새떼 이동의 모방에 의한 k-평균 군집 속도의 향상

이창영[*]

Enhancement of the k-Means Clustering Speed by Emulation of Birds' Motion in Flock

Chang-Young Lee[*]

요 약

K-평균 군집에서 수렴 속도를 향상시키기 위한 노력으로서, 우리는 새떼 이동의 개념을 도입한다. 그들 운동의 특징은 각 새가 그의 가장 가까운 이웃을 쫓아간다는 것이다. 우리는 군집 과정에 이 특징을 활용한다. 일단 한 벡터의 클래스가 결정되면, 그 근처의 몇 벡터들에게 동일한 클래스가 부여된다. 실험 결과 군집 종결에 필요한 계산 반복 횟수가 종전 방법에 비해 유의미하게 작은 것으로 나타났다. 게다가 단일 반복 계산에 소요되는 시간이 5% 이상 짧았다. 벡터와 센트로이드 사이의 거리를 누적한 값으로 군집의 품질을 평가한 바, 본 논문에서 제안한 방법과 종전 방법과의 차이는 거의 없었다. 결론적으로, 본 논문에서 제안한 방법에 의해, 보다 짧은 계산 시간으로 질적 하락 없는 군집을 수행할 수 있었다.

ABSTRACT

In an effort to improve the convergence speed in k-means clustering, we introduce the notion of the birds' movement in a flock. Their motion is characterized by the observation that each bird runs after his nearest neighbor. We utilize this feature in clustering procedure. Once the class of a vector is determined, then a number of vectors in the vicinity of it are assigned to the same class. Experiments have shown that the required number of iterations for termination is significantly lower in the proposed method than in the conventional one. Furthermore, the time of calculation per iteration is more than 5% shorter in the proposed case. The quality of the clustering, as determined from the total accumulated distance between the vector and its centroid vector, was found to be practically the same. It might be phrased that we may acquire practically the same clustering result with shorter computational time.

키워드

Clustering, K-means Clustering, Convergence Speed, Birds' Movement
군집, k-평균 군집, 수렴속도, 새의 이동

Ⅰ. Introduction

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups. It is a main task of common techniques for statistical data analysis, used in many fields including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics,

[*] Div. of Information Systems Engineering, Dongseo University(seewhy@dongseo.ac.kr)

접수일자 : 2014. 07. 08 심사(수정)일자 : 2014. 08. 21 게재확정일자 : 2014. 09. 19

and general signal processing [1-5].

Clustering is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem.

There are two major problems associated with the clustering : (1) falling into a local minimum instead of a global one and (2) computational time needed to iterative calculations of metric in multi-dimensional hyperspace.

There have been many attempts to overcome the falling into the local minimum. Application of genetic algorithm [6] or hybrid of k-means clustering algorithm and other techniques [7-9] are popular approaches.

A noble approach is based on observation and emulation of nature. Contemporary optimization algorithms, which are inspired by biology, including the wolf, firefly, cuckoo, bat and ant, simulate swarm behavior in which peers are attracted while steering towards a global objective. The ant colony optimization is an example, the aim of which is to search for an optimal path in a graph, based on the behavior of ants seeking a path between their colony and a source of food [10].

It is found that these bio-inspired algorithms have their own virtues and could be logically integrated into K-means clustering to avoid local optima during iteration to convergence [11-13].

In this paper, we concentrate on speeding up the convergence in k-means clustering. There have been lots of study on this problem [14-17]. The motivation is provided by the observation of the birds' flying in a flock.

The organization of this paper is as follows. Section Ⅱ describes the idea used in this paper. After providing the experimental procedure in Section Ⅲ, results and concluding remarks will be given in section Ⅳ and Ⅴ in turn.

Ⅱ. Clustering by Emulation of Birds' Movement

The conventional k-means clustering proceeds as follows :

1. Choose initial centroids : there might be employed several alternatives : random, picking from input vectors randomly, random numbers scaled appropriately in the input feature space.

2. Classify the class of each vector according to the chosen centroids.

3. Update class centroids according to the newly assigned vector classes.

4. Check if a convergence criterion is satisfied. If yes then stop, otherwise go to step 2.

The process of k-means clustering requires long time for big data. Our idea for speeding up the convergence begins with the observation of the birds' motion. Fig. 1 shows a group of birds flying in a flock.

Fig. 1 A group of birds flying in a flock

It is interesting to note that the flock size does not diverge indefinitely but maintains a certain amount. The secret is that a bird runs after its nearest neighbor, which is known from the birds ecology [18-19]. The nearest neighbor is changed from time to time. Though there's a certain direction driving the flock, e.g. some food, the flock is confined to a finite area. This strategy of "following the nearest neighbor" has the effect of "gathering" of the thousands of birds.

