect of late text processing. Text analysis includes Chinese lexicalization and English lexicalization. English participle only needs to delete the virtual word and the root of the word, as well as to consider the English word division too and the singular-plural change [19]; Chinese participle adopts the Chinese lexical analysis system, which can do the industry dictionaries, user-defined dictionaries, multilevel lexical annotation, keyword extraction, fingerprint extraction, etc. [20].

Text mining of new media event information needs to represent text information as numerical values or symbol vectors. For Chinese features, text mining generally adopts words as feature items. In this paper, we adopt Vector space model (VSM) [21], which reduces the processing of text content to vector operations in vector space, and expresses the semantic similarity with the similarity of the vectors in the space.VSM represents each text as an orthogonally normalized vector composed of feature terms, and the model is as follows:

$$D=\{(T_1,W_1),(T_2,W_2),(T_3,W_3),\cdots,(T_n,W_n)\} \quad (1)$$

T_i W_i T_i $(0,1)$ W_i T_i In Eq. (1), represents the value of the ith feature item; represents the vector weight of the feature item , which is generally in the range of , and the larger the value of , the more the feature item reflects the content of the text D. The vector weight of the feature item is the same as that of the feature item .

T_i The vector weight calculation of feature term is divided into clustering model weight calculation and classification weight calculation. The clustering model weight calculation adopts word frequency weighting method, and the specific calculation formula is as follows:

$$W_i=tf_i/(tf_1+tf_2+tf_3+\cdots+tf_n) \quad (2)$$

tf_i T_i In Eq. (2), denotes the number of times the feature item appears in the text.

The classification sub-model weights are calculated using the TF-IDF weighting method with the following formulas:

$$W_i=2\log(tf_i+1)\cdot(2-\log(N/(n_i+1))) \quad (3)$$

N n_i T_i In Eq. (3), is the total number of texts pushed by the website in the same time period, and is the number of texts containing the feature term .

(3) Text feature dimensionality reduction

Since the text representation is a vector space model, which makes the vector dimensions too much, thus severely limiting the computing speed of the whole system, the high dimensional vectors make the clustering overfitting. In order to reduce the overfitting of clustering and the blurring of results, this paper adopts the method of feature extraction to reduce the dimensionality of vector features with high dimensionality [22]. Analyzing the previous new media events, the situation that can reflect the whole event involves five types of words: location, person, relevant department, event, and ending, so the final vector space model dimension is five dimensions, as shown in equation (4):

$$D=\{(T_1,W_1),(T_2,W_2),(T_3,W_3),(T_4,W_4),(T_5,W_5)\} \quad (4)$$

(4) Text Clustering

The clustering model of the early warning system is the basis for carrying out the classification of new media events. The text clustering sub-model is to categorize previous new media events into a sample set, divide them into several clusters based on some strategy, where the severity of new media events in each cluster is comparable, and find the center of each cluster [23].

(5) Classification of events

Classification of new media events is different from ordinary text classification or breaking news classification, the new media event information collected through the network is not a single text, but a collection of multiple texts, the traditional text classification is not applicable to the classification of new media events. According to the clustering results, it will be used as the class label of the sample, and then a classifier will be applied to classify it [24].

3 New media event clustering model based on K-means and SO algorithm

3.1 K-means clustering algorithm

The role of clustering algorithms is to categorize large and complex data of learning activities, learning styles, and learning states into one class of similar data according to the learning attributes, which lays the foundation for the next step of intelligent learning model training. It can be said that the accurate categorization of data directly affects the construction of the sample set, so it is crucial to find a high-performance clustering algorithm.

The K-means algorithm is the most popular divisive clustering algorithm, which performs well in dealing with big data classification [25]. The algorithm determines the clustering centers and the elements to which they belong by minimizing an objective function based on squared error. The aim is to keep the cluster centers as far away from each other as possible and associate each data point to the nearest cluster center. In the K-means algorithm, the Euclidean distance is commonly used as a similarity measure. Where a small distance indicates strong similarity while a large distance indicates low similarity.

The objective function of the K-means algorithm is defined as equation (5):

$$J=\sum_{i=1}^{K}\left(\sum_{k}\|x_k-c_i\|^2\right) \quad (5)$$

K c_i x_k i k In equation (5), is the number of clusters, is the center of the cluster, and is the th data point in the th cluster.

The exact procedure of the algorithm is as follows:

K K $C=(c_1,c_1,\cdots,c_K)$ Step 1: Determine the total number of classified categories and randomly select cluster category centers .

Step 2: Compute the partition matrix. A data point belongs to the cluster whose center is closest to that data point. U Therefore, the clusters are represented by the binary division matrix . U The elements in it are determined as in equation (6):

$$u_{ij}=\begin{cases}1 & if\ \left\|x_j-c_i\right\|^2\le\left\|x_j-c_t\right\|^2,\forall t\ne i\\ 0 & otherwise\end{cases}\qquad(6)$$

u_{ij} j i In Eq. (6), indicates whether the th data point belongs to the th cluster class.

Step 3: Update the cluster centers. c_i Define each cluster class center that minimizes the objective function as being (7):

$$c_i=\frac{\sum_{j=1}^{N}u_{ij}x_j}{\sum_{j=1}^{N}u_{ij}}\qquad(7)$$

N In equation (7), denotes the number of samples.

Step 4: Compute the objective function using equation (1). Verify that the function converges or that the difference between two neighboring values of the objective function is less than a given threshold and stop. Otherwise repeat step 2.

3.2 Seasonal optimization algorithm

Most parts of the world possess the phenomenon of four seasons of the year, which include spring, summer, fall, and winter. As the weather changes during the seasons, biological features, especially trees, adapt to the weather by changing their behavior. In spring, trees begin to flourish and grow, and seeds begin to germinate and grow; in summer, trees grow by competing with each other for light, water, nutrients, and other elements; in fall, seed dispersal becomes an important aspect of the development of tree populations, and through seed dispersal trees expand their growth areas and secure their survival space; and in winter, trees begin to hibernate, and gain their survivability by enduring winter freezes. Based on the inspiration of the survival cycle of trees in different seasons, this paper proposes a Seasons optimization algorithm (SO) [26].The SO algorithm is a swarm intelligent optimization algorithm, which takes the trees as the candidate solutions, and aims at finding the strongest trees by seeking the optimal ones.

The SO algorithm has four operators which are recovery, competition, seeding, and resistance.The flowchart of the SO algorithm is shown in Fig. 2.

