Xiongjun Wen,
Qun Zhou,
Sheng Huang

Hunan International Economics University, Changsha, Hunan, China

A DIFFERENTIAL CLUSTERING ALGORITHM BASED ON ELITE STRATEGY

Сюнцзюнь Вень,
Цюнь Чжоу,
Шен Хуан

Хунаньский університет міжнародної економіки, г. Чанша, Хунань, КНР

АЛГОРИТМ ДИФФЕРЕНЦІАЛЬНОЇ КЛАСТЕРИЗАЦІЇ НА ОСНОВІ ЕЛІТАРНОЇ СТРАТЕГІЇ

Purpose. Cluster analysis is not only an important research field of data mining but also a significant means and method in data partitioning or packet processing. The research aims to further improve the effect of clustering algorithm and overcome the existing defects of differential evolution (DE). The research achievements are intended to be used in the cluster analysis to obtain better clustering effect.

Methodology. We have made in-depth research with regards to DE algorithm and cluster analysis, discussed the effect of K-means as well as the flowchart and computing method of the fitness function. The influence of different differential operations plays on the performance have been analyzed.

Findings. Firstly, we explained the fundamental ideas and methods of cluster analysis and DE algorithm. Then we illustrated how the improved DE algorithm realizes the cluster analysis. Finally, we conducted simulation experiment of cluster analysis on four artificial data through the clustering algorithm based on elite strategy DE algorithm so that we can verify the feasibility and validity of the new method.

Originality. We developed an elite strategy DE algorithm and used it in K-means cluster analysis. Since DE algorithm is a method to search the optimal solution by simulating natural evolution process, its outstanding features are its implicit parallelism and ability to utilize the global information effectively, therefore, the new and improved algorithm has stronger robustness and it can avoid getting trapped in a local optimum and greatly enhances the clustering effect. The research on this aspect has not been found at present.

Practical value. By applying elite strategy DE algorithm in K-means cluster analysis, we can improve the efficiency and accuracy of cluster analysis. The result of the simulation experiment showed that the new method presented in this paper has significantly improved the optimization performance, which verifies the feasibility and effectiveness of this new method.

Keywords: *cluster analysis, k-means, differential evolution, elite strategy, optimization performance, feasibility, effectiveness*

Introduction. Together with the rapid development of the information technology in recent years, there have emerged plenty of data, from which the existing database technology fails to extract deeper and valuable information, let alone to make an ideal decision and forecast the future development trend based on the existing data. As one of the core techniques in data mining, the cluster analysis method has become an active research topic in this field. As an unsupervised learning process, the cluster analysis method clusters the things into different classes according to certain properties, minimizes the between-class similarity and maximizes the inter-class similarity. People have paid increasing attention to the improvements and research of its relevant algorithms and clustering has been widely researched in such fields as statistics, machine learning, pattern recognition and data mining [1].

The traditional cluster analysis methods include system decomposition method, addition method, overlapping clustering and fuzzy clustering. Restricted to the fields of statistics and machine learning, most of these methods fail to take large data volume and operation cost into full consideration, making the original algorithm null and void or unable to be used in data mining. Therefore, they are not applicable to the circumstances with large data volume. People's eyes have been drawn to how to utilize and improve the traditional cluster algorithms so that they can help to find the useful information from large databases. In the future, the research on the application field of data mining will become increasingly extensive and the application of artificial intelligent algorithms will become more and more population in cluster analysis [2]. DE algorithm is a newly-emerging evolutionary computation technique raised by R. Stom and K. Price in 1995 and it is a fast evolution algorithm. For most Benchmark problems, DE algorithm has excelled itself in convergence speed and stability. Since DE is easy to understand and simple to realize, it has been widely used [3]. Therefore, to introduce the evolution algorithm with global optimization ability into cluster application field is one of the future development trends of cluster analysis. This paper discusses the clustering problems as well as the structural scheme and computing methods of the fitness function and makes research on improving the accuracy of cluster analysis by applying DE in the clustering.

Improvement strategy of differential evolution algorithm. ***Improvement of the control parameter.*** In DE operation process, there are mainly 3 parameters needing to set up, i. e. population size *NP*, mutation factor *F* and crossover factor *CR* respectively. The selection of these three parameters has a great influence on the algorithm's global optimization ability and convergence speed.

l. Population size *NP*: generally, the greater the population size *NP* is, the more numerous the individuals are, and the better the population diversity is, and the stronger the optimization ability is. However, in such case the difficulty of calculation also increases so *NP* cannot be infinitely large. In general, choose four to ten times of the solution space dimension *D* to set up the population size, which allow both, avoiding the search falling into local optimum, and quick converge to the global optimal solution.

