International Journal of Database Theory and Application
Vol.9, No.12 (2016), pp.121-130
http://dx.doi.org/10.14257/ijdta.2016.9.12.12

A Text Clustering Algorithm based on Weeds and Differential Optimization

Lipeng YANG[1], Fuzhang WANG[1] and Chunmei FAN[2]

([1]. China Academy of Railway Science,Bejing 100081 China;
[2]. China Rails Travel Technology, Co. Ltd.,Beijing 100081,China)
yanglipeng@rails.cn

Abstract

Invasive weed optimization (IWO) is a swarm optimization algorithm with both explorative and exploitive power where the diverisity of the population is obtained by allowing the reproduction and mutation of individuals with poor fitness .Differential optimization algorithm is a random parallel algorithm according to a vector change that can make individuals change toward outstanding individuals with global convergence.For k-means algorithm , the traditional algorirhm is prone to get stuck at local optimum and is sensitive to random initialization. Based on the aforementiond background a novel optimization algorithm based hybriding DE and IWO which denoted IWODE-KM is employed to optimize the parameters of k-means and is further applied to chinese text clustering. Experiment results shows that the proposed method outperforms both of its ancestors.

Keywords*: Invasive Weed Optimization; Differential Evolution optimization; K-MEANS ; text clustering*

1. Introduction

With the rapid growth of the information,the document clustering technology is an important research topic of text mining where a large amount of documentations need to be classified into different categories, with each of which containing similar documentations.Text clustering methods mainly include division-based method [1] [2], level-based approach [1] [3].Among those division-based methods, K-MEANS is easy to implement and could benefit from fast convergence, low complexity, so it is chosen as the document clustering algorithm in this paper. However, K-MEANS is sensitive to the initial cluster centers, followed with an iterative loop to update fitness value until certain stopping condition is met, hence, this algorithm might fall into local minima.Recently people are more interested in bio-inspired algorithms such as Weed evolution, Differential evolution algorithm. AR Mehrabian etal first proposed Invasive Weed Optimization algorithm (IWO)[4]. The algorithm gives the opportunity of infeasible seed breeding in the process of evolution , preserves the diversity of population, but the convergence is slow.While, convergence speed of Differential evolution algorithm (DE) [5] is rapid.

In order to solve the weakness of the K-means being sensitive to initial solution selection, we can use the combination of IWO with DE to optimize K - means algorithm by choosing a combination way which is suitable for clustering. By analyzing the combination way of [6], we can discovery that if each iteration runs IWO and DE then two mutation shappen in each iteration equivalently.The change is huge, which is not conducive to reservation of good individual.As mentioned in[7], if it runs IWO firstlyand then runs DE, since IWO searches optimization from global to local, it shows that the solution is similarwithin each local scope after running IWO.Using DE at this time will upset the distribution, so the convergence speed may be slower than using IWO algorithm directly.Because of above shortcoming,this paper proposes a new way of combination,

ISSN: 2005-4270 IJDTA

first run the IWO algorithm,secondly let all part of the individuals of IWO run DE.

2. Related Work

2.1. IWO Algorithm

2.1.1. Invasive Weed Optimization

Invasive weed optimization(IWO)[4] is an evolutionary optimization algorithm inspired from invasive and robust nature of weed sin growth and colonizing,which has the advantages of strong robustness, convergence ,simple structure and easy to implement.General processes of weed invasion are adapting to the environment, taking residence, seeding propagation, raising populations, adjusting to changing circumstances, progressive density, survival of the fittest, competition and death, during which the individual of good fitness will have more chances for survival.In some certain conditon seed begin to breed.When the number reaches to a certain extent the weeds can become more localized and be improved through natural selection and survival competition. There are a variety of ways for plants' natural evolution, among which r- and k-selections are two important ones. The r-selection corresponds to the global exploration way of IWO algorithm, while the k-selection corresponds to the local search way of IWO algorithm.