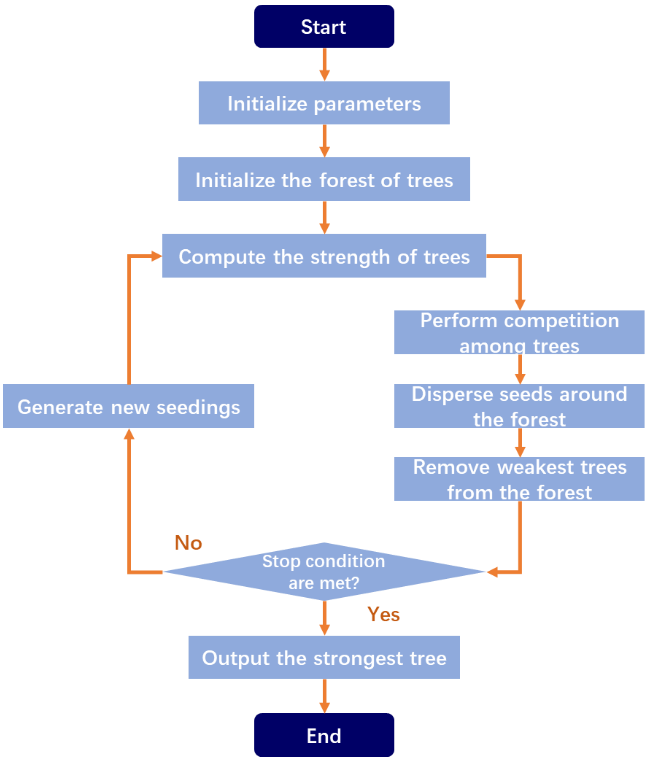

Start
Initialize parameters
Initialize the forest of trees
Compute the strength of trees
Perform competition among trees
Disperse seeds around the forest
Remove weakest trees from the forest
Generate new seedings
No
Stop condition are met?
Yes
Output the strongest tree
End

Fig. 2 Seasonal optimization algorithm flow

(1) Forest initialization

$f(T)$ Suppose is a D-dimensional optimization decision problem defined as equation (8):

$$\begin{gathered}f(T)=f(t_1,t_2,\cdots,t_D)\\ \forall t_i\in[l_i,u_i]\end{gathered}\qquad(8)$$

t_i l_i u_i In Eq. (8), denotes the ith-dimensional variable of the problem, and and denote the lower and upper bounds of the ith dimension of the decision variable, respectively.

$f(T)$ F To solve the problem, the algorithm first randomly initializes the forest :

$$F=[T_1,T_2,\cdots,T_N]\qquad(9)$$

$$T_i=[t_{i1},t_{i2},\cdots,t_{iD}]\qquad(10)$$

$$t_{ij}=l_{ij}+\varphi_{ij}\cdot(u_{ij}-l_{ij})\qquad(11)$$

φ_{ij} $[0,1]$ where denotes a uniformly distributed random number between .

Define the strength of the tree as the value of the objective function, which is defined as equation (12):

$$S_i=S(T_i)=S(t_{i1},t_{i2},\cdots,t_{iD})\qquad(12)$$

In equation (12), the strength of a tree can be thought of as the tree's ability to reproduce, survive, and store. The stronger a tree is, the more water, nutrients, and sunlight it receives, the better its chances of survival and reproduction.

(2) Recovery operator

The recovery operator mainly simulates the spring behavior of trees, and the specific mathematical model is as follows:

$$F^{y+1}=\left\{F^{y}\right\}+\left\{R\right\} \tag{13}$$

F^{y} R In Eq. (13), denotes the yth generation forest population and denotes the new seedling produced in the current generation:

$$R=\Phi\left(p_r\times A^{y}\right) \tag{14}$$

p_r A^{y} Φ In Eq. (14), denotes the recovery rate, is the number of seeds falling to the ground in the previous fall, and simulates the process of natural germination of seeds into seedlings. The process of seeds turning into seedling is affected by the number of iterations, and as the number of iterations increases, the recovery rate of seeds increases, and the probability of local search increases, which enhances the exploration ability of the algorithm, so that the forest population avoids falling into the local optimum, and tries to make up for the freezing of the weak seeds in winter. p_r Recovery rate varies with the number of iterations in the following formula:

$$p_r=p_{\max}-\frac{y}{Y}\left(p_{\max}-p_{\min}\right) \tag{15}$$

$p_{\max}$ $p_{\min}$ y Y In Eq. (15), and denote the maximum and minimum recovery rates, respectively, and and denote the current and maximum number of iterations, respectively.

(3) Competition operator

The competition operator mainly simulates the forest population behavior in summer. N_c To simulate the competitive operation, the tree populations were arranged in decreasing order according to the definition of robustness, and the first trees were selected as the core trees:

$$N_c=\left\lceil p_c\times N\right\rceil \tag{16}$$

$\Gamma=\left[T_1,T_2,\cdots,T_c\right]$ T_i $\left(S_i-S_{i+1}\right)\le\cdots\le\left(S_i-S_{N_c}\right)$ where the core tree inventory set is and satisfies .

Neighboring trees are defined below:

$$Z_i=\left|\tau_i\times N_g\right| \tag{17}$$

$$N_g=N-N_c \tag{18}$$

Z_i T_i N_g τ_i T_i where each neighboring tree belongs to the core tree, denotes the number of neighboring trees of the core tree , denotes the number of neighboring trees, and denotes the normalized robustness value of the core tree :

$$\tau_i=\frac{S_i-\min\left(I\right)}{\sum_{k=1}^{N}S_k},I=\left\{S_k\left|k=1,2,\cdots,N\right.\right\} \tag{19}$$

Z_i T_i Neighborhood trees were randomly selected from the set of neighboring trees of the core tree . T_i In each neighborhood, the mathematical model of the competitive impact of the core tree is as follows:

$$T_j^{y+1}=\frac{1}{\Lambda_j+1}\times\varphi\left(T_j^{y}\right) \tag{20}$$

$$\varphi\left(T_j^{y}\right)=T_j^{y}+\theta \tag{21}$$

$$\Lambda_j=\sum_{k=1}^{Z_i}S_k\times\Delta_{jk}^{-2}\times\lambda_{jk} \tag{22}$$

T_j^{y+1} T_j $\varphi\left(\cdot\right)$ T_j θ Λ_j Where, denotes the location information of the tree in the y+1th generation, denotes the function of the growth of the tree without the influence of the neighboring trees, is the number of uniform distributions, and is calculated based on the number of neighboring trees, distance, and robustness value. Δ_{jk} T_j denotes the distance between the core tree and the kth neighboring tree, and the calculation model is as follows:

$$\Delta_{jk}=\sqrt{\sum_{z=1}^{D}\left(T_{jz}-T_{kz}\right)^2} \tag{23}$$

λ_{jk} T_j denotes the impact value of the core tree with the kth neighboring tree, defined as follows:

$$\lambda_{jk}=\begin{cases}1 & if\left(S_k\ge S_j\right)\\ 1-\gamma & else\end{cases} \tag{24}$$

γ $\left[0,1\right]$ In Eq. (24), denotes a random misfit factor with a value ranging from , which represents a smaller degree of discounting of neighboring tree impacts.

