Landmark FN-DBSCAN: An Efficient Density-based Clustering Algorithm with Fuzzy Neighborhood*

Hao Liu, Satoshi Oyama, Masahito Kurihara, Haruhiko Sato
(Graduate School of Information Science and Technology Hokkaido University)

1. Introduction and Summary

Among various clustering techniques, the density-based clustering algorithms have several advantages as follows:

1. The number of clusters is not required before carrying out clustering.
2. The detected clusters can be represented in an arbitrary shape.
3. Outliers can be detected and removed.

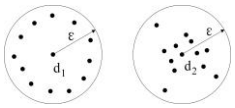
However, in general, the parameters of density-based clustering algorithms are usually difficult to select. So, in order to make the density-based clustering algorithms more robust, the extension with fuzzy set theory has attracted a lot of attentions recently. The fuzzy neighborhood DBSCAN (FN-DBSCAN) is a typical one with this idea. But FN-DBSCAN usually requires a time complexity of $O(n^2)$ where n is the number of data in the data set. This implies that the algorithm is not suitable for the work with large scale data sets.

In this study, we proposed a novel clustering algorithm called landmark fuzzy neighborhood DBSCAN (landmark FN-DBSCAN), which can provide a similar result to FN-DBSCAN, but only requires a linear time and space complexity to the size of input data set.

2. Related Work

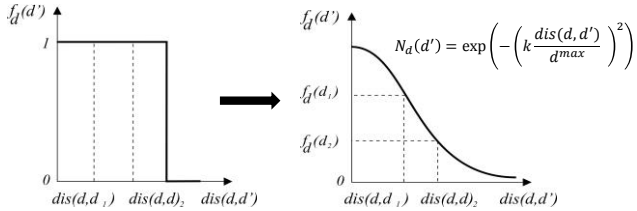
1) DBSCAN (Using a crisp neighborhood function):

$$N_d(d') = \begin{cases} 1, & if\ dis(d,d') \le \varepsilon \\ 0, & otherwise \end{cases}$$



In the right figure, data d_1 and d_2 have the same *cardinality* value which is 12, but the density values are not the same.

2) FN-DBSCAN (using a fuzzy neighborhood function)



By applying the fuzzy neighborhood function, FN-DBSCAN can provide different cardinality values for d_1 and d_2 in the above mentioned figure.

3. The Proposed Method: Landmark FN-DBSCAN

1) Landmark Concept

Given a data set $D(D = [d_{ij}]_{n\times m})$, a landmark, l, is a triplet:

$$\boldsymbol{l} = \langle \boldsymbol{V}, \boldsymbol{N_f^L}(\boldsymbol{l}), \boldsymbol{\mu} \rangle$$

- V is a m-dimensional vector referred as a data vector in D.
- $N_f^L(l) = \{d \in D | f_l(d) \ge \varepsilon_1\}$, called fuzzy neighborhood.
- μ is a positive real number, called the membership level.

2) Measuring membership degrees

- Landmark & data:

$$f_l(d) = \exp\left(-\left(r\cdot k\cdot\frac{dis(V,d)}{\Delta d^{max}}\right)^2\right)$$

- Landmark & landmark:

$$f_{l_1}(l_2) = \exp\left(-\left(k\cdot\frac{dis(V_1,V_2)}{\Delta d^{max}}\right)^2\right)$$

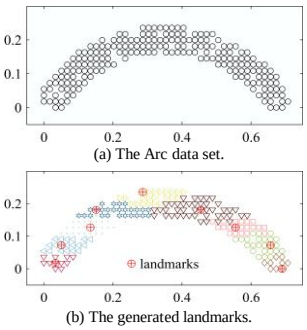
- Membership level:

$$\mu = \sum_{d\in N_f^L(l)} f_l(d)$$

3) **The landmark FN-DBSCAN Algorithm**

The algorithm consists of three steps:

I. Divide a data set into several subsets represented by generated "landmarks" (Algorithm 1).
II. Execute a modified version of FN-DBSCAN on the generated landmark set and output the index of landmarks.
III. Label data by the index of landmarks.