We try to utilize this feature of birds in clustering. Let's consider a vector $\mathbf{x}_k$. There are in general several vectors whose nearest vectors are commonly given by $\mathbf{x}_k$. Let's denote their indices by

$$k_i,\ i = 1 \cdots K \tag{1}$$

Once we get the class of $\mathbf{x}_k$, we assign the K vectors of (1) the same class as $\mathbf{x}_k$. In order not to repeat the class estimation for the K vectors of (1), it is necessary to tag the "class_assigned flag" as true. Before the clustering, we have to calculate the nearest neighbors for the vectors. The algorithm might be described as follows :

1. We examine the nearest vector to the vector $\mathbf{x}_i$ and suppose it be $\mathbf{x}_k$.
2. Increase the value of num_follower(k) by 1. Let follower_index(k,1)=i.
3. Repeat this work for all the vectors.

In step 2 above, num_follower(k) denotes the number of vectors whose nearest neighbors are commonly the vector $\mathbf{x}_k$ and follower_index(k,1) means the first vector "following" the vector $\mathbf{x}_k$.

In the clustering stage, we calculate the class for $\mathbf{x}_k$ and assign its follower vectors the same class. Then their "class_assigned flag" is set to "true", which saves the cost of the class calculations for them. Fig. 2 shows the block diagram of our work.

Our approach is reminiscent of the particle swarm optimization (PSO) [20] but is different in that, in PSO, the solution (particle) itself moves instead of the vectors.

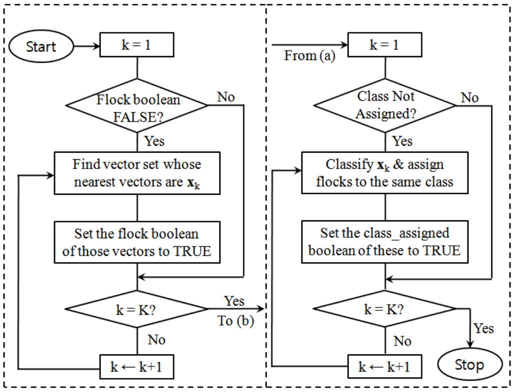
Start
k = 1
Flock boolean FALSE?
No
Yes
Find vector set whose nearest vectors are $\mathbf{x}_k$
Set the flock boolean of those vectors to TRUE
k = K?
Yes
To (b)
No
k ← k+1
From (a)
k = 1
Class Not Assigned?
No
Yes
Classify $\mathbf{x}_k$ & assign flocks to the same class
Set the class_assigned boolean of these to TRUE
k = K?
Yes
No
k ← k+1
Stop

Fig. 2 Block diagram of our work

Ⅲ. Experiment

An illustrative example would clarify the concept utilized in this paper. Fig. 3 shows an example distribution of 2-dimensional vectors.

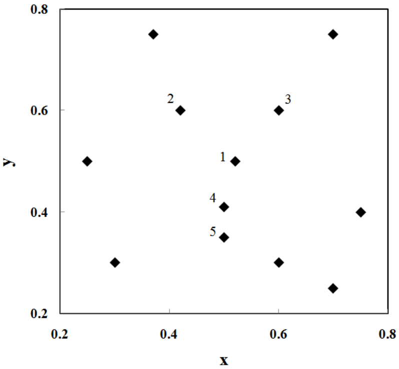

Fig. 3 An illustrative distribution of vectors to clarify the concept utilized in this paper

The vectors 2 and 3 have the vector 1 as their nearest neighbors. Therefore, once the class of the vector 1 is determined, the vectors 2 and 3 are assigned to the same class without calculation of distance metric. On the other hand, the vector 4 is assigned to the same class as that of the vector 5.

Our experiments were performed on a set of randomly generated two-dimensional vectors of numbers 20,000~50,000. Comparison between the conventional k-means clustering algorithm and the method described in the previous section was performed in two respects : the clustering quality and the convergence speed. First, the quality of the clustering result was examined by the score

$$D = \sum_k \| \mathbf{x}_k - \mathrm{c}(\mathbf{x}_k) \|^2 \tag{2}$$

where $\| \cdot \|$ denotes Euclidean norm and $\mathrm{c}(\mathbf{x}_k)$ is the centroid vector of the class to which $\mathbf{x}_k$ belongs.