For better access to core tree positions, the competing operator uses a winner-take-all strategy:

$$T_i^{y+1}=\begin{cases}T^{*} & if\left(S\left(T_i^{y}\right)\le S\left(T^{*}\right)\right)\\ T_i^{y} & else\end{cases} \tag{25}$$

T^{*} T_i In its equation (25), denotes the optimal neighboring trees around the core tree . From the above operator model, it can be seen that the competition operator strong trees and weak trees have the opportunity to improve the survival robustness.

(4) Seeding operator

The seeding operator mainly simulates the fall behavior of trees. $\Upsilon=[T_1,T_2,\cdots,T_A]$ To simulate the seeding operator model, a list of seeds is formed from a random selection of tree populations , and the seed simulation is calculated as follows:

$$A=\psi\left(p_s\times N\right) \tag{26}$$

p_s $\psi(\cdot)$ $p_s\times N$ where is the seeding rate, whose value is a random number, and is a selection function that selects the strongest trees from the forest population. A In order to stop the forest population size from increasing indefinitely, it is assumed that only one seed is selected per tree per fall, so that the number of seeds per generation is constant equal to .

T_i For the individual in the seed list , the relevant dimension variables are randomly selected to be exchanged with the new location dimension variables, and the new location of the dimension variables is computed in the following model:

$$t'_j=t_j+\ell\times r \tag{27}$$

r $\left[l_j,u_j\right]$ ℓ $\{1,-1\}$ Where is the random number in and is the binary quantity in .

(5) Resistance operator

The resistance operator focuses on modeling the winter impact behavior of trees. This operator mainly removes weak trees from the forest:

$$F^{y+1}=\left\{F^y\right\}-\{W\} \tag{28}$$

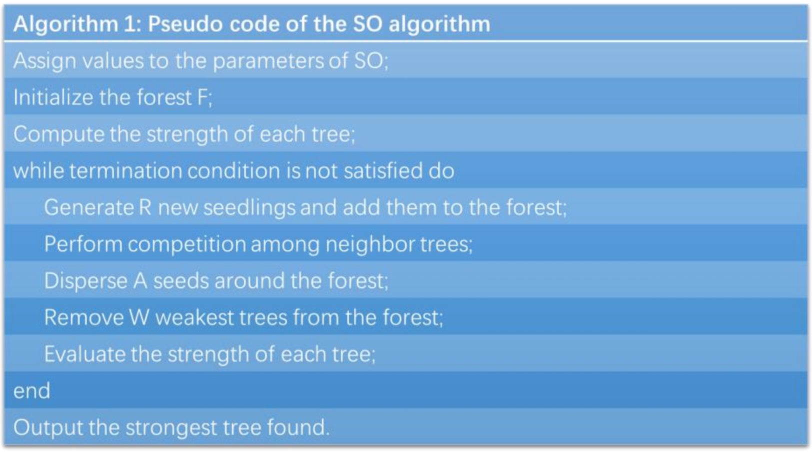
W Where denotes weak trees, the specific selection function is calculated as follows:

$$W=\chi\left(p_w\times N\right) \tag{29}$$

$$p_w=1-|\rho| \tag{30}$$

p_w $\chi(\cdot)$ ρ $[-1,0)$ Where denotes the tree resistance rate, is the weak tree selection function, and denotes the critical temperature value in the range of . ρ $-\left(1-p_s\right)$ In order to keep the forest population size constant, is usually set to .

The SO algorithm pseudo-code is shown in Fig. 3.



```
Algorithm 1: Pseudo code of the SO algorithm
Assign values to the parameters of SO;
Initialize the forest F;
Compute the strength of each tree;
while termination condition is not satisfied do
    Generate R new seedlings and add them to the forest;
    Perform competition among neighbor trees;
    Disperse A seeds around the forest;
    Remove W weakest trees from the forest;
    Evaluate the strength of each tree;
end
Output the strongest tree found.
```

Fig. 3 SO algorithm pseudo-code

3.3 New media clustering model based on SO algorithm optimized K-means

In order to increase the accuracy of the K-means clustering method, this paper uses the SO algorithm in K-means combined with the new media event clustering analysis, specifically refers to the use of the SO algorithm to find a set of optimal clustering centers to make the text in all categories of all text to change the minimum distance to the clustering center. K-means-SO algorithm decision variable for the clustering center, the sum of the intraclass distance between the text documents as the SO algorithm's fitness assessment function, as follows:

$$J=\sum_{j=1}^{K}\sum_{\forall s_j\in c_j}d\left(s_i,c_j\right)^2 \tag{31}$$

$d\left(s_i,c_j\right)$ Where, is the text similarity, which is calculated as follows:

$$d\left(s_i, c_j\right) = 1 - \cos\left(s_i, c_j\right) = 1 - \frac{s_i \cdot c_j}{\left\|s_i\right\| \cdot \left\|c_i\right\|} \quad (32)$$

In Eq. (32), the higher the text similarity, the value of cosine similarity is 1 and the value of distance between two documents is 0. The lower the text similarity, the value of cosine similarity is 0 and the value of distance between two documents is 1.

The steps of the new media event clustering model by combining the SOS algorithm with the K-mean clustering algorithm are as follows, Figure 4:

Step 1: Acquire the dataset. After the text D has been segmented and dimensionality reduced, the text data is converted to numerical data, which is used as the original dataset S for the text clustering algorithm;

Step 2: Initialize the population. $F = \left[T_1, T_2, \cdots, T_N\right]$ Initialize the number of clustering categories K and randomly initialize the forest population according to different dimensions to obtain the forest population ;

Step 3: Calculate fitness values. $d\left(s_i, X_j\right)$ Calculate the cosine similarity of each data object in the numerical new dataset S hitting the K clustering centers represented by each tree individual of the initial forest population, and then calculate the textual similarity , assign the data object to the closest class cluster, direct all the data to be assigned, and finally calculate the sum of intraclass distances of all the data objects in the class clusters of the individual trees of each tree individual, i.e., the value of the fitness function;

Step 4: Position update. Perform recovery, competition, seeding, and resistance operators according to the SO algorithm optimization strategy;

Step 5: Determine whether the K-means-SO text clustering algorithm reaches the maximum number of iterations or satisfies the convergence condition, if yes, output the optimal clustering center; otherwise, loop iterate step 3 to step 5;

Step 6: Output the text clustering results and assign the corresponding text document data to the corresponding categories according to the final clustering results.

