TELKOMNIKA, Vol. 13, No. 4, December 2015, pp. 1384~1389
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v13i4.xxxx

Text Mining Research Based on Intelligent Computing in Information Retrieval System

Yong Li
Chong Qing Technology and Business Institute, Chong Qing City, 400052, China
email: liyong0417151@126.com

Abstract

With the popularity and rapid development of the Internet, web text information has rapidly grown as well. To address the key problem of text mining, text clustering is investigated in this study. The shuffled frog leaping algorithm as a new type of swarm intelligence optimization algorithm can be used to improve the performance of the K-means algorithm, but the shuffled frog leaping algorithm is influenced by its moving step length. On the basis of this information, the shuffled frog leaping algorithm is improved, and the K-means clustering algorithm based on the improved shuffled frog leaping algorithm is introduced. Experiment results show that the proposed scheme can enhance the ability of searching for the optimal initial clustering center and can effectively avoid instability in the clustering results of the K-means clustering algorithm. The proposed scheme also reduces the chances of the algorithm falling into the local optimum. The performance of the proposed clustering scheme is found to be better than that of the clustering algorithm based on the shuffled frog leaping algorithm.

Keywords*: Text Mining, Shuffled Frog Leaping Algorithm, Clustering Accuracy*

Copyright © 2015 Universitas Ahmad Dahlan. All rights reserved.

1. Introduction

With the rapid growth of web information, people urgently need a type of technology that organizes and manages information to help them find what they need quickly and accurately; data mining combing with web mining has emerged as a response to this issue [1]. Text is the main form of information on the web, so text mining has become a research hotspot in recent years. Chinese text mining has been developed relatively late, and it falls behind English text mining in the aspects of theory research and application. Chinese text mining based on the web is therefore selected as our research object. Text classification and clustering are key technologies in text mining. Organizing and classifying text data sets can significantly address the problem of information explosion. Text classification and clustering can also be applied as the technical basis in specific fields, such as information retrieval, search engines, electronic libraries, and text databases [2]. With the advent of the information era, text classification and clustering have become increasingly popular.

Current clustering algorithms are classified into partitioning clustering, hierarchical clustering, grid-based clustering, and density-based clustering. The K-means algorithm, a classical clustering algorithm, is a local search scheme with some serious disadvantages. The K-value needs to be determined in advance, and the clustering result depends on the selection of the initial clustering center [3]. In this regard, many researchers have proposed clustering methods based on the intelligent optimization algorithm [4, 5]. This algorithm is gradually developed with the use of certain similarities between complex systems (e.g. natural or social) and optimization problems. The algorithm obtains the next feasible solutions through operation on a set of initial solutions in the search space according to certain rules of probability. Therefore, the searching mechanism of the algorithm determines its optimization performance [6]. The shuffled frog leaping algorithm is a type of new intelligent optimization algorithm. However, this algorithm also has some disadvantages, such as its poor local search ability and slow convergence speed [7–10]. To improve the convergent performance of the basic shuffled frog leaping algorithm, an improved algorithm is introduced in this study. The clustering method based on intelligent optimization can convert a particular clustering problem to an optimization problem of the objective function, which finds the optimal value of the objective function to obtain the optimal clustering scheme through repeated iteration. The key technologies in using

Received March 26, 2015; Revised September 26, 2015; Accepted October 12, 2015

intelligent computing to solve the optimization problem include coding of the problem and design of an appropriate fitness function [11-14]. In this study, the shuffled leapfrog algorithm is combined with the K-means algorithm. The main advantage of the proposed scheme is that the k-means method can be used to improve the convergence speed, and the global optimal solution can be obtained by means of the shuffled frog leaping algorithm. In particular, the proposed scheme combines the global searching characteristic of the shuffled frog leaping algorithm and the simplicity of the k-means algorithm, and accordingly, an algorithm with global searching capability is obtained [15, 16].

In the next section, the basic shuffled frog leaping algorithm is investigated. In Section 3, a type of improved shuffled frog leaping algorithm is introduced. In Section 4, the clustering algorithm based on improved intelligent computing is proposed. In Section 5, an experiment is conducted to test the performance of the proposed scheme. In the last section, conclusions are provided [17].

