A New K-Means Clustering Algorithm for Customer Classification in Precision Marketing

Xinwu Li[1*],Xiaoling Du[2]

[1]*Department of Electronic Business, Jiangxi University of Finance and Economics,Nanchang, Jiangxi, China*

[2]Institute of International Business and Economics, *Jiangxi University of Finance and Economics,Nanchang, Jiangxi, China*

**Corresponding Author*

Abstract

K-means is wildly used in data mining and clustering for its powerful data clustering ability, but its inherent limitations affect its application fields and accuracy. Theoriginal K-means algorithm is improved and applied in customer clustering in precision marketing. Firstly, integrates K-means algorithm with particle swarm optimization according to analyzing the source of the K-means calculation limitations; Secondly, improves the improved algorithm in its operation time, convergence speed, global solution exploration ability successively and redesigns the calculation procedures; Finally applies it in customer classification in precision marketing and the experiment results shows that the new algorithm can increasecustomer clustering effectiveness, validity, accuracy and has satisfactory results in practice.

Keywords*: Data mining, customer clustering, K-means, particle swarm optimization, precision marketing*

I. Introduction

With coming of the big data age, the machine learning technology has developed greatly. Data cluster algorithm is wildly used and favored in traditional machine learning, and has been applied in variousrelated fieldssuccessfully[1], for example document clustering, customer classification, image segmentation, features learning etc.,because of its simple, efficient and practical calculation advantages[2-4]. Data clustering is a vital concept in data mining too and its aim is to identifymeaningful information hidden in the data to be processed.

The K-Means, as a wildly used clustering algorithms, was presented by MacQueen in 1967, hasbeen widely used in clustering algorithms and developed a large number of improved algorithms due to its better effect and simple principle, and is still a research hotspot[5].

II. Literature Review

2.1Thecalculation principle and limitations of K-means algorithm

The calculation principle of K-means algorithm can be listed as follows: Firstly,chooses an initial clusteringseed at Assigns them to the class of the clusteringseed with the greatest similarity according to the closest distance criterion; Finally, computes the average distanceof all processed points in each group and renews the clusteringseeds until the objective criterion equationconverges[1,6]. The working and calculation thought of K-means can be indicated in Figure 1.

K-Means algorithm shows its limitations in practical calculation too. Such as, the clusters number k should be

givenfirst; The initial seedonly can bechosen at random; The effectsfromdiscrete pointsmay appear[1,6].