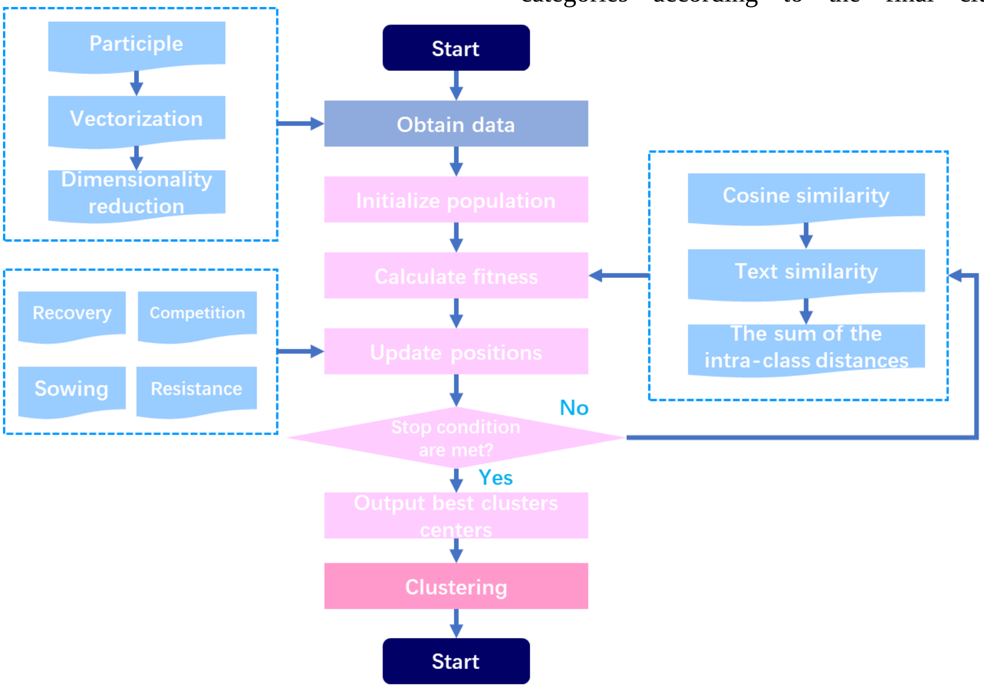


Fig. 4 Flowchart of new media clustering method based on SO algorithm optimized K-means

4 Random Forest Algorithm Based Classification Model for New Media Events

4.1 Random Forest Algorithm

Random Forest (RF) algorithm is an integrated learning algorithm based on decision trees, and its algorithmic principle is based on the Bagging integration algorithm and the random subspace method [27].The RF algorithm utilizes the samples to be classified, trains to produce a decision tree and determines the predicted classification results from all the results of all the decision trees by aggregating the results by voting, and the specific steps are as follows in Fig. 5:

Step 1: Randomly draw samples, construct different sub-datasets, and train decision trees;

Step 2: Randomly select m attributes and use the information gain strategy to select the optimal attributes for decision tree node splitting and train to form a decision tree;

Step 3: Build different decision trees according to steps 1~2 and integrate to construct a random forest model.

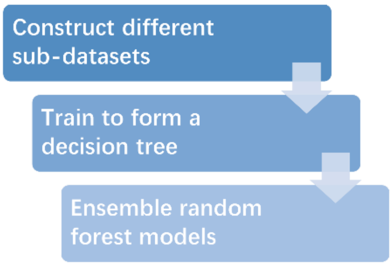

Construct different sub-datasets
Train to form a decision tree
Ensemble random forest models

Fig. 5 Schematic diagram of random forest

4.2 Categories of new media events

Through clustering it can be analyzed that new media events are classified into different types of new media events based on different class clusters. The new media events of different class clusters train the classifier to distinguish which type of new media events the new new media events belong to. In order to deal with new media events correctly, new media events can be divided into new media events that are favorable to the development and stability of society, new media events that belong to the general netizens' self-indulgence, events that have a certain negative impact on the development and stability of society, and events that seriously affect the development and stability of society according to the degree of impact [28], which are classified in the following Figure 6.

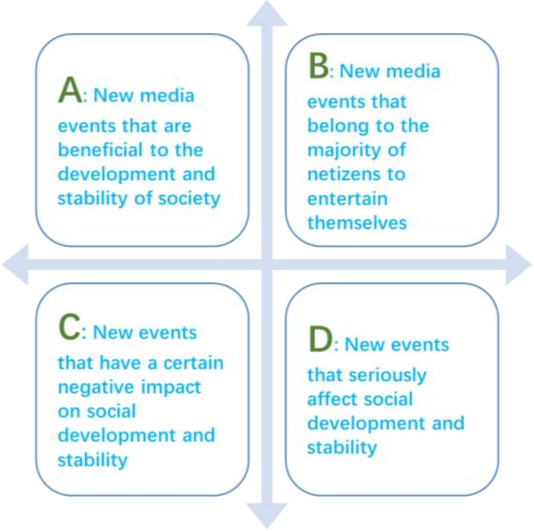

A: New media events that are beneficial to the development and stability of society
B: New media events that belong to the majority of netizens to entertain themselves
C: New events that have a certain negative impact on social development and stability
D: New events that seriously affect social development and stability

Fig. 6 Schematic diagram of new media event categorization

4.3 New media event classification model based on RF algorithm

According to the RF algorithm and the new media event classification type, the flowchart of the new media event classification model based on the RF algorithm is shown in Fig. 7, and the specific steps are as follows:

Step 1: According to the clustering can be analyzed to get the dataset with four different class clusters and labeled with new media event categories;

Step 2: Randomly draw a sample set to train the decision tree;

Step 3: Randomly select m attributes and train to form a decision tree by node splitting;

Step 4: Integrate the construction of a random forest model;

Step 5: Test the new media event classification model based on RF algorithm using test dataset.