2. Mutation factor *F*: mutation factor *F* is a real constant in the range of [0, 2]. The *F* setting affects the convergence and the diversity of the population. Too small mutation factor may cause the algorithm prematurity. The increase of *F* value improves the algorithm's ability to avoid trapping into the local optimum. However, when $F > 1$, it becomes difficult to make the algorithm converge to the optimal value rapidly, because the population convergence will become very poor when the perturbation of differential vector is larger than the distance between two individuals. Therefore, the mutation factor value is usually set in the interval from 0.5 to 0.9.

3. Crossover factor *CR* is a real constant in the range of [0, 1]. The larger the crossover factor CR is, the greater the likelihood of the crossover, and usually, the crossover factor value is selected within the range from 0.3 to 0.9. When *CR* setting is too small, the convergence speed accelerates and the algorithm easily falls into local optimum, but when *CR* setting is too big, the stability of the algorithm is reduced.

In the standard DE algorithm, three main parameters are usually fixed in the whole looping execution. This makes the DE algorithm simple compared with other evolution algorithms. Nevertheless, because parameters are usually selected appropriately based on the experience, for different optimization problems, there exist very large differences in terms of the parameter settings.

Improvement of DE mutation strategy. For the standard DE algorithm, by improving the mutation operation mode of DE algorithm, people put forward many modes. Some main modes are listed as follows:

Mode 1: DE/RAND/1 (standard DE algorithm)

$$\overrightarrow{v_{i,G}} = \overrightarrow{x_{r1,G}} + F * (\overrightarrow{x_{r2,G}} - \overrightarrow{x_{r3,G}}).$$

Mode 2: DE/BEST/1

$$\overrightarrow{v_{i,G}} = \overrightarrow{x_{best,G}} + F * (\overrightarrow{x_{r2,G}} - \overrightarrow{x_{r3,G}}).$$

Mode 3: DE/RAND to BEST/1

$$\overrightarrow{v_{i,G}} = \overrightarrow{x_{i,G}} + F * (\overrightarrow{x_{best,G}} - \overrightarrow{x_{i,G}}) + F * (\overrightarrow{x_{r1,G}} - \overrightarrow{x_{r2,G}}).$$

Mode 4: DE/BEST/2

$$\overrightarrow{v_{i,G}} = \overrightarrow{x_{best,G}} + F * (\overrightarrow{x_{r1,G}} + \overrightarrow{x_{r2,G}} - \overrightarrow{x_{r3,G}} - \overrightarrow{x_{r4,G}}).$$

Mode 5: DE/RAND/2

$$\overrightarrow{v_{i,G}} = \overrightarrow{x_{r5,G}} + F * (\overrightarrow{x_{r1,G}} + \overrightarrow{x_{r2,G}} - \overrightarrow{x_{r3,G}} - \overrightarrow{x_{r4,G}}).$$

Among the foregoing 5 modes, $\overrightarrow{x_{best,G}}$ represents the best individuals in number *G* generation population, and *r*1, *r*2, *r*3, *r*4, *r*5 are random integers which represent the individual's serial number in the population.

Through a lot of function testing, there exist certain difference among various models of DE algorithm, in which, DE algorithm of model 1 and model 4 is the best, but DE algorithm of model 1 is the simplest [4, 5].

Improvement of selection strategy. The selection strategy adopted by the standard DE is a kind of greedy strategy, namely, the comparison of fitness value between the parent and offspring generation. The following selection strategies can be adopted to replace the standard DE selection strategy, which is

$$\begin{aligned} &\textit{if}\ \ f(x_i(g)) \le k * \sigma^2(g) \\ &\textit{then}\quad x_i(g+1) = u_i(g). \end{aligned}$$

Where, $\sigma^2(g)$ is a function about the evolution algebra, *k* is a fixed constant, therefore, this algorithm would be different in each generation according to different selection criteria of $\sigma^2(g)$.

In addition, an alternative selection strategy can also be adopted, in which, the probability of selecting the offspring generation to enter the next evolution is shown as follows

$$p_{election} = \min\left(1, \frac{f(x_p)}{f(x_o)}\right).$$

In which, $f(x_p)$ is the objective function value of the parent generation, and $f(x_o)$ the objective function value of the offspring generation.