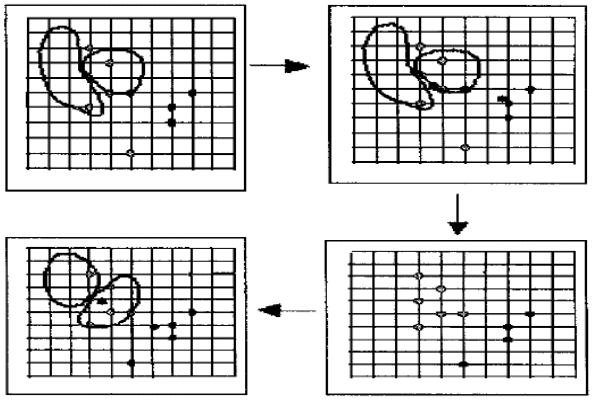

Fig 1:The working and calculation thought of K-means

2.2Literature review related toimproving K-means algorithm

In response to above limitations, scholars in various fields have proposed different improvements. ①Improve initial cluster center selection.JiaRuiyu (2018), Lei Gu (2017), M. S. Premkumar (2017) and M. E. Celebi (2013) use the methods of cluster centers at local density, subtractive clustering algorithms, feature dimensions, maximum-minimum criteria respectively to determine the initial cluster center, and then obtain the best cluster center gradually through related iterative operations[7-10]. The practical results show that the improvements are effective in clustering high-dimensional concentrated data, but not ideal in dealing with sparse data sets. ②Optimize measurement methods for distance and similarity. Euclidean distance is uesed in original k-means algorithm to calculate the similarity between points. W. Xue (2017), J. P. Singh, N. Bouguila (2017) and Chen Leilei (2015) use spatial density similarity, Aitchison distance, and compound distance (comprehensive calculation of Tanimoto, squared Euclidean, cosine, Euclidean and Manhattan distance etc.) respectively to measure the similarity between pixels. Practical results show that these methods have good effects on non-linear image data clustering, but insufficient for business data clustering obviously[11-13]. ③Improve the outliers detection. The density-based method is a popular and common outlier detection method in K-means. Ting Zhang et al. (2017) propose to add an upper limit norm and an effective iterative reweighting algorithm to reduce the outliers influence[14]; P. O. Olukanmi (2017) obtain the global threshold by detecting the distribution of the distance from the point to the centroid, then detect outliers automatically[15]. K. Zhang et al. (2009) establish a discreteindex based on distance to calculate the discrete value of discrete points[16]. Above methods can solve the cluster interference of a certain number of outliers to a certain extent, but its elimination ability is insufficient in dealing with a large amount of outlier data. ④Improvements combined with other algorithms. Chen Xiaoxue et al. (2018), Kapil et al. (2017) and Shen Yan et al. (2014) combines with genetic algorithm, particle swarm algorithm, firefly optimization weighting algorithm respectively and have achieved good clustering results in their own application fields, but these improvements have certain limitations in versatility[17-19].

To sum up: ①it is better to combine with other related algorithms to improve K-means in solving clustering problems of specific fields. ②With the theInternet age coming, the data needed to be processed has shown an exponential growth. How to cluster these massive data (especially multi-dimensional business data) efficiently has become a current research hotspot[17]. This paper tries to integrates the particle swarm optimization algorithm with K-means algorithm together and takes some measures to overcome the limitations of original K-meansalgorithm

and applies the new algorithm to the customer classification in multi-dimensional business data.

III. Improving K-means with particle swarm optimization algorithm

3.1Source analysis of the original algorithm limitations

The clustering result stability is still a big problem for K-means algorithm when used in data clustering. The cluster resultsareideal when the processed data shows the distribution of round or convex, but its results show obvious deviations and errors when sample data is scattered. Clustered data distribution inevitably appears isolation phenomenon, that is, some data are far away fromintensivedatum area, but the clustering average data (the geometric center of whole data to be processed) areregardedto replace the old seed for next round calculation. In this case, the new cluster seed may far from the realintensive datum area, thereby leading to clustering results deviation. So the K-means algorithm has great limitations caused by isolation distribution outlier data[4].

3.2Improving K-means with particle swarm hybridcluster

The paper advances an improved K-means and particle swarm optimization (referred to asPSO) hybrid clustering algorithm. The new algorithm uses population fitness variance to determine the operation time of K-means algorithm, achieve the organic integration of K-means and PSO algorithm, enhances the new algorithm's search and convergence capability. The particle position updating mechanism based on the extrapolation direction is added in the evolution process to overcome the problem of slow convergence speed of original K-means algorithm. The random mutation operation is introduced in the evolution process of PSO algorithm, only the K-means search is carried out on the particles participating in the mutation each time, to enhance the the population diversity without affecting the algorithm convergence speed and overcome the limitation of getting stuck at locally optimal valueeasily of general K-means[17].

3.2.1Optimizing particle swarm

PSO is an evolution method taking intelligent group as its base. Each possible optimization object method is a point in the searching space, and each point has its corresponding velocity, position and a rightness decided by its target definition, and the algorithm evaluates the point quality by its rightness. The new algorithm defines a swarm of random points first; then finds the optimizationmethodthroughcyclic calculation. The particle renews its own valuethrough tracing the maximum and minimum values in everycyclic calculation: one is the optimizationmethodsearched by the point, which is always named as individual maximum or minimum value $pBest$; The other is the optimizationmethodsearched by the total particle swarm, which is always named aswholemaximum or minimum value $gBest$. After getting the two $pBest$ and $gBest$ values, the pointrenews the velocity and location of its own though Equations (1) and (2)[4].

