1M−9

A Fast Density-based Clustering Algorithm Using Fuzzy Neighborhood Functions

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

***Abstract*: This paper aims to reduce the time complexity of the FN-DBSCAN algorithm. We propose a novel clustering algorithm called landmark FN-DBSCAN which provides good clustering qualities and has linear time and space complexities to the size of input data.**

***Keywords:* clustering, fuzzy neighborhood functions, FN-DBSCAN**

I. Introduction

Clustering (also called cluster analysis) aims to detect the homogeneous groups of data in a given data set. Numerous clustering techniques [1] have been proposed in the literature where density-based clustering techniques [2], [3] have several advantages, e.g. the number of clusters need not be known beforehand, the detected clusters can be represented in arbitrary shapes and outliers can be detected and eliminated. These advantages make the density-based clustering algorithms suitable for dealing with spatial data sets. However, they usually have difficulties in selecting appropriate parameters. Recently, the Fuzzy Neighborhood DBSCAN (FN-DBSCAN) extended the density-based clustering algorithms with fuzzy set theory, which makes density-based clustering algorithms more robust [4]. However, FN-DBSCAN requires a time complexity of $O(n^2)$, where n is the number of data in the data set, implying that FN-DBSCAN is not suitable for applications with large scale data sets. In this paper, we propose a novel clustering algorithm called landmark FN-DBSCAN. Here, 'landmark' represents a subset of the input data set, which makes the algorithm efficient with large-scale data sets. We present a theoretical analysis on time and space complexities, which indicates that they are linearly dependent on the size of the data set. The experiments presented in this paper also show that landmark FN-DBSCAN is much faster than FN-DBSCAN and provides good clustering qualities.

II. Landmark Fuzzy Neighborhood DBSCAN

In this section, the landmark FN-DBSCAN algorithm is proposed. We first present several basic concepts in Part A, then we detail the algorithm in Part B. Finally, we present an analysis on the time and space complexities in Part C.

A. Basic Concepts

Definition 1 (landmark): Given a data set D as an $n \times m$ matrix, where n is the number of data and m is dimensionality of data, a landmark (l), which is a triplet is defined as

$$l = \left\langle \mathbf{V}, N_f^L(l), \mu \right\rangle \tag{1}$$

where $\mathbf{V}$ is an m-dimensional vector equaling to a data in D (determined by Algorithm 1), $N_f^L(l)$ is a subset of D, containing all the data in the fuzzy neighborhood of l (Definition 2) and μ is a positive real number called the membership level of l (Equation (5)).

Definition 2 (fuzzy-neighborhood of a landmark): Given a set D where D can be a set of data or a set of $landmarks$, and a positive real number ε_1, the $fuzzy$-$neighborhood$ of a $landmark$, l, denoted as $N_f^L(l)$, is a set of data or a set of $landmarks$ defined by

$$N_f^L(l) = \{d \in D \mid f_l(d) \geq \varepsilon_1\} \tag{2}$$

where $f_l(d)$ can be the following two equations.

$$f_l(d) = \exp\left(-\left(r \cdot k \cdot \frac{dis(\mathbf{V}, d)}{\Delta d^{max}}\right)^2\right) \tag{3}$$

where r, k are positive real numbers and Δd^{max} is the maximum distance between $\mathbf{V}$ and all the other data in D.

$$f_{l_1}(l_2) = \exp\left(-\left(k \cdot \frac{dis(\mathbf{V_1}, \mathbf{V_2})}{\Delta d^{max}}\right)^2\right) \tag{4}$$

where k is a positive real number and Δd^{max} is the maximum distance between $\mathbf{V_1}$ and all other $landmarks$.

Based on Definition 2, the membership level of a $landmark$, μ, can be calculated by the following equation.

$$\mu = \sum_{d \in N_f^L(l)} f_l(d) \tag{5}$$

Definition 3 (cardinality of a landmark): Given a set of $landmarks$, L, where $l = \left\langle \mathbf{V}, N_f^L(l), \mu \right\rangle \in L$, the $cardinality$ of l can be calculated by

$$card(l) = \sum_{l' \in N_f^L(l)} l'.\mu \tag{6}$$

B. Algorithm

The landmark FN-DBSCAN algorithm is summarized as follows:

1) Divide a data set into several subsets represented by the generated 'landmarks'.
2) Execute a modified version of FN-DBSCAN (implemented by updating the method of calculating $cardinalities$ according to Definition 3 and by using Equation 4 as the fuzzy neighborhood function) on the generated landmark set and output the landmark index.
3) Label data according to the landmark index.

