2011 International Conference on Modeling, Simulation and Control
IPCSIT vol.10 (2011) © (2011) IACSIT Press, Singapore

Optimal Choice of Parameters for DENCLUE-based and Ant Colony Clustering

Niphaphorn Obthong[1], Wiwat Sriphum[2]

[1] Faculty of Computer Science, Ubon Ratchathani Rajabhat University, Thailand.
[2] Faculty of Informatics, Mahasarakham University, Thailand.

Abstract. DCAnt Clustering was an algorithm in clustering the data with the 2-step procedure. That is, the initial stage started with applying DENCLUE algorithm to firstly find density and later using Ant Colony Clustering Algorithm to cluster all of the data objects in the grid board. The result gained from DCAnt Clustering was satisfactory and gave the more significant performance when compared with a K-mean. In particular, it would be able to provide more effective performance when the user applied it with the optimal σ and ξ parameter values. As a result, this study proposed the selecting procedure of the optimal σ value in order gain the more beneficial outcome.

Keywords: Ant Clustering, DENCLUE, DCAnt Clustering.

1. Introduction.

The Ant Colony Clustering was the data clustering by simulating the ant's natural behavior to cluster the data in the 2D gird board. In practical, every moment that the ant moved to the surrounding cells, it would either grab or drop the data based on the possibility and the similarity of the data required to be clustered. This process resulted as an effective data clustering. Thus, the Ant Colony Clustering was considered the fundamentals of the development of the DCAnt Clustering algorithm by firstly finding density via DENCLUE, in order that the ant could avoid unnecessary movements as well as increase the functional speed and performance of the data clustering. The process of DENCLUE Clustering required the parameter σ and ξ as key parameters of the clustering. The σ parameter value was the determiner of the effects of the proximal point and the amount of intervening data. When the σ value was recognized, the lowest density of ξ would possibly be discovered. Indeed, DCAnt Clustering randomly selected the σ value by defining the distance between σ_{max} and σ_{min} until receiving the satisfied results.

Particularly, this study aimed to empower the performance of DCAnt Data Clustering by seeking out the proper σ and ξ values for DCAnt Clustering, in order to achieve the more-effective data clustering.

2. DENCLUE (DENsity-based CLUstEring) Algorithm.

Alexander Hinneburg and Daniel A Keim [1] suggested DENCLUE from the basis of good development in statistics and pattern recognition, well-known as "Kernel density estimation" [2]. The 2 parameters supported the performance of the data clustering including the σ and ξ. In details, the σ parameter was the determiner of the effect of the proximal point; therefore, it had influence on the proximal point and the amount of intervening data. Meanwhile, the ξ parameter contained the lowest density amongst the density-attractor.

[+] Corresponding author. Tel.: + 66 87655 5322; fax: +66 4531 3868.
E-mail address: nobthong@yahoo.com.

DENCLUE Algorithm contained 5 steps as followings.

- Step 1: Creating the data board (2σ) by defining the distance between $\sigma_{\max}$ and $\sigma_{\min}$. The σ parameter provided the effects on the data within the proximal point and the amount of intervening data.
- Step 2: Verifying the cubes as a populated ($||C_p||$) in which these populated cubes depended on the determination of the σ value. Hence, the researcher explored only the cubes that contained members.
- Step 3: Undertaking the data clustering by exploring the next populated cubes, under the requirement that the cubes $c_1, c_2 \in C_p$ and the distance from *d*(*mean* (c_1), *mean* (c_2)) ≤ 4σ. These 2 cubes would be combined together and calculated for a new mean (*mean(c)*).
- Step 4: Determining the density-attractor (X^*) through the hill-climbing; the density-attractor could be found by the equation (1).

$$x = x^0, x^{i+1} = x^i + \delta \frac{\nabla \hat{f}^D_{Gauss}(x^i)}{\left\|\nabla \hat{f}^D_{Gauss}(x^i)\right\|} \tag{1}$$

- Step 5: On the final stage, the density-attractor would be connected to each other within 4σ on a basis of the data clustering (ξ).