$$v_i(n+1) = wv_i(n) + c_1 \cdot rand_1() \cdot (pBest - p_i(n)) + c_2 \cdot rand_2() \cdot (gBest - p_i(n)) \quad (1)$$

$$p_i(n+1) = p_i(n) + v_i(n+1) \quad (2)$$

In the Equations, $v_i(n)$ represents the present point speed, $p_i(n)$ represents the present point position, $i = 1,2,3,....N$, N represents the present space number, $rand_1()$, $rand_2()$ represents a random value in the

range of [0, 1] respectively, c_1, c_2 representsan evolutionary indicator respectively, usually $c_1 = c_2 = 1$, w is a weighting coefficient, generally taken as a value in the range of [0.1, 0.9]. The existing research results show that when weighting coefficient descends with the cyclic calculationlinearly, the convergence speed of the new algorithm will be accelerated significantly. Let $w_{\max}$, $w_{\min}$ be the two extreme weighting coefficient ,maximum and minimum respectively, run is present iteration value, and $runMax$ is the whole iterations number, then Equation (3) exists[5].

$$w = w_{\max} - run\frac{(w_{\max} - w_{\min})}{runMax} \quad (3)$$

3.2.2K-means and particle swarm optimization hybrid clustering algorithm

K-means clustering hasextensiveuse in massivedata analysis, such as artificial intelligence, machine learning, business data clustering and various related fields for its easyunderstandingcalculation principle and effectivecalculation speed. But, two inherent limitations exist in the original K-means too: the influence of initial clustering seed definition is too large, and it is difficult to find the global optimum. The emergence of PSO algorithm offers a novelapproaches to overcome these limitations. How to fully utilizeand integrate the algorithm advantages of PSO (powerful global search capability) and K-means (precise local search method and capability), and improving the solution accuracy and accelerating the algorithm convergence speed become the vital factors to success of new algorithm (K-means + PSO hybrid algorithm) through analyzing previous literature[17].

3.2.3Determining the operation time for K-means

In order to combine and integrate the K-means and PSO algorithm organically, the of operation timeK-means should be determined first. When PSO search seeds globally and randomly, the K-means can and should stop work, by doing these it can use PSO algorithm to approximate to the global method subspace to accelerate the convergence process of the improved hybrid algorithm. K-means canbe used to increase the local search capacity and speed up the convergence process in PSO working state. So we can realize the organic integration of the hybridalgorithm of K-means+PSO just to determine when the PSO algorithm converges.

Whether PSO algorithm achieveslocal optimum or global optimum, points in particle swarm optimization may appear "aggregation" phenomenon, At this time, the position of each particle is the same, that is, the fitness of each particle is the same. Therefore, It can track the particle swarm state and judge whether the algorithm can converges or not to study the overall rightnesschange of total points particles in PSO. Suppose n represents the number of point group, f_i is the rightness of the first i point, f_{avg} is the meanrightness of point group at present, and the population rightness variance of particle swarm σ^2 is defined in Equation (4).

$$\sigma^2 = -\sum^{N}\left\{\frac{f_i - f_{\text{avg}}}{f_i}\right\} \quad (4)$$

The swarmrightness variance means the global optimum degree of all processed data in point group. With the decrease of particle swarm, the σ^2 has more convergence tendency. When σ^2 is zero, the group fitness is almost the same, and the PSO falls into global optimum or precocious optimum. On the contrary, the particle swarm has different rightnesswhen the algorithm isin random seed search calculation. So, the group rightness variance can be used to judge the K-means algorithm operation time. When a certain given value is more than the group

rightnessvariance, PSO algorithm begins to perform the local solution accurate search at the convergence stage, which not only improves the global solution search performance, but also accelerates the global optimal search speed of the hybrid clustering algorithm to increase the final calculation result accuracy.