(a) The Arc data set.

(b) The generated landmarks.

Algorithm 1 LandmarkGeneration

Input: D, r, k, ε_1
Output: L

```
L ← ∅;
for all x in D do
    find a landmark l = (V, N, u) ∈ L, such that l.V = min{dis(l.V, x)}.
    u ← exp(−(r · k · dis(l.V,x)/Δd^max)^2);
    if L = ∅ or u < ε1 then
        V ← x; N ← {x}; u ← 1;
        l ← (V, N, u);
        L ← L ∪ {l};
    else
        l.N ← l.N ∪ {x};
        l.u ← l.u + u;
    end if
end for
```

An Example of generated landmarks applying Algorithm 1.

4) Complexity Analysis

a)Time complexity:

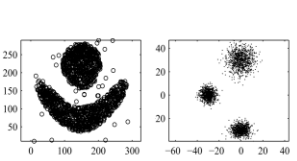
$$O(kn + k^2 + n) = O(kn + k^2) \underset{k\ll n}{\longrightarrow} O(n)$$

b) Space complexity:

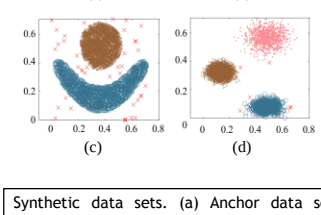
$$O(n + k + n + k) = O(n + k) \underset{k\ll n}{\longrightarrow} O(n)$$

where n is the number of data and k is the number of the generated landmarks.

4. Experimental Results:



Results of Anchor data set (r = 3). Comparison of (a) Clustering quality and (b) Clustering Efficiency.



Results of Banana data set (r = 3). Comparison of (a) Clustering quality and (b) Clustering Efficiency.

Synthetic data sets. (a) Anchor data set (2500 data). (b) Three-mixed Gaussian data set (3000 data). (c) The result for anchor data set. (d) The result for three-mixed Gaussian data set.

Gusssian Data Set (3 Clusters).

Size	Landmark FN-DBSCAN Rand-Index (r value)					FN-DBSCAN Rand-Index	Landmark FN-DBSCAN execution time(s) (r value)					FN-DBSCAN execution time(s)
	1.5	1.8	2.0	2.5	3.0		1.5	1.8	2.0	2.5	3.0	
3000	0.9953	0.9976	0.9980	0.9982	0.9987	0.9991	0.13	0.13	0.16	0.22	0.27	2.00
6000	0.9669	0.9982	0.9985	0.9990	0.9994	0.9996	0.20	0.25	0.30	0.36	0.45	6.08
9000	0.9921	0.9861	0.9814	0.9999	0.9999	0.9999	0.20	0.20	0.23	0.28	0.55	12.34
12000	0.9926	0.9928	0.9969	0.9962	0.9995	0.9996	0.25	0.28	0.31	0.36	0.44	20.55
15000	0.9983	0.9984	0.9995	0.9995	0.9995	0.9996	0.34	0.36	0.41	0.45	0.58	24.05
18000	0.9924	0.9990	0.9994	0.9996	0.9996	0.9997	0.39	0.44	0.45	0.56	0.67	38.56
21000	0.9976	0.9995	0.9996	0.9995	0.9998	0.9998	0.45	0.52	0.56	0.67	0.80	47.28
24000	0.9955	0.9998	0.9998	0.9999	0.9999	0.9999	0.53	0.59	0.66	0.75	0.92	127.11
27000	0.9943	0.9973	1.0000	1.0000	1.0000	1.0000	0.58	0.66	0.70	0.88	1.06	137.14
30000	0.9992	0.9998	0.9998	0.9998	0.9999	0.9999	0.64	0.72	0.81	0.97	1.14	218.30

References: Hao Liu, Satoshi Oyama, Masahito Kurihara, Haruhiko Sato. "Landmark FN-DBSCAN: An Efficient Density-based Clustering Algorithm with Fuzzy Neighborhood," Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.17 No.1, pp: 60-73, 2013.

* 2013 revision