At present, different combinations of other methods with the advantages of DE algorithm aimed to creation of a distinctive hybrid algorithm is a popular way to improve the current DE algorithm. It was combined with the clustering analysis, neural network, collaborative evolution algorithm, etc. [6, 7].

The flowchart of the DE is shown in Fig. 1.

K-means cluster analysis algorithm. The cluster is to classify the data objects in the data space. The data objects in the same class are usually very similar while those in different classes are quite different [8]. The inputs of a cluster analysis cluster include a group of samples and the standard to measure the similarity (or dissimilarity) of two samples and the outputs are several groups (classes) of the data sets, which constitute a partition or a partition structure [9]. The diagram of clustering analysis is shown in Fig. 2.

The basic idea of K-means cluster method includes: random selection of K objects from the dataset as the initial cluster centers; calculation of the distance of every object to every cluster and assigning the objects to the class where the nearest cluster center is located; calculation of the mean value of the data objects of every newly-formed cluster and obtaining a new cluster center. If there appears no change in the neighbor cluster centers twice, the sample adjustment ends and the cluster criterion function has converged [10]. This process will repeat itself

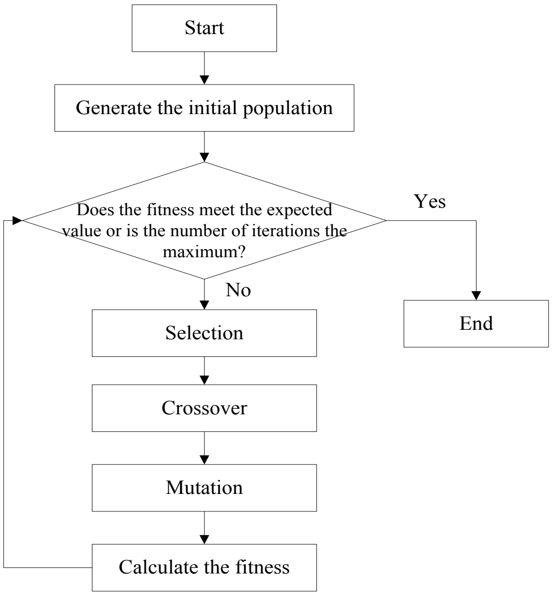
Start
Generate the initial population
Does the fitness meet the expected value or is the number of iterations the maximum?
Yes
No
Selection
End
Crossover
Mutation
Calculate the fitness

Fig. 1. The flowchart of DE

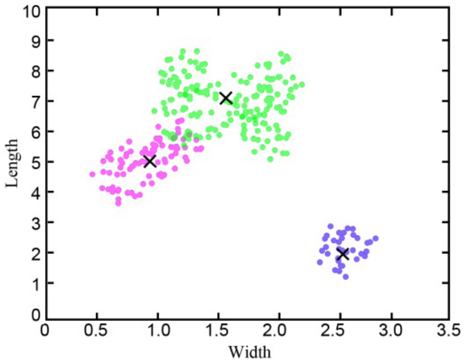

Fig. 2. Diagram of clustering analysis

continuously until it meets certain termination condition. The steps of this algorithm are as follows:

1. Input K classes and the database including N objects.

2. For the data object set, randomly select K objects as the initial cluster centers.

3. Cluster every sample to one of the nearest k samples.

4. Put every object with the most similar classes according to the mean value of the objects in the classes.

5. Calculate the mean value of every cluster and replace the original cluster center with the new mean value.

6. Repeat step 3, step 4 and step 5.

7. Meet the convergence conditions and the running ends.

The flowchart of K-means cluster is shown in Fig. 3.

Solve cluster analysis problems with differential evolution based on elite strategy. In the improved differential evolution, an individual is a cluster center and the individual fitness is calculated according to the fitness function. By using diversity radius mechanism, every generation will use the modified standard fitness sharing function to calculate the new fitness for the individual. Then conduct selection, crossover and mutation operations on the individuals of the population. Elite strategy is used to replace the worst individual in the current population.