3.2.4Improving the hybrid clustering algorithm convergence speed

In order to speed up hybrid clustering algorithm in the initial iteration phase, the extrapolation direction is introduced to update the particle position. Before the particle reaches its optimal value, If the rightness of the particle position after iteration is greater than the current rightness, another extrapolation optimal pointcan be shown in Equation (5).

$$p(n+1) = p(n+1) + k(p(n+1) - p(n)) \quad (5)$$

Where k is the adjustment coefficient. Because it has long distance from the optimum result in beginning stage, the larger regulation amplitude is conducive to accelerate evolution; When it is closer to the optimal solution in the later calculation, the adjustment range is small to gradually approach the optimal solution, so the adjustment coefficient is shown in Equation (6).

$$k = \exp(-20 \times (run / runMax)^{10}) \quad (6)$$

For the multivariable optimization problem, because each particle position has many components, it is easy to appear some very close components or even the same two particles, so the equation (5) does not work. In this case, a small random number can be added after Equation (5), which only plays a role in the later evolution stage to strengthen the fine-tuning amplitude, and the position calculation is shown in Equation (7).

$$p(n+1) = p(n+1) + k(p(n+1) - p(n)) + 10^{-6} \cdot rand \quad (7)$$

3.2.5Improving the global solution exploration ability

Because the global solution searching ability of the new algorithm presented in the paper is completely dependent on the exploration results of global solution space in the early PSO calculation stage. Therefore, introducing random mutation operation in the particle swarm can avoid falling into the local extremum and early convergence of PSO algorithm. Because the particles with good fitness do not need to be mutated, this paper just carries out random mutation operation on the particles with poor fitness, and other preferred particles continue to carry out local search while maintaining the original population structure, thus realizing the balance between improving the convergence speed and maintaining the population diversity, as shown in Equation (8).

$$if \quad (r_i \le C_v) \quad then \quad v_{id} = r_2 \times r_3 \times V\max / C_m \quad (8)$$

In Equation 8, r_i ($i \in M$) and r_2 representtwo variableswhich are randomly and uniformly and its value is in the range of [0,1], M represents a part of points with poor rightness after sorting. r_3 represents a random variabledistributedrandomly used to control the particle flight direction, and it is 1 when the random number is

less than 0.5, it is -1 when the random number is greater than 0.5.

3.2.6Designing the calculation procedures for the presented algorithm

The presented hybrid clustering algorithm in the paper (K-means+PSO) adopts the coding mode based on clustering center. That is, every position of every point is consist of m clustering seeds. Besides their positions, all points also have their own speed and rightness. Let the vector number of data to be processedas d, so the velocity and position of all points are $m \times d$ dimensionvariables, and everypoint has its own rightness f_i too. In this way the particle can adopt the encoding structure as $c_1^1 c_1^2 ... c_1^d ... c_m^1 c_m^2 ... c_m^d v_1^1 v_1^2 ... v_1^d ... v_m^1 v_m^2 ... v_m^d f_i$. When determining the cluster center, The clustering division is determined by the nearest neighbor rule as Equation (9), if x_i, c_j satisfy equation (9), then x_i belongs to class j.

$$\left\| x_i - c_j \right\| = \min \left\| x_i - c_k \right\| \qquad k = 1,2...,m \quad (9)$$

For some specific point, it can compute its rightness according to Equation (10), and L represents the number of specimendata and x_i represents the input specimen data.

$$fitness_i = \sum_{i=1}^{L} \sum_{j=1}^{m} \left\| x_i - c_j \right\|^2 \quad (10)$$

The specific algorithm flow is as follows:

(1) Initializing a point group, i.eparticle swarm. Assigns each specimen data to a specific group as first cluster seed, computes the succeedingcluster seeds for every group and takes the calculation results to encode its first position; calculates the pointrightness and takes the results as its individual optimal position, and randomly initializes the particle velocity;Repeats above calculation N times, and each time calculation can get one first point group;

(2) For every point, their rightness can be compared with the optimalrightness position in their calculation, and renew the optimal position of the particle for the better.