2. Basic Shuffled Frog Leaping Algorithm

The function optimization problem is transformed into the minimum value problem of the objective function $f(x)$ in the feasible domain, where x is the solution vector. First, *F* number of points is selected as the initial value from the feasible domain randomly. The i-th frog is represented by $x^j = (x_1^j, x_2^j, \ldots, x_n^j)$, and the objective function value of each frog is calculated. Each frog is ordered decreasingly according to its objective function value, and then the entire frog swarm is divided into *S* number of subgroups that contain m number of frogs.

In the iteration process, the first solution enters into the first subgroup, the second solution enters into the second subgroup, and the rest can be conducted in the same manner. Then, the (*s*+1)-*th* solution enters into the first subgroup again, and the (*s*+2)-*th* solution enters into the second subgroup until all the solutions are assigned completely. In each subgroup, the solution with the best objective function value and that with the worst objective function value are labeled as $x^b = (x_1^b, x_2^b, \ldots, x_n^b)$ and $x^w = (x_1^w, x_2^w, \ldots, x_n^w)$, respectively. The solution with the best objective function in the swarm is labeled as $x^g = (x_1^g, x_2^g, \ldots, x_n^g)$. In each iteration, x^w is updated by formula 1.

$$D^j = r_1 \cdot (x^b - x^w) \tag{1}$$

$$x^{w'} = x^w + D^j. \quad (-D_{\max} \le D^j \le D_{\max}) \tag{2}$$

$r_1 \in U(0,1)$, $j = 1, 2, \ldots, S$, $D_{\max}$ represents the maximum moving step length of the frog. If the objective function value of $x^{w'}$ is better than that of x^w, $x^{w'}$ is replaced by x^w. If the solution is not improved, x^b is replaced by x^g, and formulas 1 and 2 are executed repeatedly. If the solution is still not improved, a new solution from the feasible domain is generated to replace the original x^w. The above operations are performed within a specified number of times, and the location of the worst frog in each group is updated. Accordingly, the first iteration of the shuffled frog leaping algorithm is completed. As observed from formula 1, the size of the moving step directly affects the global convergence of the algorithm. When the size is large, the frog can conduct global searching, but it may skip the optimal solution. When the size is small, the frog can search finely in the local area, but it can easily fall into the local optimum. Therefore, moving the step length affects the optimization performance of the shuffled frog leaping algorithm to a certain extent. [18, 19]

3. Improved Shuffled Frog Leaping Algorithm

For each subgroup, the state of surrounding frogs can affect the behavior of the worst frogs to some extent, so repulsion against the worst frog occurs. The worst frog moves with the guide of information, and the frogs in the subgroup encourage one another to improve

performance through competition and cooperation; in this way, co-evolution of groups can be realized.

The improved shuffled frog leaping algorithm ensures that each frog brings a certain amount of charge, which is decided by the optimized objective function value. Then, the resultant force imposed on the worst frog from the other frogs in the subgroup is calculated to determine the moving step length. In the subgroup, the q^i of frog i is calculated by

$$q^i = \exp\left(-n \frac{f(x^i) - f(x^g)}{\sum_{k=1}^{m} (f(x^k) - f(x^g))} \right) \tag{3}$$

$i = 1, 2, \ldots, m$, $f(x^i)$ represents the current objective function value of frog i in the subgroup, and $f(x^g)$ represents the current optimal objective function value in the swarm. The component force imposed on the worst frog in the subgroup is calculated by

$$F_j^w = q^j q^w (x^j - x^w) \tag{4}$$

After the resultant force F^w is calculated, the force imposed on each frog is normalized to ensure the feasibility of frog shift, and the proposed update strategy is

$$D^j = r_1 \cdot (x^b - x^w) + a^w \tag{5}$$

$$x^{w'} = x^w + D^j. \quad (-D_{\max} \le D^j \le D_{\max}) \tag{6}$$

$a^w = r_2 \cdot F^w / \left\| F^w \right\|_2$, $r_2 \in U(0,1)$, and the proposed scheme increases the diversity of the subgroup to prevent the occurrence of premature convergence; as a result, the entire group can benefit. The process involves in the improved shuffled frog leaping algorithm is as follows:

Step 1. Initialize the swarm and parameters, such as the total number of frogs, *F*; the number of subgroups, *S*; the iteration times within the subgroup, It; and the hybrid iteration times, $N_{\max}$.