2.1.2 IWO Algorithm Process

IWO algorithm execution process[8] includes population initialization, reproduction, spatial diffusion and competitive exclusion.

(1) Initialize a population

A finite number of seeds are being dispread randomly on the d-dimensional search area as the initial solutions.

(2) reproduction

A member of the population of plants is allowed to produce seeds depending on its own and the colony's lowest and highest fitness: the number of seeds that each plant produce increases linearly.The number of seed produced by each strain of weed is determined by the following equation. (2.1)

$$w_n = \frac{f - f_{min}}{f_{max} - f_{min}} \left(S_{max} - S_{min} \right) + S_{min} \quad (2.1)$$

Among which, the W_n: number of seeds produced by per strain of weed;f: fitness value of weed; f_{min}: minimum fitness value of weed; f_{max}: maximum fitness value of weed;S_{max}: constant, maximum number of seeds produced by each strain of weed;S_{min}: constant, minimum number of seeds produced by per strain of weed.

(3) Spatial diffusion

The generated seeds are being randomly distributed over the D-dimensional search space by normally distributed random numbers with zero mean and variancer σ^2.However, the standard deviation σ is made to decrease over the generations so that the algorithm gradually moves from exploration to exploitation with increasing generations[9].If the σ_{init} and σ_{final} respective are initial and final standard deviation, then the standard deviation in particular generation (or iteration) is given by (2.2)

$$\sigma_{cur} = \frac{\left(iter_{max} - iter \right)^n}{\left(iter_{max} \right)^n} \left(\sigma_{init} - \sigma_{final} \right) + \sigma_{final} \quad (2.2)$$

Where n represents the non-linear modulation index, iter is the current iteration number and

$Iter_{max}$ is the maximum number of iterations allowed

(4) Competitive exclusion

If a plant leaves no offspring then it would go extinct, otherwise they would take over the world. Thus, there is a need of competition between plants to limit the maximum number of plants in a population. Initially, the plants reproduce fast and all the produced weeds will be included in the colony, until the number of plants reaches to a maximum value of pop. From then on, only the fitter plants among the existing ones and the reproduced ones are taken in the colony .So in every generation the population size must be less than or equal to the max population. This method is known as competitive exclusion and is the selection procedure of IWO.

2.2. DE Algorithm

Differential Evolution algorithm(DE)is a simple but effective algorithm based stochastic search technique for solving global optimization problems .It can dynamically track the current search situation to adjust their search strategies in order to implement adaptive optimization.DE optimization mechanism is to generate new candidate individuals between different individuals according to the distance and direction information to achieve population evolution.

The basic process of differential evolution algorithm：

(1) Initial population. Initial population of N individuals are genetated randomly.

$$P_i(t) = \left[P_{i,1}(t), P_{i,2}(t), \cdots P_{i,d}(t)\right]$$

where d represent the dimension of vector, $i \in \{1, 2, \ldots n\}$ represent ith individual.

(2) Mutation

For every target vector, in any generation t, a mutant vector V_i is generated. Now there are many mutation policies.The paper use the DE/current-to-best/bin policy as shown (2.3)

$$v_i(t) = P_i(t) + F^*\left(P_{best}(t) - P_i(t)\right) + F^*\left(P_j(t) - P_k(t)\right) \quad (2.3)$$

Where i, j, and k are random and mutually exclusive integers generated in the range [1, N], which should also be different from the trial vector's current index i. F is weight factor for scaling differential vectors and $P_{best}(t)$is the individual vector with best fitness value in the population at generation t.

(3) Crossover

This operation involves binary crossover between the target vector $P_{i,j}$ (t)and the mutant vector $V_{i,j}(t)$produced in the previous step. The crossover operation is done as follows (2.4).

$$u_{i,j}(t) = \begin{cases} v_{i,j}(t), \text{if } \text{rand}(0,1) < CR \text{ or } \text{rand}(j) = j \\ P_{i,j}(t), \text{otherwise} \end{cases} \quad (2.4)$$

where CR is a user-specified crossover constant in the range [0, 1)and j_{rand} is a randomly chosen integer in the range [1, d]to ensure that the trial vector $u_i(t)$ will differ from its corresponding target vector $P_i(t)$ by at least one parameter.