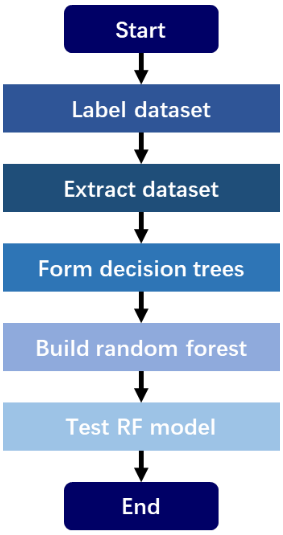


Fig. 7 Flowchart of new media event classification model based on RF algorithm

5 Flow of new media event early warning method based on K-means-SO+RF model

In order to improve the accuracy of new media event warning, combining K-means with SO algorithm and RF algorithm, this paper proposes a new media event warning method based on K-means-SO+RF model, the specific flow chart is shown in Fig. 8, and the specific steps are as follows:

Step 1: Collect data. Capture textual information about the event from the web and count the characters;

Step 2: Pre-processing of text information. Combine all the contents of the message into a string, removing all useless characters;

Step 3: The characters are disambiguated using a disambiguation system, and after feature dimensionality reduction, the key features are obtained;

Step 4: Calculate the weight of each feature term and analyze the feature term semantic term strength;

Step 5: Set the number of clusters K=4 and utilize the K-means-SO algorithm for cluster analysis of the text dataset;

Step 6: According to the clustering results, analyze the clustered class clusters, label the labels, and construct the new media event classification model based on RF algorithm;

Step 7: Evaluate the new media event alerting methodology using the test set.

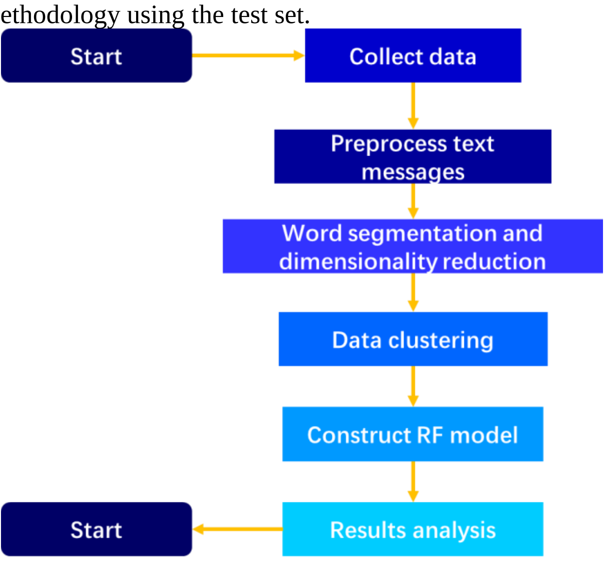


Fig. 8 Flowchart of new media event warning method based on K-means-SO+RF algorithm

6 Experiment and Result Analysis

6.1 Experimental environment setup

In order to verify the advantages and disadvantages of the new media event early warning methods proposed in this paper, five early warning algorithms are selected for comparison, and the specific parameters of each algorithm are set as in Table 1.The experimental simulation environment is Windows 10, with a CPU of 2.80GHz, 8GB of RAM, and a programming language of Matlab2023a.The data used in this paper are from the Internet Haimen Event Crawl data.

Table 1 Parameter settings for new media event warning methods

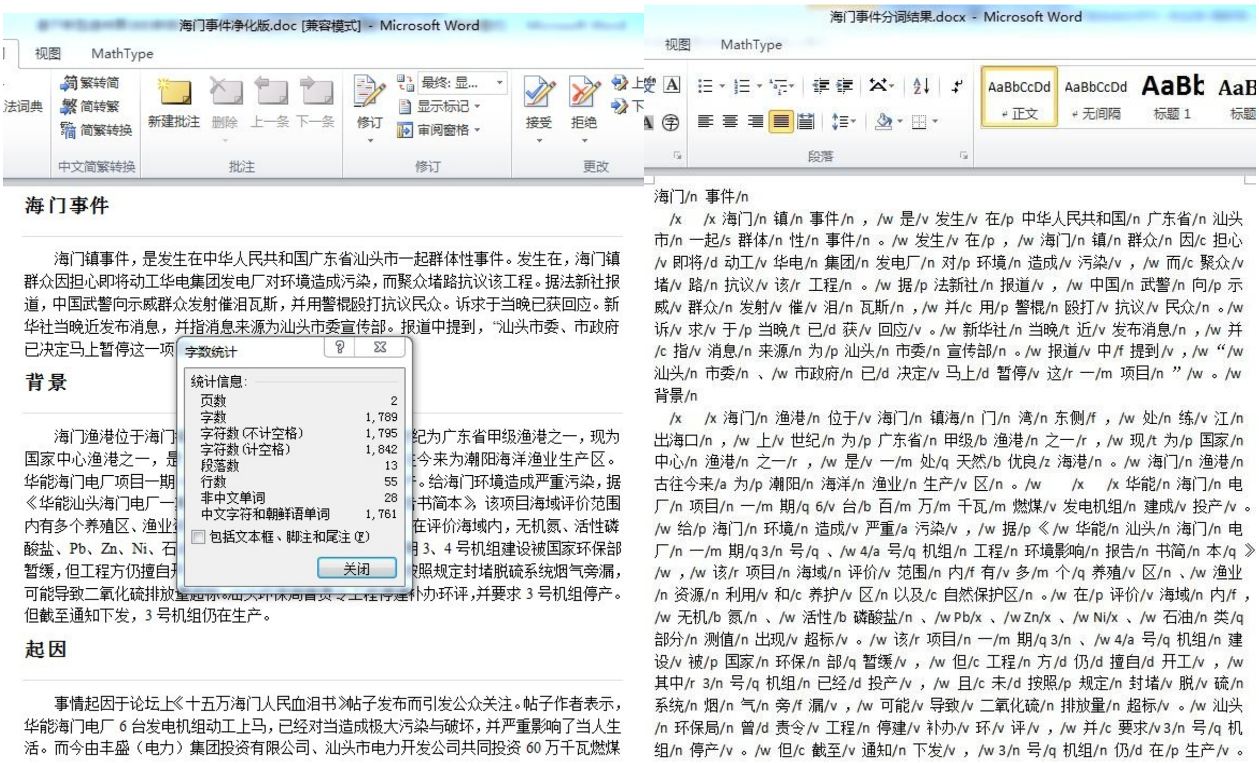
arithmetic	parameterization
K-means+SVM	Number of clusters K=4; C=10, ε=0.05
K-means+BP	The number of clusters K=4; the hidden layer nodes are 50, and the activation function is radial basis function;
K-means+RF	Number of clusters K=4; N_tree=500, m_try=floor(80.5)
K-means-SO+SVM	Number of clusters K = 4; SO algorithm population is 50, number of iterations is 500; C = 10, ε = 0.05
K-means-SO+BP	The number of clusters K=4; the population of SO algorithm is 50, the number of iterations is 500; the number of hidden layer nodes is 50, the activation function is radial basis function
K-means-SO+RF	Number of clusters K=4; SO algorithm population is 50, number of iterations is 500; N_tree=500, m_try=floor(80.5)

6.2 Description of experimental results

Following the methodology flow, the following experimental results are given in this section in turn:

(1) Organize the text information about the Haimen Incident pushed by the background of the network, which is counted as 2012 characters, including 1,852 Chinese characters.