3. Ant Colony Clustering Algorithm.

The Ant Colony Clustering was clustering by using the 2D grid board in which its size would depend on the number of data object required to be clustered, as m2 = 4n. Indeed, number of the cells on the board would be lower than number of the object ($m^2 \geq n$), and number of the ants used for clustering could be discovered by *n/3*. After that, the heap would be defined by considering the groups of over 2 objects and selecting the proper parameters as followings.

- $D_{\max}(H)$ was the widest distance of the object within the heap and could be gained from the equation (2).

$$D_{\max}(H) = \max_{O_i, O_j \in H} D(O_i, O_j) \tag{2}$$

- $O_{center}(H)$ was the center of the object group within the heap and could be found by using the equation (3).

$$O_{center}(H) = \frac{1}{n_H} \sum_{O_i \in H} O_i \tag{3}$$

- $O_{dissim}(H)$ was the object with the widest distance from the center of the heap.
- $D_{mean}(H)$ was the average distance of a single object to the center of the heap and could be found by using the equation (4).

$$D_{mean}(H) = \frac{1}{n_H} \sum_{O_i \in H} D(O_i, O_{center}(H)) \tag{4}$$

The functional procedure of the mentioned ant algorithm was that the ant would randomly walk toward the 8 surrounding cells. While walking, the ant would either grab or drop the objects along the way. The algorithm would stop working after the whole procedure was fully completed as illustrated in Figure 1.

4. DCAnt Clustering.

The function of DCAnt Clustering was divided into 2 sections including:

1. Clustering the data objects according to the area that contained the highest density through DENCLUE.

2. Applying the Ant Colony Clustering to cluster the data objects in the grid board by adjusting the equations, O_{center} and D_{mean} , as followings.

$$O_{center}(H) = \frac{1}{n_1 + \sum_{O_i \in H} w_i} \sum_{O_i \in H} o_i \cdot w_i \tag{5}$$

$$D_{mean}(H) = \frac{1}{n_1 + \sum_{o_i \in H} w_i} \sum_{O_i \in H} D(O_i \cdot w_i, O_{center}(H)) \tag{6}$$

If w_i was the weight of the clustered objects according to the highest-density area (c_i where $i=1,\ldots m$), w_i= number of the clustered objects (c_i).

```
1. Initialize randomly the ants positions,
2. Repeat
3. For each ant $ant_i$ Do
    1. Move $ant_i$,
    2. If $ant_i$ does not carry any object Then look at the 8 cells in the neighborhood of
       $ant_i$ location and possibly pick up an object,
    3. Else ( $ant_i$ is already carrying an object O) look at 8 cells around $ant_i$ and
       possibly drop O,
4. Until stopping criterion.
```

Fig. 1: Ant Colony Clustering Algorithm [3]

5. Optimal choice of parameter σ and ξ for DENCLUE.

Entropy is a measure of uncertainty [6] for spreading density by defining the σ value. If density of the data point was comparatively equal to another data point, it would increase uncertainty of the density spread and entropy. On the other hand, if the data point offered different density spreads, the uncertainty of the spread and entropy would be much decreased. As a consequence, the optimal σ parameter value by decreasing the value of entropy measure would be as the below statement.

n data points were defined in the data space $D = \{x_1, \ldots, x_n\}$ density entropy value could be found by using the equation (7)

$$H = -\sum_{i=1}^{n} \frac{f_B^D(x_i)}{Z} \log\left(\frac{f_B^D(x_i)}{Z}\right) \tag{7}$$

where $Z = \sum_{i=1}^{n} f_B^D(x_i)$ and for $\sigma \in [0, +\infty]$, density entropy (*H*) satisfies are $H \geq 0, H \leq \log(n)$ and $H = \log(n) \Leftrightarrow f_B^D(x_i) = \cdots = f_B^D(x_n)$.