(4) Selection

The fitness value of each trial vector f ($u_i(t)$)is compared to that of its corresponding target vector f ($P_i(t)$) in the current population and the population for the t+1 generation formed as follows(2.5): (for a minimization problem)

$$P_i(t+1) = \begin{cases} u_i(t), if \ \text{f}\left(u_i(t)\right) < f\left(P_i(t)\right), \\ P_i(t), otherwise \end{cases} \quad (2.5)$$

Where f (.)is the objective function.

2.3. Text Technology

(1) Participle

Common methods used in Chinese word segmentation are based on the dictionary of the string matching [10],based on statistical word segmentation method,based on rules. The Chinese lexical analysis system, which is developed by Chinese Academy of Sciences, (Institute of Computing Technology Chinese Lexical Analysis System) uses a layered hidden Markov model, whose main features include the Chinese word segmentation, part of speech tagging, named entity recognition, new word recognition and has a very high accuracy of word segmentation.

(2)Vector Space Model (VSM)

The VSM is a kind of representation of text proposed by Salton etc and it's applied into information retrieval system [11]. Considering from efficiency and effect, we gain the conclusion that VSM is the best used in large-scale language processing [12]. The VSM includes some concepts about text, feature item, the weight of feature item, the presentation of feature vector and similarity calculation etc. The feature item is the character, word or phrase processed from text. In VSM, the feature vector represents text, the number of feature items is the dimension of feature vector. The similarity is calculated to measure the distance among texts. Now,VSM is still the most popular tool of text representation.

(3)The Representation of Feature Weight TF-IDF

TF-IDF is also employed to express the weight of feature vector namely of words selected by the aforementioned preprocessing steps, which could be calculated by (2.6),(2.7),(2.8)。

$$tfidf_{i,j} = tf_{i,j} \times idf_i \quad (2.6)$$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (2.7)$$

In the above formula, $n_{i,j}$ is the word occurrences in the document, and the denominator is the sum of occurrences of all words in a document.

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}| + 1} \quad (2.8)$$

|D|: the total number of files in the corpus, the denominator means the number of files containing the corresponding word.

The cosine function is used as the measure of distance between two distinct documents,which could be calculated by(2.9)

$$\sin(d_1, d_2) = \frac{\vec{d_1} * \vec{d_2}}{\left\|\vec{d_1}\right\| * \left\|\vec{d_2}\right\|} \quad (2.9)$$

3. IWODE-KM Text Clustering Algorithm

3.1. Algorithm Description

Web document clustering problem can be described as follows. Let $X_{nxd} = \{X_1, X_2, \ldots, X_n\}$ be a collection of documents consisting of n articles, where X_i is a d-dimensional file vector (d is determined by the numbers of feature words from the corpus), the Web document clustering problem can be described as seeking a partition on the set X denoted $C = \{C_1, C_2, \ldots, C_k\}$which could minimize a clustering criterion function denotes $f(X_{nxd}, C)$, satisfy the following conditions:

(1) $C_i \neq \varnothing, \forall i \in \{1, 2, \cdots, k\};$

(2) $C_i \cap C_j = \varnothing, \forall i \neq j \text{ and } i, j \in \{1, 2, \cdots, k\};$

(3) $\bigcup_{i=1}^{k} C_i = X$

Where the number of the members of C_i is N_i.IWODE-KM first initialize IWO, reproduction, Spatial diffusion,secondly run k-means for every seed,compute the fitness and sort the seeds according to the degree of adaptation,finally remove outside of the largest population of individuals .The selected row back 20% of the individuals from the individual retained, mutate, cross, do k-means partition with the new cluster center, select according to fitness, then new individuals will replace the original ones which are at the back of the population, repeat the above process until meet the maximum of iterations.