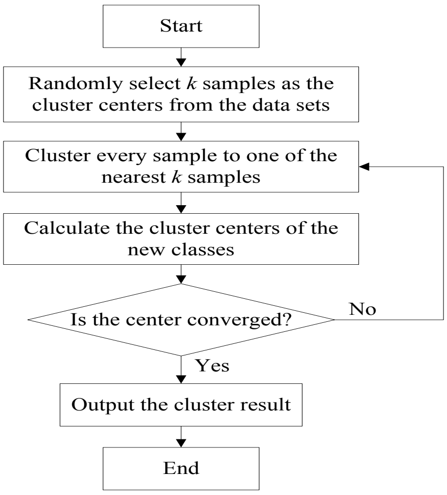
Start
Randomly select k samples as the cluster centers from the data sets
Cluster every sample to one of the nearest k samples
Calculate the cluster centers of the new classes
Is the center converged?
No
Yes
Output the cluster result
End

Fig. 3. The flowchart of K-means cluster

K-means clustering based on DE algorithm has extraordinary clustering effect on some representative expression data. By simulating the evolution process of the natural world, K-means clustering based on DE adopts the principle of ‘survival of the fittest’ and searches the optimal solution with replication, crossover and mutation. Since it integrates the simplicity of K-means and the global search ability of DE, this algorithm has both, the global search capacity and hill-climbing capability. In the meanwhile, it makes up for the shortcomings of K-means algorithm and DE.

This research aims to design a differential evolution K-means clustering algorithm based on the elite strategy. This algorithm can not only select the cluster center through crossover operation but also automatically learn the value of K via mutation. At the same time, it overcomes the locality of the K-means clustering algorithm because of the globality of the DE algorithm.

The steps of the differential evolution clustering algorithm based on elite strategy are as follows.

Step 1: Initialize the encoding and population.

Set the relevant parameters, including the initial number of clusters k, the population size m, the crossover probability p_c, the mutation probability p_m and the maximum iterations t. Randomly select m data points. Every data point is comprised of v-dimensional constants and it represents an individual.

Step 2: Select the fitness function.

Select the population of the next generation according to the fitness of the individuals, which depends on the value of the fitness function. The objective is to minimize the mean square error function so that the individual with a minimal value of mean square error function is more probable to be preserved and that individual will have bigger fitness value. The fitness function is defined as follows

$$fitness = \frac{d_{\min}}{c_{avg}(x)}.$$

Here, $d_{\min}$ is the minimum between-class distance and $c_{avg}(x)$ is the average within-class distance and their definitions are as follows

$$d_{\min} = \min_{i,j=1}^{k} \left\| c_i - c_j \right\|^2; \qquad (1)$$

$$c_{avg}(x) = \frac{1}{k} \sum_{i=1}^{k} \left(\sum_{j=1}^{n_i} \left\| x_j - c_i \right\|^2 \Big/ n_i \right),$$

where, c_i is the class center and x_j is the distance. The fitness function mainly demonstrates that the between-class distance shall be as loose as possible while the within-class distance shall be as compact as possible. When the fitness function is optimized in the initial center, it can inspire the K value to learn the optimal number of clusters automatically. Take this idea as the foundation of the design of fitness function. If the value of the objective function is smaller, the sum of within-class scatter is smaller and the cluster partitioning quality is better.

Step 3: Selection Operation.

Selection operation is to select the individuals according to the value of the fitness function. The individuals with bigger fitness function value are more probably selected; otherwise, they may be eliminated. This research adopts the elite selection strategy, namely to select the individual with the optimal fitness value from the parent individuals and the experimental individuals as the individuals of the next generation.

Step 4: Crossover Operation.

Crossover operation is the relevant part of the multiple individuals combined together to form a new individual. Different populations use different initial crossover probabilities.

$$p_c = \begin{cases} p_{c1} - \dfrac{(f' - f_{avg})(p_{c1} - p_{c2})}{f_{\max} - f_{avg}}, & f_{avg} \le f' \\ p_{c1}, & f_{avg} > f' \end{cases}.$$

The values of p_{c1} and p_{c2} in different sub-populations are different. f' is the bigger fitness value in the two crossover individuals, $f_{\max}$ is the largest fitness value in the population and f_{avg} is the average fitness value of every generation of the population.

Step 5: Mutation Operation.

Mutation refers to individual change and obtain new individuals. Newly generated individuals continue to be evaluated. Select relatively good individuals from the parent and offspring populations to form new populations.

$$p_m = \begin{cases} p_{m1} - \dfrac{(f_{\max} - f)(p_{m1} - p_{m2})}{f_{\max} - f_{avg}}, & f_{avg} \le f \\ p_{m1}, & f_{avg} > f \end{cases}.$$

The values of p_{m1} and p_{m2} are different for the sub-populations with different mutation probabilities. f is the fitness value of the mutated individuals.