Step 2. Calculate the objective function value $f(x^j)$ of each frog.

Step 3. Order the *F* number of objective function values, and divide it into *S* number of subgroups.

Step 4. Determine the individual x^b with the best objective function value and the individual x^w with the worst objective function value in each group. Within the given iteration times, It, update the worst solution according to formulas 5 and 6.

Step 5. For each subgroup, arrange the individuals in descending order according to the objective function value to constitute a new group.

Step 6. Determine whether the termination condition is met; if this condition is satisfied, the related information of the optimal objective function value is the output. Otherwise, return to step 2.

4. Clustering Algorithm Based on the Improved Shuffled Frog Leaping Algorithm

4.1. Principle of the K-Means Algorithm

The input is the data set $X = \{x_1, x_2, \ldots, x_n\}$, and the number of clusterings is *K*. The output is *K* number of clusterings, C_j. The process involved in the K-means algorithm is as follows:

Step1. *K* number of random data is chosen as the initial clustering center $z_1, z_2, \ldots, z_k$ from *X*.

Step2. Calculate the distance between each data point x_i and C_j, $j \in \{1,2,\ldots,k\}$. If the condition of $\|x_i - z_j\| < \|x_i - z_m\|$, $m = 1,2,\ldots,K$, $m \neq j$ is met, then $x_i \in C_j$.

Step3. Calculate the new center point $z_1^*, z_2^*, \ldots, z_k^*$ with

$$z_i^* = \frac{1}{n_i} \sum_{X_j \in C_i} X_j \quad , \ i = 1,2,\ldots,K \tag{7}$$

Step4. If $z_i^* = z_i, i = 1,2,\ldots,K$, the algorithm stops, and the current center point is selected as the result of the clustering partition. Otherwise, return to step 2.

4.2. Clustering Scheme Based on the Improved Shuffled Frog Leaping Algorithm

The shuffled frog leaping algorithm uses the fitness value of each individual in one population to search the optimal value. Therefore, choosing suitable fitness function affects the convergence rate of this algorithm. The fitness function is

$$f = 1 / \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - z_i\|^2 \tag{8}$$

The clustering algorithm based on the improved shuffled frog leaping algorithm is as follows:

Step 1. Set the initial parameters of the algorithm.

Step 2. Initialize the entire swarm and *K* number of initial clustering centers.

Step 3. Calculate the distance between *X* and *K* number of centers corresponding to the frog. *X* is classified according to distance.

Step 4. Calculate the fitness value of each frog according to classification, and arrange the frog swarm in descending order according to the size of the fitness value to generate D^j randomly.

Step 5. Divide the frog swarm into m number of subgroups. Calculate the best solution x^b, the worst solution x^w, and the global optimal solution x^g.

Step 6. Within the given iteration times, It, update the worst solution according to formulas 5 and 6.

Step 7. Each subgroup is combined to form a new frog swarm. Arrange the frog swarm in descending order according to its fitness value, increase the global iteration times by one, and return to step 5. Repeat the above process until the maximum iteration time is achieved.

5. Experiment Results

The University of California, Irvine (UCI) data set is commonly used in the algorithm performance testing of machine learning, information processing, and data mining. The data of this data set are strictly labeled, so this data set is typically used as the evaluating standard of several algorithms. The Iris and Wine data sets of UCI are selected to test the performance of the proposed scheme. The number of subgroups m is 3, the number of frogs in the subgroup n is 20, the iteration time It within the subgroup is 30, and the global search time is 1000. The clustering performance of the proposed scheme and the traditional k-means based on the shuffled frog leaping algorithm is shown in Table 1. This table depicts that the average clustering accuracy of the proposed scheme is higher than that of the k-means algorithm based on the traditional shuffled frog leaping algorithm.