In order not to have empty cluster, which incurs critical problem in clustering procedure, we slightly modify the centroid update according to the work of Pakhira [21] as follows. In usual k-means clustering, centroid update is performed by

$$\mathrm{c}_k = \frac{1}{n_k} \sum_{\mathbf{x}_j \in C_k} \mathbf{x}_j \tag{3}$$

where n_k is the number of vectors belonging to the class c_k. However, in the new scheme, the update is done by

$$\mathrm{c}_k = \frac{1}{n_k + 1} \left[\sum_{\mathbf{x}_j \in \mathrm{c}_k} \mathbf{x}_j + \mathrm{c}_k^{\mathrm{old}} \right] \tag{4}$$

In this prescription, a fictitious vector is added as if it belongs to the updated cluster. This has the effect of removing the possibility of having empty cluster, which sometimes happens due to unlucky choice of initial centroids.

Termination of the clustering was done when no change of the classes of the vectors has occurred. Convergence comparison was performed in two respects : (1) the number of iterations required for termination and (2) the elapsed time per iteration.

Ⅳ. Results and Discussion

Fig. 4 shows the total distance as defined by Eq. (2) as the iteration proceeds for conventional and proposed methods. The two cases follow practically the same pattern, which asserts that the quality of the clustering result for our proposed method is not inferior to the conventional method.

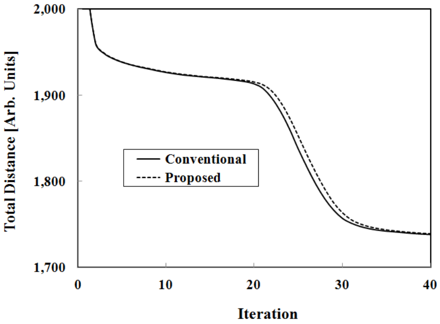

Fig. 4 Total distance vs. iteration for conventional (solid line) and proposed (dotted line) methods

In order to compare the convergence speed, we counted the number of iterations needed for termination of the centroid change for the same set of test vectors. Fig. 5 shows the result. The abscissa and ordinate represent the number of iterations for termination for the conventional and proposed methods respectively.

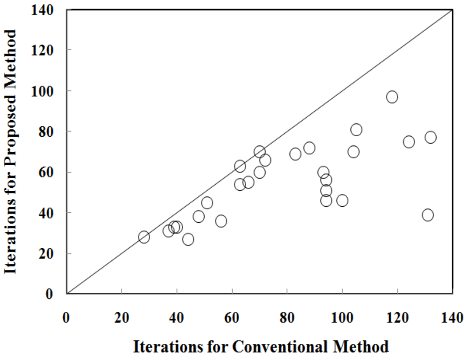

Fig. 5 The number of iterations for termination for the conventional and proposed methods

The data reside on the lower right side which means that the proposed method terminates earlier than the conventional method. In conventional k-means clustering, small fluctuations lead often long tail of iterations. In the method proposed in this paper, however, such a case is considerably reduced.

The time for each iteration has also measured. The result has shown that it is shorter by more than 5% in our case compared to the case of using conventional k-means algorithm. This should be the case by the fact that, in our case, much of the calculations of metric distance and "argmin" loop is replaced by a simple looking up a table. The results are summarized in Table 1.

Table 1. Summary of comparison between the conventional method and the one proposed in this paper

Item	Conventional k-Means	Proposed In This Paper
Total Distance	Little Difference	
Iterations	Larger	Smaller
Time / Iteration	-	More than 5% shorter

Ⅴ. Conclusion

In this paper, we studied on a method of speeding up the convergence in k-means clustering. The idea is based on emulating the birds' flying in a flock. Once a class of a certain vector is determined, a group of vectors with that vector as their nearest vectors are assigned to the same class. In this way, we could achieve savings in computational cost. From a series of experiments, the method proposed in this paper have the following properties compared to the conventional method.

1. The required iterations for termination are smaller.
2. The time of computation per iteration is shorter.
3. The clustering quality is of little difference.

References

[1] N. Krishnan and S. N. N. Sujatha, "Segmentation of cervical cancer images using active contour models," *IEEE Int. Conf. on Computational Intelligence and Computing Research*, Tamilnadu, India, Dec. 2010, pp. 1-8.

[2] T. Zhicun and L. Ruihua, "Application of Improved Genetic K-Means Clustering Algorithm in Image Segmentation," *First Int. Workshop on Education Technology and Computer Science*, vol. 2, Hubei, China, Mar. 2009, pp. 625-628.

[3] W. Zhang and H.-J. Suh, "Analysis and Simulation of Signal Acquisition of GPS Software Receiver," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 6, no. 1, 2011, pp. 27-33.