Step 6: K-means Cluster Operation.

For the given matrix M, use the formula of the minimum between-class distance (1) to calculate the cluster center. This formula ensures that there is at least one object in a class.

The centroid of the k class $C_k = \{C_{k1}, C_{k2}, \ldots, C_{kd}\}$. Here,

$$C_{kj} = \frac{\sum_{i=1}^{n} \gamma_{ik} x_{ij}}{\sum_{i=1}^{n} \gamma_{ik}}.$$

The Within-Class Variation (WCV) of the kth class is defined as

$$\mathrm{WCV}(M) = \sum_{i=1}^{n} \gamma_{ik} \sum_{j=1}^{d} (x_{ij} - c_{kj})^2,$$

where, x_{ij} is the jth property of the object x_i, $i = 1, 2, \ldots n$, if the ith individual belongs to the kth class, γ_{ik} is 1, otherwise, it is 0. Redistribute every object to the class whose nearest cluster center belongs to and generate a new matrix.

Step 7: Preserve the optimal individuals.

If the evolution iterations have been reached, find the optimal individuals in every sub-population, select the optimal individual from them and then spread to all the sub-populations; otherwise, turn to Step 3. In every iteration, record the individual with the minimum within-class variation. If the current optimal individual meets the convergence conditions, end it and output the result. The optimal individual obtained represents the optimal cluster result.

The mutation operation is mainly to finish the increase and decrease of the individual genes in the population, namely to clarify the final number of clusters. The mutation probability determines the implementation frequency of mutation operation and its value affects the population diversity and the convergence performance of the algorithm. This research uses fixed value in the selection of mutation probability, however, after integrating the actual algorithm, it can be seen that the selection of K value in the first place is not certain and it is greatly variable, therefore, the mutation probability at the beginning can be bigger. However, as the K value goes towards the optimal K value continuously, the mutation probability will reduce afterward. Besides, the mutation probability will directly affect the final clustering effect in the entire implementation process.

Experiment result and analysis. Four artificial data sets and the setting of their experiment parameters were described. These four data sets include different numbers of clusters and distribution. These data points conform to a normal distribution; they are generated surrounding the cluster centers. The diagram of the four data sets is shown in Fig.4.

The other control parameters of the four data sets are the same, namely that the crossover probability is 0.8 and the mutation probability is 0.03. Fig. 5 shows the cluster analysis results of the improved algorithm on the four data sets respectively, "–" is the location of the data point and "+" is the cluster center. It can be seen that the elite strategy differential clustering algorithm can figure out the numbers of clusters of the data sets and the location of the cluster center successfully.

Fig. 5 demonstrates the cluster analysis results, which the algorithm makes on the data sets in different initial radiuses. In the Fig. 5, the initial radius of data set in (*a*) is larger, the data set is less, the clustering is more dispersed and the variance is bigger, thus making data between any

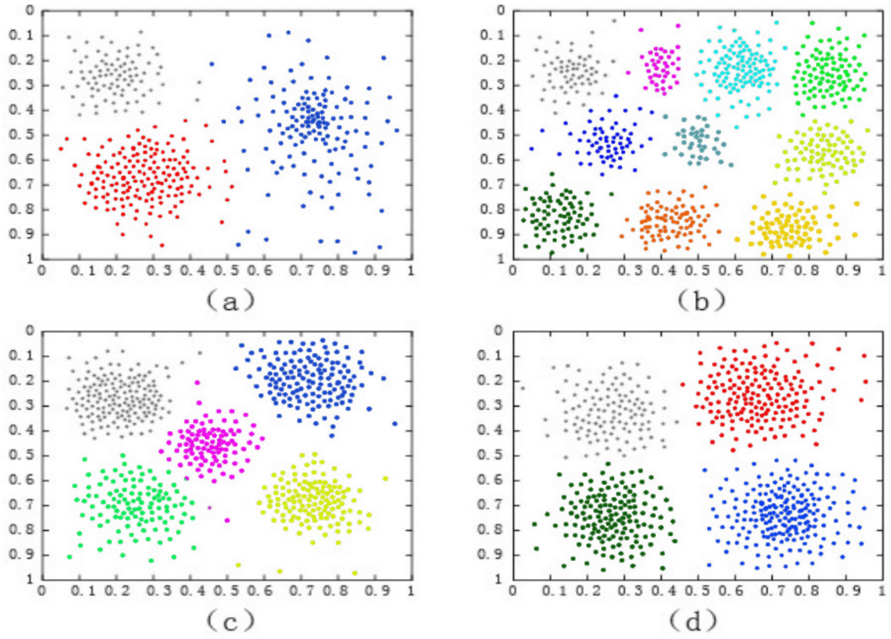

Fig. 4. The diagram of the data sets: a – is the 1st data set, which includes 200 data set with data points and 3 cluster centers; b – is the 2nd 700 data points and 10 cluster centers; c – is the 3rd set with 400 data points and 5 cluster centers; d – is the 4th set with 300 data points and 4 cluster centers