Table 1. Clustering performance comparison of the two algorithms

Data set	algorithm	right	wrong	accuracy
Iris	shuffled frog leaping algorithm	138	12	92.00%
	proposed scheme	141	9	94.00%
Wine	shuffled frog leaping algorithm	146	32	82.02%
	proposed scheme	157	21	88.20%

Table 2. Fitness function value comparison of the two algorithms

Data set	algorithm	fitness function	inner class distance	between-class distance
Iris	shuffled frog leaping algorithm	0.604±0.072	3.493±0.224	0.891±0.037
	proposed scheme	0.409±0.016	3.101±0.109	0.985±0.019
Wine	shuffled frog leaping algorithm	1.088±0.015	4.462±0.217	2.967±0.432
	proposed scheme	1.015±0.015	4.197±0.396	3.651±0.306

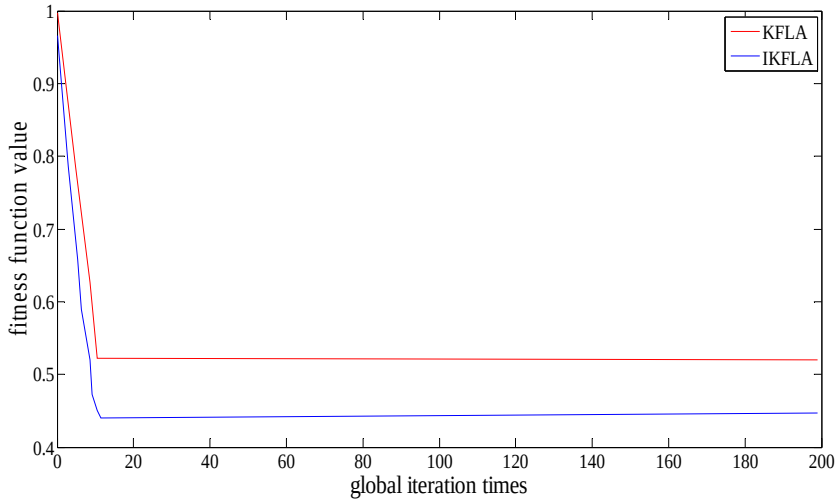

Figure 1. Convergence comparison of two algorithms on Iris dataset

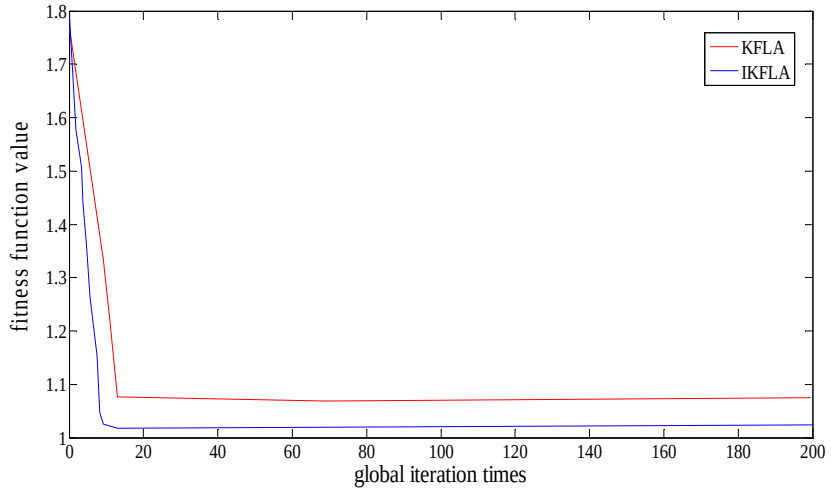

Figure 2. Convergence comparison of two algorithms on Wine dataset

The comparison of the fitness function value is shown in Table 2. The average fitness function value of the proposed scheme is smaller than that of the traditional scheme. The inner-class distance is also smaller, and the between-class distance is larger. The convergence comparison of the two algorithms in the Iris and Wine data sets is shown in Figures 1 and 2, respectively. The k-means algorithm based on the shuffled frog leaping algorithm is affected by the initial value. This scheme can easily fall into the local optimal solution. Compared with the traditional scheme, the proposed clustering algorithm based on the improved shuffled frog leaping algorithm has higher convergence speed and higher clustering accuracy.

6. Conclusion

To solve the key problem of text mining, we conducted a study on text mining based on intelligent computing in an information retrieval system. An improved shuffled frog leaping algorithm was proposed, the detailed process of this algorithm was presented, and a novel K-means clustering method based on the improved shuffled frog leaping algorithm was introduced. The experiment results show that the proposed scheme has higher clustering accuracy and convergence speed than the k-means algorithm based on the shuffled frog leaping algorithm. This study can serve as a useful and meaningful reference for text classification and clustering, particularly in information retrieval and intelligent computing.