3.2. Fitness Function

In order to better reflect clusters effect,we choosethe ratio of cluster separation and tightness in the cluster, use CS measure [13] as a clustering criterion function, such as (3.1) shown.

$$CS(p_k) = \frac{\sum_{i=1}^{k}\left[\max_{j \in k, j \neq i}\left\{\cos\left(m_i, m_j\right)\right\}\right]}{\sum_{i=1}^{k}\left[\frac{1}{n_i}\sum_{x_j \in C_i}\cos\left(x_j, m_i\right)\right]} \quad (3.1)$$

Where cos (m_i, m_j) = $m_i \bullet m_j$ / ($\| m_i \|$ $\| m_j \|$) represents the cosine of the angle between two vectors of documents, the bigger m_i and m_j cosine, the greater the degree of similarity.CS- measure can be regarded as the ratio of cluster separation and clustertightness, the smaller the CS measure value, indicating better clustering effect, therefore, IWODE-Kmeans uses (3.1) as a fitness function to find the global minimum function.

3.3. Coding Scheme

The ith cluster C_i is regarded as an object, in which the members of the cluster and centers being properties of the object as shown Figure 1.As a consequence, each individual in the population will be composed of an array of objects, with each element of which being a collection of information of a cluster.

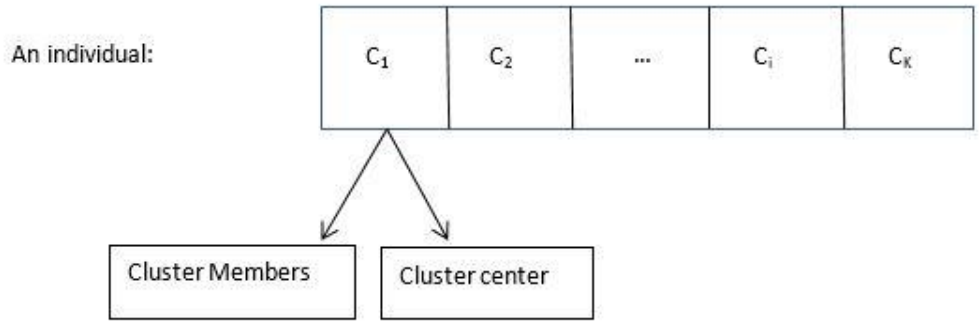

Figure 1

3.4. IWODE-KM Algorithmic Process

The parameters of input and output are shown in Table 1

Table 2

Input	initial population number N_0	The maximum population size N_{max}	the maximum number of iterations IWO $iter_{max}$,	the number of seeds generated limit S_{max} and S_{min}
	σ_{init} and σ_{final} which are the standard deviation of the initial and final values	the initial search space X	the number of clusters K the nonlinear index n	crossover rate CR Scaling factor F
output	Cluster K	Cluster members	Cluster center	

Step1：An initial individual is generated by randomly selecting K samples from all the samples, each of which being the center of a specific cluster with coding scheme shown in Figure1. Repeat the above steps N_0 times to obtain initial population for all individuals.

Step2：breeding individuals,according to (2.1), to obtain w_n and clone the parent for each seed produced.

Step3：spatial diffusion: new seeds are updated according to (2.2) .Then run K-MEANS algorithm with information encode in individuals to update C, then according to the (3.1) calculate its fitness value.

Step4：individual exclusion: To determine whether the population has reached the maximum size, if not, jump to Step2.Otherwise,sort the individuals in ascending order of fitness,only the first N_{max} individuals in the sorted result are left.

Step5： Individuals which rows behind 20% are regarded as the initial population of DE, mutation according to formula (2.3) , crossover according to formula(2.4).