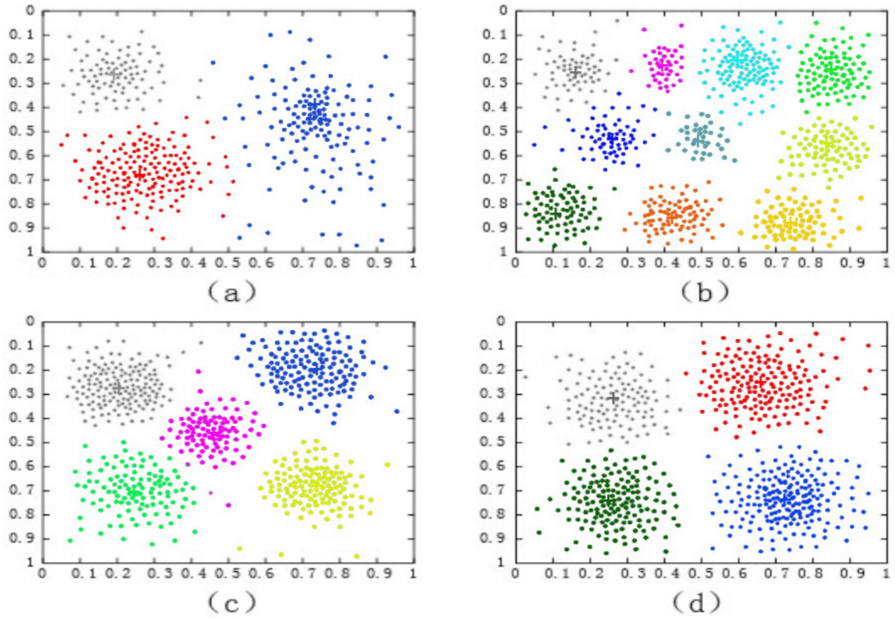

Fig. 5. Cluster centers obtained by the four data sets

two classes of data does not completely belong to one class, but belongs to several classes. The experimental simulation results showed that the final clustering center results do not result in the deviation. The number of clustering sample points in (*b*) is bigger, the data set is also larger, and the variance is smaller. The results showed that the clustering center distribution of each data set is more reasonable. The number of clustering sample points in (*c*) is bigger, but the variance of each data set is smaller. The results showed that the optimal class center result of the proposed algorithm met the need with better effect. The number of sample points in (*d*) is bigger, but the data set is smaller, and there are a number of points drifting away each data. The method proposed was adopted to search the clustering center of each data set. The results showed that the deviation of clustering center is small with relatively ideal effect.

It can be seen clearly that no matter what the initial radius is, the cluster results are very satisfactory. Therefore, a conclusion can be made that the algorithm can overcome the problem of the initial radius and it has excellent robustness whether in the experiment functions or in the practical applications.

Conclusion. The improved K-means cluster algorithm based on DE algorithm has been developed. As an algo-

rithm that searches the optimal solution by simulating the process of natural evolution, DE makes itself stand out with its implicit parallelism and ability to utilize the global information effectively. Therefore, the new improved method has better robustness, avoids being trapped in local optimum, enhances clustering effect greatly, and has such characteristics as avoiding prematurity and fast convergence. The experiments have verified the effectiveness of the new method.

Acknowledgements. This work was financially supported by National Natural Science Foundation of China (Grant No. 61103202).

References / Список літератури

1. Enmei Tu, Longbing Cao, Jie Yang and Nicola Kasabov, 2014. A novel graph-based K-means for nonlinear manifold clustering and representative selection. *Neurocomputing*, vol. 143, no. 2, pp. 109–122.

2. Michio Yamamoto and Yoshikazu Terada, 2014. Functional factorial image-means analysis. *Computational Statistics & Data Analysis*, vol. 79, no. 11, pp. 133–148.