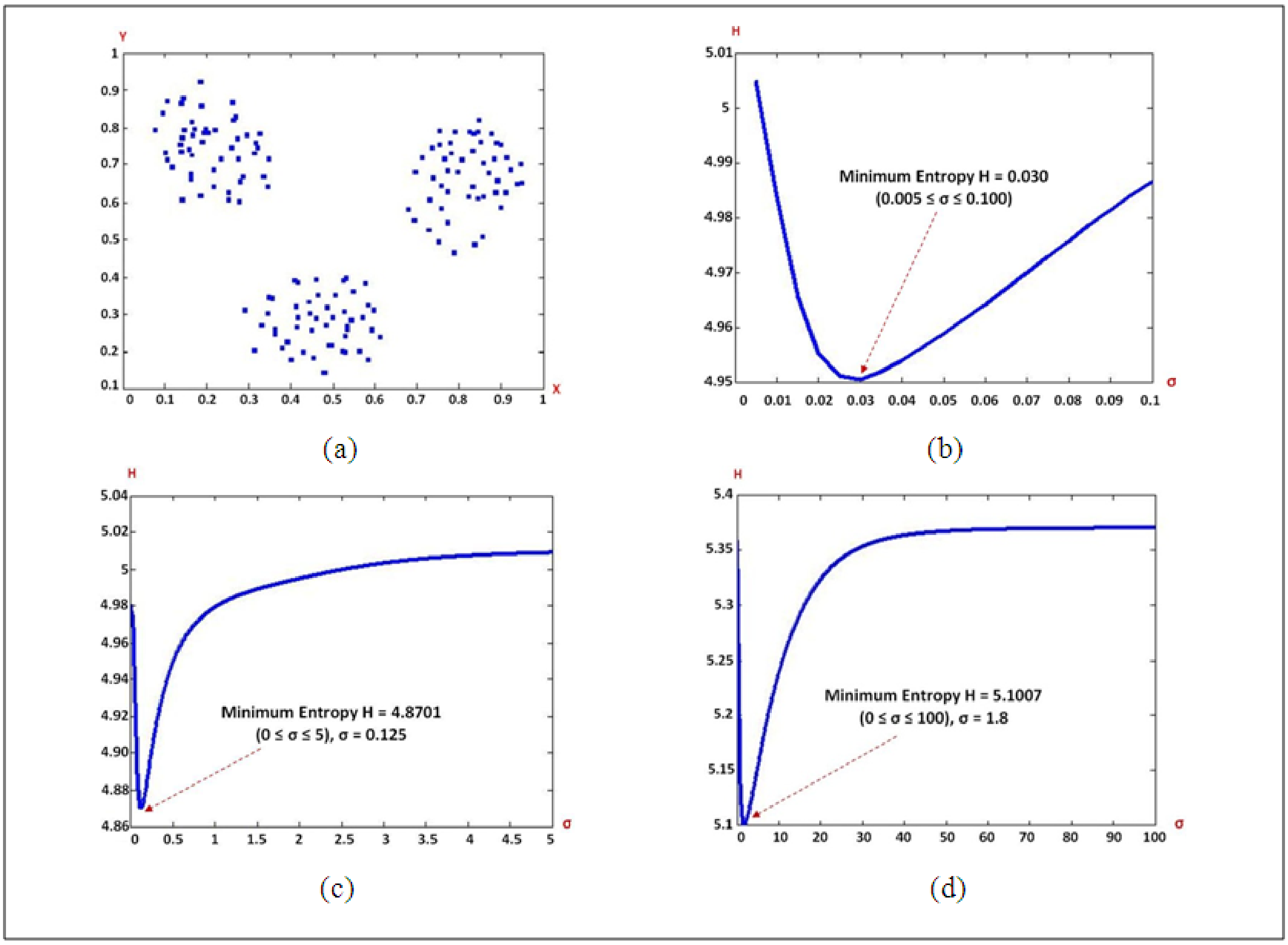
If the σ value was too low or high, the *H* value would get closer to the maximum entropy; some σ values would make *H* value the global minimum, which was in accordance with the demand of the optimal σ value. After receiving the proper σ parameter value, it would be possible to find out the minimum density level for a density attractor (ξ). Thus, the results gained from the data clustering would depend on the density threshold (ξ) which could be found by using the equation (8).