(2) Text information preprocessing, remove all the useless characters, the remaining characters 1842, Chinese characters for 1761, specific as shown in Figure 9.



海门事件

海门镇事件，是发生在中华人民共和国广东省汕头市一起群体性事件。发生在，海门镇群众因担心即将动工华电集团发电厂对环境造成污染，而聚众堵路抗议该工程。据法新社报道，中国武警向示威群众发射催泪瓦斯，并用警棍殴打抗议民众。诉求于当晚已获回应。新华社当晚近发布消息，并指消息来源为汕头市委宣传部。报道中提到，"汕头市委、市政府已决定马上暂停这一项

字数统计

统计信息:

页数	2
字数	1,789
字符数(不计空格)	1,795
字符数(计空格)	1,842
段落数	13
行数	55
非中文单词	28
中文字符和朝鲜语单词	1,761

☐ 包括文本框、脚注和尾注(F)

关闭

背景

海门渔港位于海门 纪为广东省甲级渔港之一，现为
国家中心渔港之一，是 今来为潮阳海洋渔业生产区。
华能海门电厂项目一期 。给海门环境造成严重污染，据
《华能汕头海门电厂一 书简本》，该项目海域评价范围
内有多个养殖区、渔业 在评价海域内，无机氮、活性磷
酸盐、Pb、Zn、Ni、石 3、4 号机组建设被国家环保部
暂缓，但工程方仍擅自开 按照规定封堵脱硫系统烟气旁漏，
可能导致二氧化硫排放 补办环评，并要求 3 号机组停产。
但截至通知下发，3 号机组仍在生产。

起因

事情起因于论坛上《十五万海门人民血泪书》帖子发布而引发公众关注。帖子作者表示，华能海门电厂 6 台发电机组动工上马，已经对当造成极大污染与破坏，并严重影响了当人生活。而今由丰盛（电力）集团投资有限公司、汕头市电力开发公司共同投资 60 万千瓦燃煤

海门/n 事件/n

/x /x 海门/n 镇/n 事件/n ，/w 是/v 发生/v 在/p 中华人民共和国/n 广东省/n 汕头市/n 一起/s 群体/n 性/n 事件/n 。/w 发生/v 在/p ，/w 海门/n 镇/n 群众/n 因/c 担心/v 即将/d 动工/v 华电/n 集团/n 发电厂/n 对/p 环境/n 造成/v 污染/v ，/w 而/c 聚众/v 堵/v 路/n 抗议/v 该/r 工程/n 。/w 据/p 法新社/n 报道/v ，/w 中国/n 武警/n 向/p 示威/v 群众/n 发射/v 催/v 泪/n 瓦斯/n ，/w 并/c 用/p 警棍/n 殴打/v 抗议/v 民众/n 。/w 诉/v 求/v 于/p 当晚/t 已/d 获/v 回应/v 。/w 新华社/n 当晚/t 近/v 发布消息/n ，/w 并/c 指/v 消息/n 来源/n 为/p 汕头/n 市委/n 宣传部/n 。/w 报道/v 中/f 提到/v ，/w "/w 汕头/n 市委/n 、/w 市政府/n 已/d 决定/v 马上/d 暂停/v 这/r 一/m 项目/n "/w 。/w

背景/n

/x /x 海门/n 渔港/n 位于/v 海门/n 镇海/n 门/n 湾/n 东侧/f ，/w 处/n 练/v 江/n 出海口/n ，/w 上/v 世纪/n 为/p 广东省/n 甲级/b 渔港/n 之一/r ，/w 现/t 为/p 国家/n 中心/n 渔港/n 之一/r ，/w 是/v 一/m 处/q 天然/b 优良/z 海港/n 。/w 海门/n 渔港/n 古往今来/a 为/p 潮阳/n 海洋/n 渔业/n 生产/v 区/n 。/w /x /x 华能/n 海门/n 电厂/n 项目/n 一/m 期/q 6/v 台/b 百/m 万/m 千瓦/m 燃煤/v 发电机组/n 建成/v 投产/v 。/w 给/p 海门/n 环境/n 造成/v 严重/a 污染/v ，/w 据/p 《/w 华能/n 汕头/n 海门/n 电厂/n 一/m 期/q 3/n 号/q 、/w 4/a 号/q 机组/n 工程/n 环境影响/n 报告/n 书简/n 本/q 》/w ，/w 该/r 项目/n 海域/n 评价/v 范围/n 内/f 有/v 多/m 个/q 养殖/v 区/n 、/w 渔业/n 资源/n 利用/v 和/c 养护/v 区/n 以及/c 自然保护区/n 。/w 在/p 评价/v 海域/n 内/f ，/w 无机/b 氮/n 、/w 活性/b 磷酸盐/n 、/w Pb/x 、/w Zn/x 、/w Ni/x 、/w 石油/n 类/q 部分/n 测值/n 出现/v 超标/v 。/w 该/r 项目/n 一/m 期/q 3/n 、/w 4/a 号/q 机组/n 建设/v 被/p 国家/n 环保/n 部/q 暂缓/v ，/w 但/c 工程/n 方/d 仍/d 擅自/d 开工/v ，/w 其中/r 3/n 号/q 机组/n 已经/d 投产/v ，/w 且/c 未/d 按照/p 规定/n 封堵/v 脱/v 硫/n 系统/n 烟/n 气/n 旁/f 漏/v ，/w 可能/v 导致/v 二氧化硫/n 排放量/n 超标/v 。/w 汕头/n 环保局/n 曾/d 责令/v 工程/n 停建/v 补办/v 环/v 评/v ，/w 并/c 要求/v 3/n 号/q 机组/n 停产/v 。/w 但/c 截至/v 通知/n 下发/v ，/w 3/n 号/q 机组/n 仍/d 在/p 生产/v 。

Fig. 9 File information after event preprocessing

(3) For the collected new media event text, after the word separation with the word separation ICTCLAS2011 system, the vectorized text feature items are subjected to feature dimensionality reduction, and the words with the highest frequency counts after the dimensionality reduction are: T1 mass (13 times), T2 demonstration (11 times), T3 blockage (5 times), T4 protest (5 times), and T5 contamination (4 times), and the specific results of the word separation are shown in Fig. 10.

$$W_1 = 13/38 \quad W_2 = 11/38 \quad W_3 = 5/38 \quad W_4 = 5/38$$

$W_5 = 4/38$ (4) Calculate the weight of each feature item, and the calculation results are shown in Fig. 11, , , , , .