3. Basu, M., 2014. Improved differential evolution for economic dispatch. *International Journal of Electrical Power & Energy Systems*, vol. 63, no. 12, pp. 855–861.

4. Ali Wagdy Mohamed, 2014. RDEL: Restart differential evolution algorithm with local search mutation for global numerical optimization. Egyptian Informatics Journal, vol. 15, no. 3, pp. 175–188.

5. Pratyay Kuila and Prasanta K. Jana, 2014. A novel differential evolution based clustering algorithm for wireless sensor networks. *Applied Soft Computing*, vol. 25, no. 12, pp. 414–425.

6. Das, S., Konar, A., Chakraborty, U.K., 2005. Improved differential evolution algorithms for handling noisy optimization problems. In: IEEE. *The 2005 IEEE Congress on Evolutionary Computation*, vol.2, pp. 1691–1698.

7. Qingya Zhou, 2014. The research of differential evolution under dynamic environment. *Zhengzhou University, China*.

8. Md Anisur Rahman and Md Zahidul Islam, 2014. A hybrid clustering technique combining a novel genetic algorithm with K-Means. *Knowledge-Based Systems*, vol. 71, no. 11, pp. 345–365.

9. Grigorios Tzortzis and Aristidis Likas, 2014. The MinMax K-Means clustering algorithm. *Pattern Recognition*, vol. 47, no. 7, pp. 2505–2516.

10. Velmurugan, T., 2014. Performance based analysis between K-means and fuzzy C-means clustering algorithms for connection oriented telecommunication data. *Applied Soft Computing*, vol. 19, no. 6, pp. 134–146.

Мета. Кластерний аналіз – це не тільки важливий напрям досліджень у сфері інтелектуального аналізу даних, але й важливий засіб і метод поділу даних або обробки пакетів. Дослідження присвячене подальшому поліпшенню результативності алгоритму кластеризації та усуненню існуючих недоліків диференціальної еволюції (DE). Результати дослідження призначені для застосування у кластерному аналізі з метою отримання кращого ефекту кластеризації.

Методика. Проведені глибокі дослідження DE-алгоритму та кластерного аналізу, розглянуто вплив методу k-середніх, а також блок-схем і методу розрахунку функції пристосованості. Проаналізовано вплив різних диференціальних операцій на продуктивність.

Результати. По-перше, розглянуті основні ідеї й методи кластерного аналізу та DE-алгоритму. По-друге, продемонстрована реалізація кластерного аналізу поліпшеним DE-алгоритмом. По-третє, проведене експериментальне моделювання кластерного аналізу на чотирьох наборах змодельованих даних за допомогою алгоритму кластеризації на основі DE-алгоритму з елітарною стратегією, що дало можливість перевірити доцільність і обґрунтованість нового методу.

Наукова новизна. Розроблено DE-алгоритм з елітарною стратегією для застосування у кластерному аналізі за методом k-середніх. Так як DE-алгоритм являє собою метод для пошуку оптимального рішення шляхом імітації природного еволюційного процесу, його відмінною рисою є його прихований паралелізм і здатність ефективно використовувати глобальну інформацію, таким чином, новий і покращений алгоритм більш стійкий і може уникнути попадання в пастку локального оптимуму та значно посилити ефект кластеризації. Дослідження цього аспекту раніше не проводилися.

Практична значимість. Застосування елітарної стратегії DE-алгоритму може підвищити ефективність і точність кластерного аналізу за методом K-середніх. Результат експериментального моделювання показав, що новий метод, представлений у цій роботі, значно поліпшив продуктивність оптимізації, що доводить його доцільність та ефективність.

Ключові слова: *кластерний аналіз, метод K-середніх, диференціальна еволюція, елітарна стратегія, оптимізація продуктивності, доцільність, ефективність*

Цель. Кластерный анализ является не только важным направлением исследований в сфере интеллектуального анализа данных, но и важным средством и методом разделения данных или обработки пакетов. Исследование посвящено дальнейшему улучшению результативности алгоритма кластеризации и устранению существующих недостатков дифференциальной эволюции (DE). Результаты исследования предназначены для применения в кластерном анализе с целью получения лучшего эффекта кластеризации.

Методика. Проведены глубокие исследования DE-алгоритма и кластерного анализа, рассмотрено влияние метода k-средних, а также блок-схем и метода расчёта функции приспособленности. Проанализировано влияние различных дифференциальных операций на производительность.