References

[1] Samaan M, Rahman F. PTH-200 Amino acid profiles in patient with intestinal failure: preliminary data on biochemical insights: Abstract PTH-200 Figure. *Gut*. 2015; 64(1): A1-A584.
[2] Rozanski MR, Mitchell T. Greetings of Peace. Claritas: *Journal of Dialogue and Culture*. 2015; 4(2): 5.
[3] Wang F, Geng C, Su L. Parameter identification and prediction of Jiles–Atherton model for DC-biased transformer using improved shuffled frog leaping algorithm and least square support vector machine. *IET Electric Power Applications.* 2015; 9(9): 660-669.
[4] Wang T, Zhao X, Zhou Y. Community detection in social network using shuffled frog-leaping optimisation. *International Journal of Security and Networks.* 2015; 10(4): 222-227.
[5] Lei D, Guo X. A shuffled frog-leaping algorithm for hybrid flow shop scheduling with two agents. *Expert Systems with Applications*. 2015; 42(23): 9333-9339.
[6] Pasha MFK, Yeasmin D, Rentch JW. Dam-lake Operation to Optimize Fish Habitat. *Environmental Processes*. 2015; 2(4): 631-645.
[7] Gómez-González M, Ruiz-Rodriguez FJ, Jurado F. Metaheuristic and probabilistic techniques for optimal allocation and size of biomass distributed generation in unbalanced radial systems. *IET Renewable Power Generation*. 2015; 9(6): 653-659.
[8] Geem ZW. Multiobjective optimization of water distribution networks using fuzzy theory and harmony search. *Water.* 2015; 7(7): 3613-3625.
[9] Youa Z, Caoa X, Wanga Y. An Unequal Clustering Strategy for WSNs Based Urban Intelligent Transportation System. *Journal of Information and Computational Science*. 2015; 12(10): 4001-4012.
[10] Sedighizadeh M. Optimal Reconfiguration and Capacitor Allocation in Radial Distribution Systems Using the Hybrid Shuffled Frog Leaping Algorithm in the Fuzzy Framework. *Journal of Operation and Automation in Power Engineering*. 2015; 3(1): 56-70.
[11] Zhao J, Lv L. Two-Phases Learning Shuffled Frog Leaping Algorithm. *International Journal of Hybrid Information Technology*. 2015; 8(5): 195-206.
[12] Li Y, Hu L. A fast exact simulation method for a class of Markov jump processes. *The Journal of Chemical Physics.* 2015; 143(18): 184105.
[13] Wang F, Geng C, Su L. Parameter identification and prediction of Jiles–Atherton model for DC-biased transformer using improved shuffled frog leaping algorithm and least square support vector machine. *IET Electric Power Applications.* 2015; 9(9): 660-669.
[14] Elbeltagi E, Hegazy T, Grierson D. A modified shuffled frog-leaping optimization algorithm: applications to project management. *Structure and Infrastructure Engineering*. 2007; 3(1): 53-60.
[15] Tang Z. Improved K-means Clustering Algorithm based on Genetic Algorithm. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2014; 12(3):1917-1923.
[16] Mohan RNVJ, Rao KRS. Efficient K-Means Fuzzy Cluster Reliability on Angle Oriented Face Recognition. *International Journal of Informatics and Communication Technolog*y, 2012, 2(1):38-45.
[17] Niknam T, Farsani EA. A hybrid self-adaptive particle swarm optimization and modified shuffled frog leaping algorithm for distribution feeder reconfiguration. *Engineering Applications of Artificial Intelligence*. 2010; 23(8): 1340-1349.
[18] Niknam T, Nayeripour M. An efficient multi-objective modified shuffled frog leaping algorithm for distribution feeder reconfiguration problem. *European Transactions on Electrical Power.* 2011; 21(1): 721-739.
[19] Eusuff MM, Lansey KE. Optimization of water distribution network design using the shuffled frog leaping algorithm. *Journal of Water Resources Planning and Management*. 2003; 129(3): 210-225.