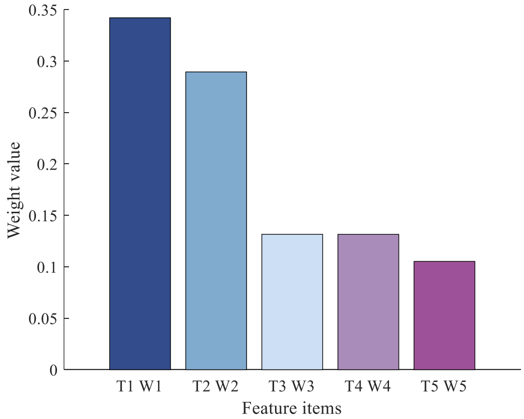


Fig. 10 Results of weight calculation

(5) Setting the number of clustering class clusters 4, the new media event feature vector is clustered and analyzed, and the clustering results are shown in Figure 11. Using the clustering results, the data are labeled, with cluster 1 indicating events that seriously affect the development and stability of society, cluster 2 indicating events that have a certain negative impact on the development and stability of society, cluster 3 indicating events belonging to the general netizens' self-indulgence, and cluster 4 indicating events that are beneficial to the development and stability of society.

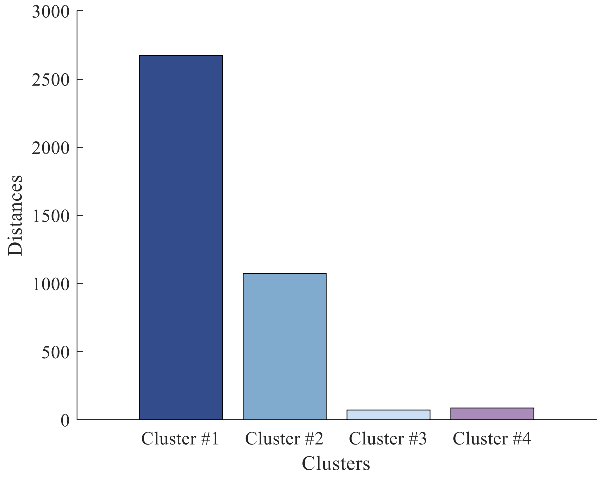


Fig. 11 Clustering results

(6) The clustering results are composed into a classification model dataset, which is fed into the training RF classifier to obtain the new media event classification model. From the analysis of the above results, it can be seen that the Haimen incident belongs to the third category of new media events, which is a group event triggered by the normal demands of the masses, and there is a tendency to evolve to the fourth category of new media events.

6.3 Experimental Performance Analysis

In order to verify the effectiveness and superiority of the new media event early warning method based on K-means-SO+RF algorithm, the new media event early warning method based on K-means-SO+RF algorithm is compared with K-means+SVM, K-means+BP, K-means+RF, K-means-SO+SVM, and K-means-SO+ BP algorithms are compared and the performance results of each model are shown in Figures 12, 13.

The accuracy of new media event warning methods based on each algorithm is given in Figure 12 respectively. From Fig. 12, it can be seen that the mean value of the accuracy of each algorithm increases with the increase of the number of samples; the accuracy of the new media event warning method based on the K-means-SO+RF algorithm is the largest and the standard deviation is the smallest for different number of samples; the new media event warning method based on the K-means-SO+RF algorithm is compared with the K-means-SO+SVM, K-means-SO+BP algorithm, which shows that the RF classification model is better than SVM and BP model and has better robustness; the new media event warning method based on K-means-SO+RF algorithm is compared with K-means+RF, which shows that SO algorithm improves the effect of K-means clustering. In summary, IK-means-User-R-2 algorithm has better accuracy and better stability than other algorithms.

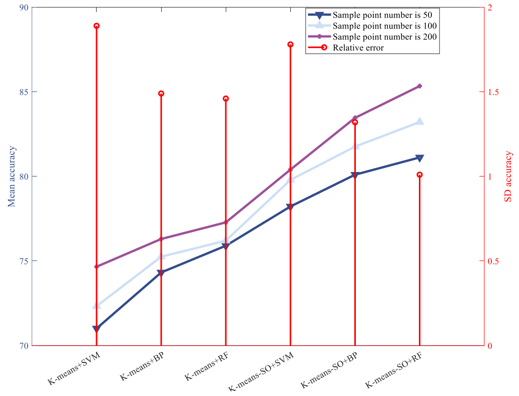


Fig. 12 Comparison results of the accuracy of new media event warning methods based on each algorithm

Figure 13 gives the results of time comparison of new media event warning methods based on each algorithm. From Fig. 13, it can be seen that the new media event warning time based on K-means-SO+RF algorithm is the least; the time of new media event warning method based on K-means-SO+RF algorithm is the smallest with the smallest standard deviation for different number of samples; the time of K-means-SO+SVM, K-means-SO+BP, K-means SO+RF warning time is more than K-means+SVM, K-means+BP, and K-means+RF, respectively, indicating that the optimization process of SO algorithm increases the time overhead. Although the K-means-SO+RF algorithm warning time is not the best, it satisfies real-time.

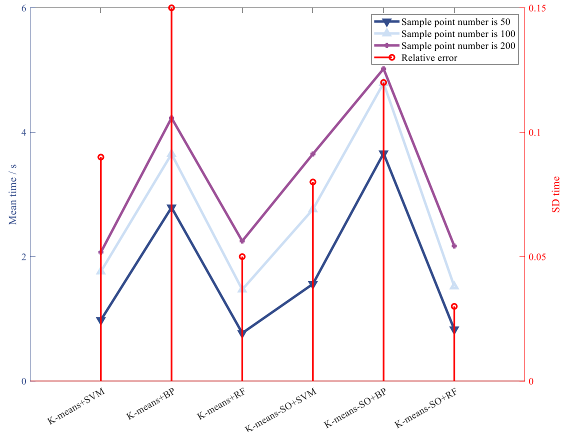


Fig. 13 Results of time comparison of new media event warning methods based on each algorithm