Результаты. Во-первых, рассмотрены основные идеи и методы кластерного анализа и DE-алгоритма. Во-вторых, продемонстрирована реализация кластерного анализа улучшенным DE-алгоритмом. В-третьих, проведено экспериментальное моделирование кластерного анализа на четырех наборах смоделированных данных с помощью алгоритма кластеризации на основе DE-алгоритма с элитарной стратегией, что дало возможность проверить целесообразность и обоснованность нового метода.

Научная новизна. Разработан DE-алгоритм с элитарной стратегией для применения в кластерном анализе по методу k-средних. Так как DE-алгоритм представляет собой метод для поиска оптимального решения путем имитации естественного эволюционного процесса, его отличительной особенностью является его скрытый параллелизм и способность эффективно использовать глобальную информацию, таким образом, новый и улучшенный алгоритм более устойчив и может избежать попадания в ловушку локального оптимума и значительно усилить эффект кластеризации. Исследования этого аспекта ранее не проводились.

Практическая значимость. Применение элитарной стратегии DE-алгоритма может повысить эффективность и точность кластерного анализа по методу K-средних. Результат экспериментального моделирования показал, что новый метод, представленный в этой статье, значительно улучшил производительность оптимизации, что доказывает его целесообразность и эффективность.

Ключевые слова: *кластерный анализ, метод K-средних, дифференциальная эволюция, элитарная стратегия, оптимизация производительности, целесообразность, эффективность*

Рекомендовано до публікації докт. техн. наук В. В. Гнатушенком. Дата надходження рукопису 17.04.15.

Liu Ning Shangluo University, Shangluo, China

ENSEMBLE CLASSIFICATION ALGORITHM BASED IMPROVED SMOTE FOR IMBALANCED DATA

Лю Нін Шанлонський університет, м. Шанло, КНР

ПОКРАЩЕНА SMOTE-СТРАТЕГІЯ КЛАСИФІКАЦІЇ НЕЗБАЛАНСОВАНИХ ДАНИХ НА ОСНОВІ АНСАМБЛЕВОГО АЛГОРИТМУ

Purpose. In practical application, the accuracy of the minority class is very important and the research on imbalanced data has become one of the most popular topics. In order to improve the classification performance for imbalanced data, the classification algorithm based on data sampling and integration technology for imbalanced data was proposed.

Methodology. Firstly, the traditional SMOTE algorithm was improved to K-SMOTE (an over-sampling method based on SMOTE and K-means). In K-SMOTE, the dataset was to perform clustering operation, and the interpolation operation was performed on the connection of the cluster center and the original data point. Secondly, ECA-IBD (an ensemble classification algorithm based improved SMOTE for imbalanced data) was proposed. In ECA-IBD, over-sampling was conducted by K-SMOTE, and random under-sampling was carried out to reduce the problem scale to form a new dataset. A number of weak classifiers were generated and integration techniques were used to form the final strong classifier.

Findings. Experiment was carried out on the UCI imbalanced dataset. The results showed that the proposed algorithm was effective by using the F-value and G-mean value as the evaluation indexes.

Originality. In the paper, we improved the SMOTE algorithm and combined over-sampling technology, under-sampling technology and boosting technology to solve the classification problem for imbalanced data.

Practical value. The proposed algorithm has important value in imbalanced data classification. It can be applied in the field of different kinds of imbalanced data classification, such as fault detection, intrusion detection, etc.

Keywords: *imbalanced data, ensemble learning, over sample, under sample, data classification*

Introduction. Classification problem is one of the most important in the field of data mining. Traditional classification methods have achieved good results on balanced datasets, but the actual datasets are often imbalanced. For the traditional classifier, it aims at pursuing the overall classification accuracy. The imbalance of the dataset is bound to cause the classifier to pay more attention to the majority class samples so that the classification performance of the minority class samples declines [1,2]. However, in practical application, people are more concerned about the minority class data, and the cost of the error in its classification is usually larger than that of the majority. For example, if the cancer patients were diagnosed as normal, it would delay the optimal timing of treatment, resulting in life threatening for patients. If the fault is identified as normal, it leads to failure undetected and may lead to major accidents. In network intrusion detection, if the network intrusion behavior is sentenced to normal behavior, it will have the potential danger to cause major network security incidents. Therefore, in practical application, it is more needed to improve the classification accuracy of the minority class samples. The research on imbalanced data has become one of the most popular topics [3].

The imbalanced classification is such a problem where the number of training samples in the class distribution is not balanced and the number of samples in one class is far