Algorithm 1 LandmarkGeneration

Input: D, r, k, ε_1
Output: L

$L \leftarrow \phi$;
for all d in D **do**
  find a landmark $l = (V, u, s) \in L$, such that $l.V = \min\{dis(l.V, d)\}$.
  $u \leftarrow \exp\left(-\left(r \cdot k \cdot \frac{dis(l.V,d)}{\Delta d^{max}}\right)^2\right)$;
  if $L = \phi$ or $u < \varepsilon_1$ **then**
    $V \leftarrow d$; $N \leftarrow \phi$; $u \leftarrow 0$;
    $l \leftarrow (V, N, u)$;
    $L \leftarrow L \cup \{l\}$;
  else
    $l.N \leftarrow l.N \cup \{d\}$;
    $l.u \leftarrow l.u + u$;
  end if
end for

C. Complexity Analysis

Theorem 1: The time complexity of landmark FN-DBSCAN is $O(kn + k^2)$, where n is the number of data and k is the number of generated landmarks.

However, in practice the number of generated landmarks is much lesser than the number of data in the data set, i.e. $k \ll n$. In this case, the time complexity of landmark FN-DBSCAN reduces to $O(n)$, which indicates that it is suitable for large-scale data sets.

Theorem 2: The space complexity of landmark FN-DB-SCAN is $O(n + k)$, where n is the number of data and k is the number of generated landmarks.

Similar to the time complexity, the space complexity will reduce to $O(n)$ if $k \ll n$.

III. Experiments

We used an artificial data set, Anchor, which contains 20000 data including noisy data, to evaluate the clustering quality and efficiency of the proposed algorithm. The clustering quality was evaluated by Rand-Index[5]. All results were compared with FN-DBSCAN.

We used eight groups of Anchor data sets with different scales. Generally speaking, both landmark FN-DBSCAN and FN-DBSCAN can provide good clustering quality, i.e two clusters can be found and noisy data can be detected. The detailed results of clustering quality with different sizes are shown in Fig. 1(a). We observe that the landmark FN-DBSCAN algorithm and the FN-DBSCAN algorithm achieved similar results, and both obtained Rand-Index values of approximately 0.99. However, there were substantial differences in their efficiencies. The time cost of the FN-DBSCAN algorithm increased very rapidly, while that of the landmark FN-DBSCAN algorithm increased slowly (Fig. 1(b), $r = 3$). For example, when the size of the data set was 2500, the landmark FN-DBSCAN algorithm saved approximately 85% of the time of FN-DBSCAN and provided almost the same quality ($r = 3$). On increasing the number of data to 20000, it saved 95.5% of the time of FN-DBSCAN and even provided a slightly better quality ($r = 3$).

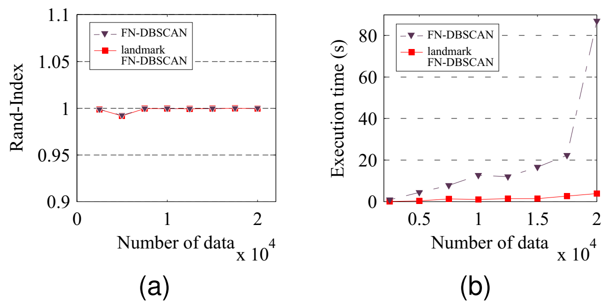

Fig. 1. Results of Anchor data set ($r = 3$). Comparison of (a) Clustering quality and (b) Clustering Efficiency.

IV. Conclusion

In this paper, we proposed a novel clustering algorithm called landmark FN-DBSCAN. We presented a theoretical analysis on the time and space complexities, which showed that both were linearly dependent on the size of data set. The experiments presented in this paper also showed that the landmark FN-DBSCAN algorithm was much faster than the original FN-DBSCAN algorithm and was able to provide a very similar clustering quality.

References

[1] R. Xu and D. C. W. II, "Survey of clustering algorithms." *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, 2005.
[2] J. Sander, "Density-based clustering," in *Encyclopedia of Machine Learning*, 2010, pp. 270–273.
[3] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 231–240, 2011.
[4] E. N. Nasibov and G. Ulutagay, "Robustness of density-based clustering methods with various neighborhood relations," *Fuzzy Sets and Systems*, vol. 160, no. 24, pp. 3601 – 3615, 2009.
[5] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "Cluster validity methods: Part i." *SIGMOD Record*, vol. 31, no. 2, pp. 40–45, 2002.