(3) Adjusting the velocity and the position of all points based on the PSO algorithm;

(4) Generate optimal extrapolation particles by Equation (7);

(5) Adding a random mutation operation with small probabilities by Equation (8) and carrying out K-mean operation on the particles which are involved in participation mutation;

(6) Judging whether the current particle swarm has reached the convergence state according to Equation (4). if the group fitness variance σ^2 is less than the threshold $thre\sigma$, select P_m optimal particle to carry out a K-means local search, so as to jump out of the precocious convergence trap; For selected particles, ①follow the following

K-means algorithm calculation procedures to optimize: Determining the cluster division corresponding to the particles according to a nearest neighbor rule, and taking the particle cluster center code as an initial value; ②Calculating new cluster centers based on the clusteringgroup to substitute for the old code; Because K-means has strong local searchingcapacity, the global maximum searchingcapacity of the improved algorithm in the paper can be assured; ③If calculation reach the largestcyclicnumber or local optimum, it can be regarded as the new cluster center rightness. And when the old point is worse than the new one, then renew the old point and finish this roundcalculation, otherwise turn to step (2).

IV.Experimental verification

4.1Experimental data processing and user feature dimension determination

Customer classification of precision marketing is an essential application and research field of business datumanalysis. Here, the algorithm improved in this paper is applied to customer classification of precision marketing to illustrate the validity and scientificity of this new hybrid improved algorithm.

Sample selection: Considering the privacy and availability of customer data, this paper selects China Unicom Communications Corporation customer data as the research sample, including 4000 valid 4G, 5G, fixed-line and broadband users respectively, totally valid 20000 customers.

Customer characteristic dimension determination:combined with relevant literatures, this paper finally determines the dimensions as follows: monthly average caller call duration, average calls per month, monthlyaverage calls on different networks, monthlyaverage data traffic, monthly average wireless traffic, monthlypeer-to-peer SMS/ MMS volume, network access time, monthlyaverage payment amount, monthly average total communication consumption and terminal equipment brand.

Data preprocessing: ①Remove specific user data, such as removing some items with a small amount of data or some special user data; This step can overcome the influence of the scattered data; ②Data structure and integration, adjusts all the data types of the original data table through extraction, merging, and deriving to obtain the intermediate table and convert the data from text type to digital data.

4.2Experimental results

The paper realized the improved algorithm with the user characteristic dimension designed before and collected data. The verificationoutcomes are indicated in the Table from 1 to 3. Revenue contribution proportion in the Table 1 is the statistical data by the China Unicom Communications Corporation corresponding to variouskinds of users, . Table 2 is the verificationoutcomes with the different number of users and Table 3 is the verificationoutcomes realized by the various algorithms which have relatively good practical application effects. And E in table 2.and table 3. means the square errors in user clustering.

Table 1 Usertype classification results

Customer Type	Number of Customers	Percentage %	Revenue Contribution Proportion %
The most important	1434	7.17	55.76
Important	2984	14.92	28.16
Ordinary	5019	25.10	10.71
Less important	6578	32.89	4.31
Least important	3985	19.93	1.06
Total	20000	100.00	100.00

Table 2User classification effectiveness of different number

Algorithm	20000 users	5000 users	1000 users
Accuracy rate	98.86 %	99.14%	98.14%
E value	103.24	102.88	104.17

Table 3Performance comparison of different algorithms

Algorithm	The new hybrid algorithm	Original K-means	K-means in reference17	K-means in reference 2
Accuracy rate	98.86 %	80.22%	92.19%	93.87%
E Value	103.24	176.66	120.96	119.96

V. Conclusion

According to the verificationoutcomes above, it can get following conclusions. ①The improved algorithm in the paper has an satisfactory application results when used in business data mining such as user clustering practically, see in table 1. ②Inthe practice of business data analysis, the amount of data has less influence on the improved algorithm, and the improved algorithmalso has a satisfactory effect in dealing with big data, see in table 2. ③In comparison with the some current widely used algorithm, the new hybrid algorithm in the paper has higher datum analysis accuracy and lower errors,indicated in table 3.