7 Conclusion

Aiming at the problems of incomplete research and inconspicuous model effect of new media event early warning method, this paper proposes a new media event early warning method combining base group intelligent optimization algorithm, K-means clustering algorithm, and random forest algorithm. The method constructs a new media event warning model based on the combination of clustering model and classification model by analyzing the construction process of the new media event warning system, extracting the feature items and processing them in dimensionality reduction, improving the K-means clustering algorithm by using seasonal optimization algorithm, and combining with the random forest algorithm. Using the Haimen event crawler dataset and retrograde simulation experiments, the following conclusions are drawn:

(1) The proposed method is better than other algorithms by comparing K-means-SO+RF with other algorithms warning accuracy;

(2) By comparing the accuracy of K-means-SO+RF with K-means-SO+SVM and K-means-SO+BP algorithms, the accuracy can be improved by using random forest;

(3) By comparing the accuracy of K-means-SO+RF and K-means+RF algorithms, the seasonal optimization algorithm optimizing K-means can improve the accuracy of warning classification;

(4) The recommendation time of K-means-SO+RF algorithm is 0.83, 1.52, and 2.17s under the conditions of 50, 100, and 200 samples, respectively.

There are deficiencies in the public opinion semantic analysis technology in the method proposed in this paper, which is analyzed by simple questionnaire categorization, making the survey sample not comprehensive enough and lacking the semantic value to be given to the science. In the future work, the intelligentization of public opinion semantic analysis technology will be considered to make the effect of the early warning system effectively improved.

References

[1] Ma W , Hu X , Chen C , Wen S, Choo K R, Xiang Y. Social Media Event Prediction using DNN with Feedback Mechanism[J].ACM transactions on management information systems, 2022.

[2] Sobral V , Fairley S , O'Brien D .Factors influencing event media personnel's frame building process at the 2018 FIFA World Cup Russia[J]. .Tourism management, 2022(10):92.

[3] S. S , Thomas M G .Metaheuristic Enabled Hot Event Detection and Product Recommendation in Social Media Data Streams[J]. Communication Networks and Distributed Systems, 2023.

[4] Liu J , Wang Y , He S , Shangguan W, Wang T. A Quantitative Analysis of the Relationship Between the Public and News Media Attentions to Hot Network Events in China[J].International Journal of Crowd Science, 2022, 6(2):53-62.

[5] Li L .Public Opinion Management of Public Crisis Events in the New Media Era - Taking 3.21 China Eastern Airlines Flight Accident as an Example[J].Modern Economics & Management Forum, 2022, 3(2):70-74.

[6] Ma X , Xue P , Li M ,et al. Detection and analysis of emergency topic in social media considering changing roles of stakeholders[J].Online information review, 2023.

[7] Ruozhou W .Research on the Communication Mechanism of Online Rumors Under the Empowerment of New Media Technology-Take the "Nabobess Having an Affair With Courier" Rumor Incident as an Example[J].. Psychological Research: English Edition, 2022(004):012.

[8] Mukherjee P , Badr Y , Jansen B J .Analysis of Formality in Second Screen Postings for Television Broadcast of In-Real-Life Events[J]. Communication and Media Studies, 2022.

[9] Xia L .Historical profile will tell? A deep learning-based multi-level embedding framework for adverse drug event detection and extraction[J]. Decision Support Systems, 2022.

[10] Winkle C V , Corrigan S .Communicating on social media during a #Festival Emergency[J]. 2022, 13(2):144-163.

[11] Afanasiev V . New warning over Russian oil output cuts[J].Upstream: The International Oil & Gas Newspaper, 2023.

[12] Gao J , Xia L I , Xinxing W U ,et al. Analysis on Radar Characteristics of a Short-time Heavy Rainfall Event in Wuchuan County[J]. Meteorology and Environmental Studies:English Edition, 2022, 13(6):35-41.

[13] Coatings F O P .Smart sensing coatings for early warning of degradations[J].Focus on powder coatings, 2023.

[14] Spiegl T , Yoden S , Langematz U , Sato T, Chhin R, Noda S. Modeling the Transport and Deposition of 10Be Produced by the Strongest Solar Proton Event During the Holocene[J].Journal of Geophysical Research: Atmospheres, 2022, 127.

[15] Marvin H J P , Hoenderdaal W , Gavai A K ,Mu W, Van d B L M, Liu N. Global media as an early warning tool for food fraud; an assessment of MedISys-FF[J].Food Control, 2022.

[16] Oshimi D , Yamaguchi S .Leveraging strategies of recurring non-mega sporting events for host community development: a multiple-case study approach[J].Sport, Business and Management: An International Journal, 2022.

[17] Pullen E , Jackson D , Silk M .Paralympic Broadcasting and Social Change: an Integrated Mixed Method Approach to Understanding the Paralympic Audience in the UK:[J].Television & New Media, 2022, 23(4):368-388.

[18] Zhang C , Lei Y , Xiao X , Chen X. Cross-media video event mining based on attention graph structure learning[J].Neurocomputing, 2022.

[19] Koch T K , Frischlich L , Lermer E .Effects of fact-checking warning labels and social endorsement cues on climate change fake news credibility and engagement on social media[J].Journal of Applied Social Psychology, 2023.

[20] Ruozhou W .Research on the Communication Mechanism of Online Rumors Under the Empowerment of New Media Technology-Take the "Nabobess Having an Affair With Courier" Rumor Incident as an Example[J].Psychology Research, 2022, 12(4):208-217.

[21] Hong-Xiao F , Chong J , Li-Juan X U .User Profile Based on Dendriform Vector Space Model[J].

[22] Xue G , Liu S , Ren L , Ma Y, Gong D. Forecasting the subway passenger flow under event occurrences with multivariate disturbances[J].Expert Systems with Applications, 2022, 188:116057-.

[23] Papaioannou A , Vainio R , Raukunen O ,Jiggens P, Aran A, Dierckxsens M. The probabilistic solar particle event forecasting (PROSPER) model[J]. Journal of Space Weather and Space Climate, 2022, 12:24.

[24] Xu S , Li S , Huang W , Wen R. Detecting spatiotemporal traffic events using geosocial media data[J].Computers, environment and urban systems, 2022(1):94.

[25] Jiachao Chen, Min Lu, Weijian Ding, Zhihui Chen. Collaborative clustering recommendation algorithm based on weighted Slope One padding[J]. Modern Information Technology, 2023(022):007.

[26] Emami H. Seasons optimization algorithm[J]. Engineering with Computers, 2022, 38(2): 1845-1865.

[27] MENG Qingsen, HAN Hao, LI Yi. A study on highway secondary accident prediction based on Bayesian optimized random forest[J]. China Safety Production Science and Technology, 2023, 19(7):205-210.