ATLANTIS PRESS
Advances in Computer Science Research, volume 71
4th International Conference on Machinery, Materials and Information Technology Applications (ICMMITA 2016)

Cloud Computing K-Means Text Clustering Filtering Algorithm based on Hadoop

Huang Suyu
School of Computer Science, Wuhan Donghu University, Wuhan Hubei, 430212

Key words: clustering; K average; text; cloud computing; big data; filtering

Abstract. the partition and hierarchy methods are the most popular clustering technology of the clustering algorithm. Providing that the k-means is sensitive to the initial clustering center and is likely to become partially optimal, an advanced clustering algorithm based on the partial swarm is presented in this essay through determining the number of clusters and the initial clustering center dynamically with the method shown in Literature [1] combined with the method of Literature [2], so as to optimize the normalization of sample set, weight adjustment of particle swarm, computation of dissimilarity matrix and colony fitness variance. Through this algorithm, the initial clustering center is determined through the density and the max/min distance to eliminate k-means being sensitive to the initial value and partially optimal. The colony fitness variance is introduced through normalization of the dimension properties of sampling set to work out the further optimized hybrid algorithm. According to the test results, this algorithm is featured with higher accuracy and stronger convergence ability.

Introduction

Clustering is an unsupervised learning widely used in many fields, including data mining, pattern recognition, computer vision and bioinformatics as one of the important methods for data analysis. Clustering refers to dividing the data set into multi sets according to its dependency in order to let the data points in the same set more similar to each other while the data points of different sets less similar to each other [1]. The partition and hierarchy methods are the most popular clustering technology of the clustering algorithm. The clustering algorithm based on partition mainly includes k-means and its optimization algorithm; and the clustering algorithm based on hierarchy mainly includes UPGMA and its improved algorithm. In 1967, MacQueen proposed the k-means algorithm based on objective function, which has become a representative partition method [2] featuring simple, quick and effective treatment of big data set. However, k-means is sensitive to the initial clustering center as the different clustering centers will cause different clustering results, moreover, this method is likely to be partially optimal. The rough set is introduced in Literature [3] for effective treatment of boundary fussy data, besides, the core concept is introduced to improve the clustering accuracy [4].

Text Algorithm

Standardization of data set

The data preprocessing consists of selection of quantity, type and scaling of features which are relied on feature selection and extraction. For the feature selection, an important feature is selected for the representation, and for feature extraction, the input feature will be converted into a new typical feature. The two methods are usually used to obtain a proper feature set in order to avoid clustering of curse of dimensionality. The data preprocessing also excludes the outlier from the data set. The outlier is the data not attached to normal data behaviors or models, which may lead to clustering results with deviation. Therefore, the outlier needs to be eliminated in order to obtain the correct clustering.

Providing that the metric and meaning for byte property of most data sets are different, and the change scope of relevant data value is relatively large. If the Euclidean distance is directly used for computation, the property of relative large value will have a great impact on the dissimilarity of

sample set, leading to consequence on the correctness of the clustering. According to Literature [7], the standardization of data set is carried out by linear standardization "range" treatment, to be specific, the standardization of dimensions are conducted as shown in Eq. (1), and the results will be spread to the range of [0, 1], and as a result the impact of data fluctuation of the dimensions on the clustering correctness will be reduced.

$$m' = \frac{m - \min_m}{\max_m - \min_m} \tag{1}$$

Wherein, $\min_m$ is the minimum value of the property m, and $\max_m$ is the maximum value of the property m.