Step6：use k-means division according to the current cluster centers, calculate the fitness of every individual according to formula (3.1), select outstanding individuals from the old and new population according to the formula (2.5), then replace the original individual of the IWO.

Step7：determine whether the maximum number of iterations is reached, and if so, the algorithm ends,the current best individual is recorded as the optimal solution, otherwise return Step2.

4. Experimental

4.1. Text Processing

For the training data of this paper, we use the two corpus which are provided by Fudan University and Sogou.we select history, transportation, medical care, sports, arts from Fudan University and health, sports and education from Sogou corpus. Fudan corpus can be obtained by http://www.nlpir.org/download/tc-corpus-answer.rar. Fudan corpus is provided by the group of international database center natural language processing in computer information and technology department of Fudan University, so the corpus have a certain authority. Sogou corpus is provided by Sogou laboratory.

For the processing of the document, we read the document and record the belonging label of each article. Because ICTCLAS of Chinese Academy of Sciences has a very good segmentation effect, so we choose the software for word segmentation. Afterwards We will do some processing for the segmented words, such as removing the spaces, filtering the according to the stoplist, removing the stop words, removing spaces through regular expressions, punctuation, and meaningless words, choosing the words whose frequencies are in a certain range. Finally, we get very clean data. We use the current most popular VSM to represent text, use TF-IDF to represent the weight of the feature words, and use the cosine function formula (2.9) to measure the distance of the text.

4.2. Results Evaluation Methods

The template Evaluation results using three criteria:

(1)The aforementioned CS measure is employed as one of the evaluation criteria in this paper.

(2)In this paper,a common externalevaluation method F-measure[14] is also adopted,which combining of precision and recall. A cluster j, associated with its real class label i (the original one class) .with precision p and recall r is defined as:

$$p\left(i,j\right) = \frac{N_{i,j}}{N_i} \quad (4.1)$$

$$r(i,j) = \frac{N_{i,j}}{N_j} \quad (4.2)$$

Where $N_{i,j}$ is the number of correct classificaions of i incluster j, N_i denotes the number of all objects in the ith clust-er N_j is the number of objects belonging to original class j. The F-measure of the ith original class is defined as.

$$F\left(i\right) = \frac{2pr}{p+r} \quad (4.3)$$

The i may have multiple corresponding clusters, only the cluster with highest F measure is chosen as the correct one. The total average F-measure values is defined as weighting the F-measure of each category:

$$F_k = \frac{\sum_i \left(N_i \cdot F\left(i\right)\right)}{\sum_i N_i} \quad (4.4)$$

4.3. Experimental Setup and the Results Analysis

All experiments are run on Window 7 system and computer memory is 4G. The algorithms used in our experiments contain the k-means algorithm improved by using DE (DE-KM) and k-means algorithm improved by using IWO(IWO-KM). Both DE-KM and IWO-KM are implemented using java and the integrated development environment is MyElipse10.7. In order to verify the optimization effect of the combination of the measures[7,15] on k-means, we use java to implement the optimization of k-means,respectively IWODE-KM1, IWODE-KM2.

In our experiments,50 documents, 100 documents and 200 documents from Fudan and 90 documents and 180 documents from sougou are selected.X(test50), X(test100), X (test200), X(test90) and X(test180) respectively represent the selected data.The corpus of Fudan University is set into a uniform distribution, and the distribution of the sogou corpus is set not uniform.The experimental parameters are set as follows: Scaling factor F=0.5,CR=0.1,N_{max}=40,N_0=30,S_{max}=5,S_{min}=0,σ_{init} =5,σ_{final}=0.02,n=3.Then, the experimental result is output to a file, including cluster name, cluster member, nearest

member of the cluster center, and fitness.