$$\xi = \|D_N\| \cdot c \cdot \sqrt{2\pi\sigma^2}^{\,d} \tag{8}$$

where $\|D_N\|$ contains numbers of noise , c is a constant $0 < c \leq 1$, and *d* is a dimensional of data.

6. Experiments and Conclusion.

In practical, this study carried out the experiment on the 3 data sets consisting of the Synthesized Data, the Iris Data, and the Data of Thyroid Patients. Figure (2-a) was a sample of Synthesized Data. In details, the Synthesized Data including 150 data sets and 2 attributes was conducted with the σ min and σmax defined as 0.005 - 0.1 whereas the lowest density entropy (*H*) = 4.9505, and the optimal σ value was 0.03 as presented in Figure (2-b). The Iris Data comprising 150 data sets and 4 attributes was conducted with the σ min and σmax defined as 0 – 5 and the lowest density entropy (*H*) = 4.8701, and the optimal σ value resulted as 0.125 as illustrated in Figure (2-c). The Data of Thyroid Patients including 215 data sets and 5 attributes was conducted with the σ min and σmax defined as 0 - 100 and the lowest density entropy (*H*) = 5.1007, and the optimal σ was 1.8 as showed in Figure (2-d).



(a) (b) (c) (d)

Fig. 2: Minimum of density entropy value gave the optimal σ

As a conclusion, the selecting procedure of the optimal σ parameter value for DCAnt Clustering Algorithm provided the outcome of the data clustering with a few-increased correctness, as illustrated in Table 1.

Datasets	**K-mean**	**AntClass**	**DCAnt**	**DCAnt(Optimal Parameter)**
Synthesis	100%	100%	100%	100%
Iris	88.67%	90.67%	91.33%	91.93%
Thyroid	84.19%	86.05%	86.51%	87.90%

Table 1: Comparing result of cluster.

7. References.

[1] Alexander Hinneburg, Daniel A. Keim. *An Efficient Approach to clustering in Large Multimedia Databases with Noise*. American Association for Artificial Intelligence, 1998

[2] JiaWei Han, Micheline Kamber. Data Mining Concepts and Techiques. Simon Fraser University, 2000.

[3] N.Monmarche, M.Slimance. *AntClass: Discovery of Clusters in numeric data by a hybridization of an ant colony with the Kmeans algorithm*. 1999, http://citeseer.nj.nec.com/monmarche99antclass.html

[4] Niphaphorn Obthong, Sratra Wongtanawasu. *A DENCLUE-based Ant Colony Clustering.* In Proceedings of The 11th National Computer Science and Engineering Conference (NCSEC), 2007.

[5] Shang Liu, Zhi-Tong Dou, Fei Li, and Ya-Lou Huang. *A new ant colony clustering algorithm based on DBSCAN.* In Proceeding of the Third International Conference on Machine Learning and Cybernatics, page 1491-1496. IEEE press, Shanghai, 2004.

[6] Wenyan Gan, Deyi Li. Optimal Choice of Parameters for a Density-Based Clustering Algorithm.

[7] Xiao-Gao Yu, Yin Jian. *A New Clustering Algorithm based on KNN and DENCLUE*. Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 Agust 2005.

[8] Ester M., Kriegel H.-P., Sander J., Xu X.:*'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databased with Noise'*, Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining, AAAI Press, 1996.

[9] Niphaphorn Obthong. *Density-based and Antcolony Clustering Algorithm*. Proceedings of Conference and Exhibition on Information Application in Training and Education, 2009.

[10] Database Machine Learning, http://www.ics.uci.edu/~mlearn/databases, May, 2011.