Particle swarm and its optimization algorithm

The particle swarm algorithm is a kind of evolutionary computation method proposed by the U.S. social psychologist James Kennedy and electric engineer Russel Eberhart in 1995 based on the clustering creatures in order to simulate the social behaviors. The particle swarm algorithm is initialized into a group of random particles where every potential solution is a particle in the searching space owning the specific speed, position and adaptability, and the optimal solution will be worked out through iteration. In each iteration process, the particle is self-updated through renewing the two "extreme values": the first is the optimal solution found out by the particle itself called as individual extreme value pbest; the second is the extreme value found out by the whole particle swarm called as global extreme value gbest. The speed and position of each particle are updated through the following equations:

$$\begin{aligned} v_t &= wv_{t-1} + c_1 rand_1() \cdot (pbest - x_{t-1}) \\ &+ c_2 rand_2() \cdot (gbest - x_{t-1}) \end{aligned} \tag{2}$$

$$x_t = x_{t-1} + v_t \tag{3}$$

Wherein: v_t is the current speed of the particle, x_t is the current position of the particle. c_1 and c_2 are acceleration constants normally as $c_1 = c_2 = 2$ i. $rand_1()$ and $rand_2()$ is the random numbers ranged in [0, 1]. The initial position and speed of each particle are determined randomly.

Among the adjustable parameters for particle swarm algorithm, the inertia weight is one of the important parameters, in which the large value is used for quickening the global searching process of PSO front section and the small value for quickening the partial searing process of PSO rear section. The early convergence of PSO algorithm is very quick, so the global searching shall be replaced by partial searching in order to improve the operation efficiency of the algorithm. According to Literature [5], the ideal effect will be obtained if the exponential decrease of non-linear weight decrease strategy is used for the inertia weight, and the curve variation equations are:

$$w = w_e \left(w_s / w_e \right)^{\frac{1}{\left(1 + \frac{ct}{t_{\max}}\right)}} \tag{4}$$

Wherein, w_s and w_e are respectively 0.95 and 0.4, c is 10, t is iterations and $t_{\max}$ is the maximum iterations. In the early days, the PSO is made to enter local search as soon as possible through accelerating the decrease speed of inertia weight to obtain higher solution efficiency.

Improved k-means Clustering Algorithm

k-means initialization method

Firstly set the density parameters (MP and ε) to ensure the data point in a high density area, i.e. what meets the data point in ε field is at least a data point of MP to get a high density area D. Choose an object in a high density area from D as the 1st clustering center C_1, choose the high

density point farthest from C_1 as the 2nd clustering center C_2, then calculate the distances between any data point Xi in area D and C_1 and C_2 - $d(X_i, C_1)$, $d(X_i, C_2)$, then the 3rd clustering center C_3 is $\max(\min(d(X_i, C_1), \min(d(X_i, C_2)))$ and the data point is Xi.

Therefore, the Kth clustering center shall meet the data point Xi of

$$\max(\min(d(X_i, C_1), \min(Xi, C_2), \cdots, \min(d(X_i, C_k)))) \tag{5}$$

The rest can be done in the same manner, so the Kth initial clustering center can be automatically determined.

Dissimilarity matrix

The dissimilarity matrix among n data points of data set is defined as s by literature [11][12], which is $n \times n$ matrix

$$D = \begin{bmatrix} 0 & d(1,2) & d(1,3) & \cdots & d(1,4) \\ d(2,1) & 0 & d(2,3) & \cdots & d(2,n) \\ d(3,1) & d(3,2) & 0 & \cdots & d(3,n) \\ \vdots & \vdots & \vdots & 0 & \vdots \\ d(n,1) & d(n,2) & d(n,3) & \cdots & 0 \end{bmatrix} \tag{6}$$

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + \cdots + |x_{id} - x_{jd}|^2)} \tag{7}$$

As for any data point, the dissimilarity degree between x_i and x_j is $s(i, j)$; the lower the value is, the higher the dissimilarity degree between two data points is; the higher the value is, the lower the dissimilarity degree is. $d(i, j) = d(j, i), d(i, i) = 0$. Therefore, the dissimilarity matrix is symmetrical and reflexive.