By comparing the results of cluster with the results of the original classification, we can get the following experimental results. When juding documentsas whichoriginal classification, we take the nearest member to the cluster center as the division of the category as far as possible. However, if the category contains too many menbers of other categories, we should select other corresponding category. The result of CS measure is obtained by using formula 3.1, and the result of F–measure is get by formula 4.4. Because the optimization of the K-means algorithm has a certain randomness, each data set runs 30 times, and then we take the average value as the results. The results are shown as follows: CS-measure's results are shown in Figure 5,the accuracy of the results is shown in Figure 6, F-measure results are shown in Table 2, figure 7.

Table 2. F-measure

F-measure	test50	test100	test200	test90	test180
k-means	0.84	0.57	0.54	0.57	0.6
DE-KM	0.72	0.6	0.58	0.57	0.62
IWO-KM	0.78	0.72	0.69	0.8	0.63
IWODE-KM1	0.75	0.69	0.66	0.73	0.53
IWODE-KM2	0.7	0.67	0.58	0.44	0.39
IWODE-KM	0.87	0.84	0.79	0.82	0.68

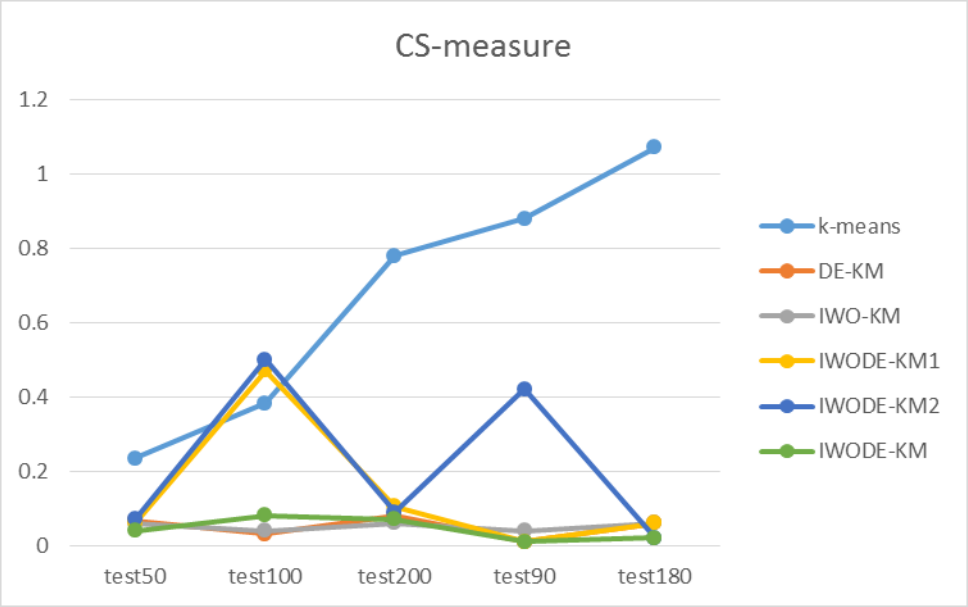

Figure 5

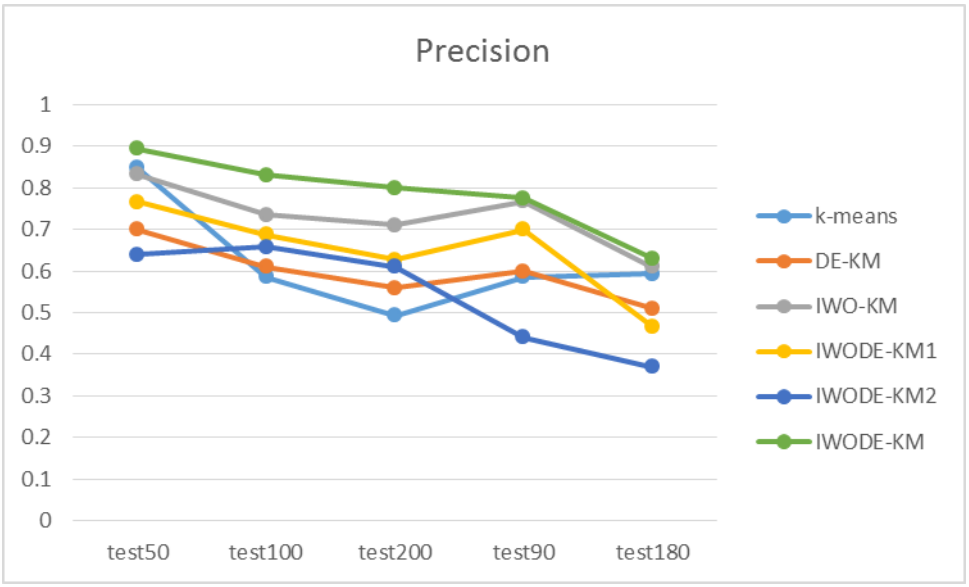

Figure 6

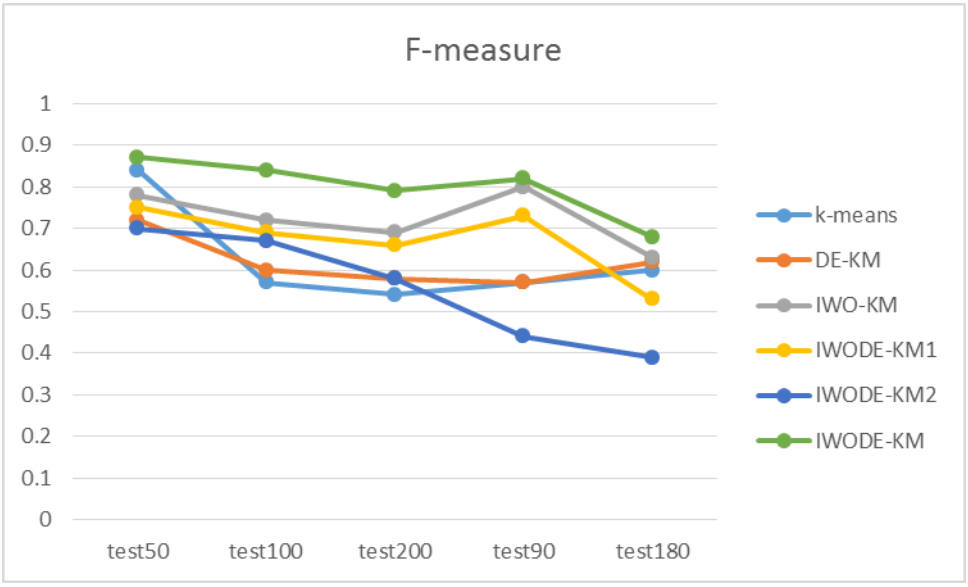

Figure 7

From the analysis results of the Figure 5 we can see that CS-measure of k-means algorithm increase along with the increase of samples. It stays almost the same level for the CS-measure results of DE, IWO, IWODE to optimize k-means algorithm. As can be seen in Figures 6 and 7, accuracy and F value of the IWODE-KM algorithm is significantly better than that of the IWO-KM algorithm and DE-KM algorithm for the equilibrium and non-equilibrium data set. However the IWODE1 and IWODE2 added DE algorithm on the basis of the optimized IWO, and the result is not as good as k-means results optimized only with IWO algorithm. It means that IWODE-KM1 approach does cause upheaval of cluster centers, and is not conducive to optimization of cluster centers. At the same time, IWODE-KM2 will use DE when the IWO is about to converge, and it also destroys the original convergence so that the result is lower than IWO-KM. Of cource The results of these optimization algorithms are superior to that of the k-means

algorithm.In the course of the experiment, the number of iterations are shown as table 3.We can see obvious improvement of convergence speed of IWODE-KM algorithm, so not only the accuracy of IWODE-KM optimization methods has been improved, but also the convergence speed is greatly improved.