Acknowledgements

Theresearch is funded by the The 12th Five Year Plan of Social Sciences in Jiangxi Province（hosted by Ye Hankun) and education department of Jiangxi Province (GJJ150458).

References

[1] K. J. Anil,“Data clustering:50Years beyond K-means,” Pattern Recognition Letters, vol.31, no. 8, pp. 651-666,2010.
[2] M. Tleis, R. Callieris, R. Roma,“Segmenting the organic food market in Lebanon: An application of K-means cluster analysis,”British Food Journal, vol.119, no. 7, pp.1423-1441,2017.
[3] N. Gokilavani, B. Bharathi, “Test case prioritization to examine software for fault detection using PCA extraction and K-means clustering with ranking,”Soft Computing（prepublish), pp. 1-10, 2021.
[4] F. Ji, W. Siyu,W.gChenchengyue, “Peer to peer lending platform risk identification method based on factor analysis and K-means cluster algorithm,” Journal of Chongqing Normal University (Natural Science), vol. 37, no. 5, pp.96-102,2020.
[5] N. Dhanachandra, K. Manglem, Y. J. Chanu,“Image segmentation using K-means clustering algorithm and subtractive clustering algorithm,” Procedia Computer Science, no. 54, pp.764-771, 2015.
[6] A. Rodriguez, A. Laio,“Clustering by fast search and find of density peaks,” Science, vol. 344, no. 6191, pp.1492-1496,2014.
[7] J. Ruiyu, L. Yulong,“K-means algorithm of clustering number and centers self-determination,”Computer Engineering and Applications, vol. 54, no. 7, pp.152-158,2018.
[8] L. Gu, “A novel locality sensitive K-means clustering algorithm based on subtractive clustering,”IEEE International Conference on Software Engineering and Service Science, pp.836-839, 2016.
[9] M. S. Premkumar, S. H. Ganesh, “A median based external initial centroid selection method for K-Means clustering,”World Congress on Computing and Communication Technologies,IEEE Computer Society, pp.143-146,2017.
[10] M. E. Celebi, H. A. Kingravi, P. A. Vela, “A comparative study of efficient initialization methods for the K-Means clustering algorithm,” Expert Systems with Applications, vol. 40, no. 1, pp.200-210, 2013.

[11] W.Xue, R.l.Yang, X. Y. Hong, “A novel K-Means based on spatial density similarity measurement,”The 29th China Conference on Control and Decision Conference, IEEE, pp.7782-7784, 2017.

[12] J. P. Singh, N. Bouguila, “Proportional data clustering using K-means algorithm:A comparison of different distances,”IEEE International Conference on Industrial Technology, IEEE, pp.1048-1052, 2017.

[13] C. Leilei, “Text clustering study with K-Means algorithm of different distance measures,” Software, vol. 36, no. 1, pp.56-61, 2015.

[14] T. Zhang, F. Yuan, L. Yang, “Capped robust K-means algorithm,”International Conference on Machine Learning and Cybernetics. IEEE, pp.150-155,2017.

[15] P. O. Olukanmi, B.Twala. K-meanssharp, “Capped robust K-means algorithm,” 2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-Robmech), Bloemfontein, pp.14-19, 2017.

[16] K. Zhang, M. Hutter, H. Jin, “A new local distance-based outlier detection approach for scattered real-world data,” Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, pp.813-822,2009.

[17] C. Xiaoxue, W. Yongqing, R. Min, M. Yuanyuan, “Weighted K-means clustering algorithm based on firefly algorithm,” Application Research of Computers, vol. 35, no. 2, pp.466-470,2018.

[18] S. Kapil, M. Chawla, M.D. Ansari, “On K-means data clustering algorithm with genetic algorithm,” 2016 Fourth International Conference on Parallel, Distributed and Grid Computing, IEEE,pp.202-206, 2017.

[19] S. Yan, Y. Donghua, W. Haolei, “Improvement of K-means based on particle swarm clustering algorithm,”Computer Engineering and Applications, vol. 50, no. 21, pp.125-128, 2014.