Determination of the best time for PSO algorithm and K-means algorithm operation

To combine PDO and k-means organically, the operation time of k-means shall be determined. According to literature [13], it can be seen that it's not necessary to carry out k-means calculation during universal search of PSO and that the k-means calculation is started when PSO enters convergence condition, thus to start local search, which can effectively improve the operating efficiency of mixed algorithm and shorten the time required for algorithm running. The convergence starting point of PSO is determined and judged with the overall change of particle fitness.

Definition: overall fitness variance of particle swarm

$$\sigma^2 = -\sum_{i=1}^{n} \left(\frac{f_i - f_{avg}}{f} \right)^2 \tag{8}$$

Wherein, n is the number of particles, f_i is the fitness value of particle i, f_{avg} is the average fitness of particle swarm at present; when $\sigma^2 < m$, m is a certain confirmed threshold value, showing that PSO has entered the convergence phase when K-means algorithm starts to be carried out.

Experimental Results and Analysis

Simulation environment: software: operating system Windows XP, compile software Matlab7.0.1; hardware: Intel（R）Core(TM) i5-2450M CPU @ 2.5GHz, memory: 4GB.

Data sets for experiment: adopt artificial data sets and Iris and Wine data sets in UCI for test; Iris data set has 3 types in total, each type having 50 samples and each sample having 4 properties in total. Wine data set has 3 types in total which have 178 samples, each sample having 13 proprieties.

The parameter setting is as follows: the number of particle swarm is 20 and the maximum iterations is 50 times; learning factor $c_1 = c_2 = 2$, $w_s = 0.95$, $w_e = 0.4$, $c = 10$, and the threshold

value of overall fitness variance of particle swarm $m = 2$. It can be seen from literature [16], in the tests of the artificial data set and Iris data set, ε=2.35 and MP=37.74, and in the test of wine data set, ε=310.62, MP=47.84, and the effect is relatively good. The three algorithms operate for 10 times to calculate their own average values to reduce the influence of error on clustering results.

Algorithm comparison

To verify the effectiveness and feasibility of algorithms in the paper, artificial data sets randomly generated are adopted to carry out tests. There are 150 artificial data sets generated, which are classified into 3 types, and the proprieties are 4 dimensions. See Fig. 1, 2 and 3 for the clustering results of K-means algorithm, PSO+k-means algorithm and the algorithm in the paper on the artificial data set respectively.