Table 3

iterations	Test(50)	Test（90）	Test（100）	Test（180）	Test（200）
IWO-KM	30	45	45	45	50
IWODE-KM	10	25	25	25	25

5. Conclusion

This paper takes the advantages of the diversity of the DE and rapid convergence of IWO to improve the K-means algorithm with low complexity. The algorithm is proposed that text clusting algorithm of using DE to optimize part individuals of IWO,and firstly apply it to the Chinese text clustering. Using the model described in this article to pretreat the text corpus, Using the ratio between the ratio of cluster separation and tightness in the cluster as the evaluation function, different data experiments have been done. The results show that IWODE-KM algorithm can improve convergence speed greatly and can perform better on the experimental data than all the other methods mentioned in this paper

References

[1] L. Wang, “Phrase message clustering related technology research”, National University of Defense Technology, **(2008)**.
[2] J. Liu, J. P. Y. Lee, L. Li, Z. Q. Luo and K. M. Wong, “Online Clustering Algorithms for Radar Emitter Classification”, IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, **(2005)**, pp. 1185-1196.
[3] H. V. D. Parunak, R. Rohwer, T. C. Belding and S. Brueckner, “Dynamic Decentralized Any-Time Hierarchical Clustering”, Engineering Self-Organizing Systems. Heidelberg: Springer Press, Berlin, **(2007)**, pp. 66-81.
[4] A. R. Mehrabian and C. Lucas, “A novel numerical optimization algorithm inspired from weed colonization”, Ecological Informatics, vol. 1, no. 4, **(2006)**, pp. 355-366
[5] C. Xinye, H. Z. Zhou and F. Zhun, “A novel memetic algorithm based on invasive weed optimization and differential evolution for constrained optimization”, Soft Computing, vol. 17, no. 10, **(2013)**, pp. 1893-1910.
[6] L. Xiaoyan, J. Renquan, L. Shaobo and X. Kun, “Prediction model of stock price based on differential evolution invasive weed optimization algorithm”, Journal of University of Science & Technology Liaoning, **(2014)**.
[7] H. Chen, Y. Zhou and W. Zhao, “Multi-population invasive weed optimization algorithm based on chaotic sequence”, Journal of Computer Applications, vol. 32, no. 6, **(2012)**, pp. 1958-1961.
[8] Y. Han, J. Cai and L. Li, “Invasive Weed Optimization and its Advances”, Computer Science, vol. 38, no. 3, **(2011)**, pp. 20-23.
[9] W. Ni, “An Effective Distributed k-Means Clustering Algorithm Based on the Pretreatment of Vectors' Inner-Product”, Journal of Computer Research and Development, vol. 42, no. 9, **(2005)**, pp. 1493-1497.
[10] G. Salton, A. Wong and C. S. Yang, “A vector space model for automatic indexing”, Commun. ACM, vol. 18, no. 11, **(1975)**, pp. 613-620
[11] W. M. Shaw Jr., R. Burgin and P. Howell, “Performance standards and evaluations in IR test collections: vector-space and other retrieval models”, Inf. Process. Manage, vol. 33, no. 1, **(1997)**, pp. 15-36.
[12] L. Sen and R. Xiaona, “An Improved Invasive Weed Optimization Algorithm with Differential Evolution Strategy and Application of Function Optimization”, Computer Applications & Software, **(2014)**.
[13] S. Das, A. Abraham and A. Konar, “Automatic Hard Clustering Using Improved Differential Evolution Algorithm”, Metaheuristic Clustering. Springer Berlin Heidelberg, **(2009)**, pp. 137-174.
[14] A. Abraham, S. Das and A. Konar, “Document clustering using differential evolution”, Proceedings of the 2006 IEEE Congress on evolutionary computation, **(2006)**, pp. 1784-1791.
[15] H. Ayad and M. Kamel, “Topic discovery from text using aggregation of different clustering methods”, CohenR, SpencerBed. Advances in artificial intelligence: 15th conference of the Canadian society for computational studies of intelligence. Calgary, **(2002)**, pp. 161-175