[4] C.-K. Ryu and C.-B. Park, "A Novel Clustering Method with Time Interval for Context Inference based on the Multi-sensor Data Fusion," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 8, no. 3, 2013, pp. 397-402.

[5] J.-H. Cho, "Psychology Analysis using Color Histogram Clustering," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 8, no. 3, 2013, pp. 415-420.

[6] Z. Zhu, Y. Tian, J. Xu, X. Deng, and X. Ren, "An Improved Partitioning-Based Web Documents Clustering Method Combining GA with ISODATA," *Fourth Int. Conf. on Fuzzy Systems & Knowledge Discovery*, vol. 2, 2007, pp. 208-213.

[7] Y.-G. Liu, K.-F. Chen, and X.-M. Li, "A hybrid genetic based clustering algorithm," *Proc. of*

2004 Int. Conf. on Machine Learning and Cybernetics, vol. 3, 2004, pp. 1677-1682.

[8] W. Zhang, C. Chang, H. Yang, and H. Jiang, "A Hybrid Approach to Data Clustering Analysis with K-Means and Enhanced Ant-Based Template Mechanism," *Int. Conf. on Web Intelligence and Intelligent Agent Technology*, vol. 1, Toronto, Canada, 2010, pp. 390-397.

[9] S. Wang and S. Fan, "Hybrid of k-means and Chevyshev neural network," *Int. Conf. on Automatic Control and Artificial Intelligence*, Changsa, China, Mar. 2012, pp. 1596-1600.

[10] M. Dorigo, "Optimization, Learning and Natural Algorithms," Ph.D thesis, *Milano : Italy*, 1992.

[11] T. Rui, S. Fong, X. Yang, and S. Deb, "Integrating nature-inspired optimization algorithms to K-means clustering," *7th Int. Conf. on Digital Information Management*, Macau, China, Aug. 2012, pp. 116-123.

[12] E. Saka and O. Nasraoui, "Simultaneous Clustering and Visualization of Web Usage Data Using Swarm-Based Intelligence," *20th IEEE Int. Conf. on Tools with Artificial Intelligence*, vol. 1, Dayton, OH, Nov. 2008, pp. 539-546.

[13] V. Krishnaveni and G. Arumugam, "A novel enhanced bio-inspired harmony search algorithm for clustering," *Int. Conf. on Recent Advances in Computing and Software Systems*, Chennai, India, Apr. 2012, pp. 7-12.

[14] Y. Zhao, G. Tang, D. Wei, X. Zhou, and G. Zhang, "A Clustering Algorithm Based on Probabilistic Crowding and K-means," *The 6th World Congress on Intelligent Control and Automation*, vol. 2, Dalian, China, June 2006, pp. 5892-5895.

[15] R. Salman and V. Kecman, "The effect of cluster location and dataset size on 2-stage k-means algorithm," *10th Int. Workshop on Electronics, Control, Measurement and Signals*, Liberec, Czech Republic, June 2011, pp. 1-5.

[16] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science Magazine*, vol. 315, 2007, pp. 972-976.

[17] Y. Zhu, J. Yu, and C. Jia, "Initializing K-means Clustering Using Affinity Propagation," *Ninth Int. Conf. on Hybrid Intelligent Systems*, vol. 1, Shenyang, China, Aug. 2009, pp. 338-343.

[18] W. C. Reynolds, "Flocks, herds and schools : A distributed behavioral model," *ACM SIGGRAPH Computer Graphics*, vol. 21, no. 4, 1987, pp. 25-34.

[19] C. Delgado-Mata, J. Ibanez, S. Bee, R. Ruiz, and R. Aylett, "On the use of Virtual Animals with Artificial Fear in Virtual Environments," *New Generation Computing*, vol. 25, no. 2, 2007, pp. 145-169.

[20] J. Kennedy, *Swarm Intelligence.* Eberhart : Morgan Kaufmann, 2001.

[21] M. K. Pakhira, "A Modified k-means Algorithm to Avoid Empty Clusters," *Int. J. of Recent Trends in Engineering*, vol. 1, no. 1, 2009, pp. 220-226.

저자 소개

이창영(Chang-Young Lee)

1982년 2월 서울대학교 물리교육학과 졸업(이학사)
1984년 2월 한국과학기술원 물리학과 졸업(이학석사)
1992년 8월 뉴욕주립대학교 (버펄로) 물리학과 졸업(이학박사)
1993년~현재 동서대학교 시스템경영공학과 교수
※ 관심분야 : 패턴인식, 신호처리