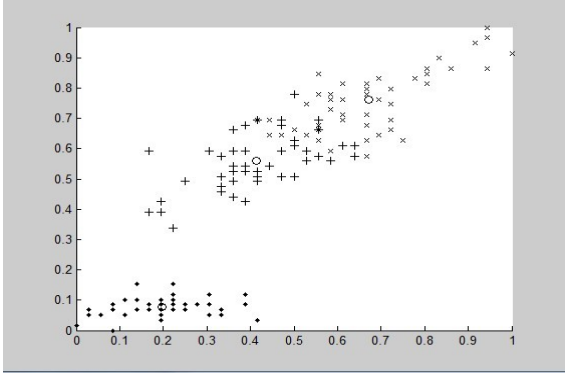

Fig. 1 K-means Algorithm

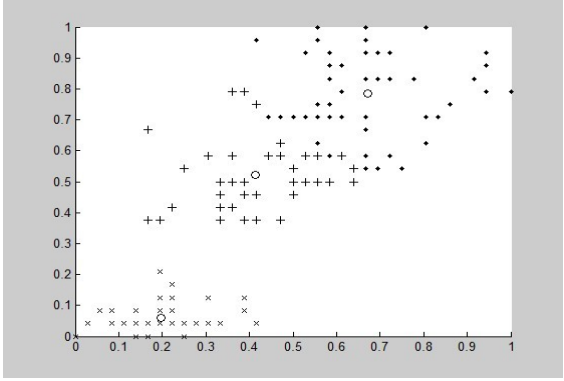

Fig.2 PSO+k-means Algorithm

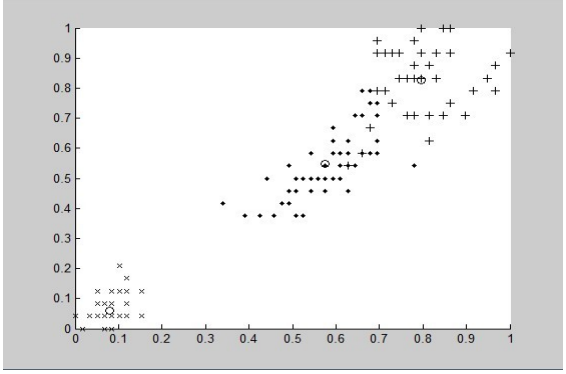

Fig.3 Algorithm in the Paper

Experimental result: the clustering accuracy of k-means algorithm, PSO+k-means algorithm and the algorithm in the paper on the artificial data set is respectively 76.68%, 80.33% and 89.76%. See Fig. 1, 2 and 3 for the simulation diagram of k-means algorithm, PSO+k-means algorithm and the algorithm in the paper on the artificial data set respectively, which intuitively shows that the clustering effect of the algorithm in the paper on the artificial data is the best.

Test of UCI standard data set

To further verify the effectiveness of the algorithm in the paper, k-means algorithm, PSO+k-means algorithm and the algorithm in the paper are tested on the Iris and Wine data sets of UCI standard data set to calculate the accuracy rate. See Table 1 for characteristic indexes of Iris and Wine data sets and see Table 2 for comparison between clustering accuracy after 3 algorithms run on two data sets. See Fig. 4 and 5 for convergence curves of fitness values of three algorithms on Iris and Wine data sets respectively.

Table 1 Information Statistics of Data Sets for the Experiment

Data set	No. of samples	No. of types	No. of proprieties	Type distribution
Iris	150	3	4	50,50,50
Wine	178	3	13	59,71,48

Table 2 Comparison between Clustering Accuracy of Three Algorithms

Algorithm / Data set	k-means	PSO+k-means	Algorithm in the paper
Iris	0.7812	0.8953	0.9285
Wine	0.6821	0.7334	0.9317

Conclusion

The algorithm in the paper determines the initial clustering center based on the density and the maximum and minimum distances, which resolves the problem that the K-means algorithm needs to be determined in advance; data sets are normalized, reducing the influence of property value fluctuation of each dimension of data set on the accuracy of clustering results. Meanwhile, by obtaining dissimilarity matrix and using the good global convergence ability of PSO, the defect of K-means algorithm that you are easily caught in the local optimum is removed. And the effectiveness of the algorithm in the paper has been proven through the experiment. However, the algorithm in the paper is highly valued on some small lower dimension data set, and further studies are needed on how to carry out effective clustering analysis on large-high dimension data.

Acknowledgement

The science and technology research project of Education Department of Hubei Province, Research and improvement of Hadoop clustering algorithm based on K_Means.

Reference

[1] Lin K, Li X, Zhang Z, et al. A K-means clustering with optimized initial center based on Hadoop platform[C]// International Conference on Computer Science & Education. 2014:263-266.

[2] Zhao W Z, Hui fang M A, University X, et al. Research on Parallel k-means Algorithm Design Based on Hadoop Platform[J]. Computer Science, 2011, 38(10):166-167.

[3] Zhao W, Ma H, He Q. Parallel K-Means Clustering Based on MapReduce[C]// International Conference on Cloud Computing. Springer-Verlag, 1970:674-679.

[4] Geng Y, Zhang L. K-Means Clustering Algorithm for Large-Scale Chinese Commodity Information Web Based on Hadoop[C]// International Symposium on Distributed Computing and Applications for Business Engineering and Science. 2015:256-259.

[5] Xu Y, Zhang Y, Ma R. K-means Algorithm Based on Cloud Computing[C]// Fifth International

Symposium on Computational Intelligence and Design. IEEE Computer Society, 2012:363-365.