CS3FCM

In recent years, there has been growing interest in the field of safe semi-supervised clustering (S3C) among researchers. This is mainly because, in practical applications, some samples may have inaccurate labels provided by users, which can negatively impact the effectiveness of clustering methods. As a result, the main aim of safe semi-supervised clustering is to use labeled samples safely. However, S3C operates under the assumption that all labeled samples have an equally negative impact on clustering performance. To address this limitation, Gan et al. [1] introduced the confidence-weighted safe semi-supervised clustering (CS3FCM) algorithm in 2019. The main idea of this algorithm is that different samples should have varying impacts or confidences on clustering effectiveness. Initially, this algorithm uses unsupervised clustering to divide the dataset and then calculates the normalized confusion matrix based on the clustering results. The probability distribution obtained from the normalized confusion matrix is used to evaluate the safe confidence of each labeled sample, assuming that a correctly clustered sample should have a high confidence. Then, a local graph is created to show the relationship between the labeled samples and their nearest unlabeled samples, as determined from the clustering results. Finally, a trust-weighted fidelity term and a graph-based regularization term are incorporated into the objective function of the unsupervised clustering.

The objective function of the CS3FCM algorithm is written as Eq. (1):

|  |  |
| --- | --- |
|  | (1) |

Subject to:

|  |  |
| --- | --- |
|  | (2) |

In Eq. (1), represents the partition matrix and signifies the Euclidean distance between the *n*th sample and the *k*th cluster center . and represent the regularization parameters determined by the user. illustrates the fuzzy membership degrees of labeled samples, where if (where denotes the label of labeled sample ), and 0 otherwise. , which is the weight of , is defined as follows:

|  |  |
| --- | --- |
|  | (3) |

In Eq. (3), is the partition matrix obtained by FCM, and indicates the predicted cluster labels. Moreover, The map function is denoted as . The confusion matrix can be represented as:

|  |  |
| --- | --- |
|  | (4) |

where and . denotes the percentage in the ith class that be clustered into the jth class.

Given that a local graph is employed to represent the interrelation between labeled and unlabeled samples, it is imperative to identify the homogeneous nearest unlabeled neighbors corresponding to the labeled samples based on the Euclidean distance and the clustering outcomes derived from FCM. Subsequently, the edge weight of the local graph may be computed as:

|  |  |
| --- | --- |
|  | (5) |

where signifies the collection of *p* closest neighbors corresponding to , and and , respectively, denote the samples that are labeled and those that remain unlabeled.

By resolving the Lagrange equation, the updating equations (Eq. (6) to Eq. (8)) are derived for , and , correspondingly:

|  |  |
| --- | --- |
|  | (6) |

where and .

|  |  |
| --- | --- |
|  | (7) |

where and .

|  |  |
| --- | --- |
|  | (8) |

[1] H. Gan, Y. Fan, Z. Luo, R. Huang, and Z. Yang, "Confidence-weighted safe semi-supervised clustering," *Engineering Applications of Artificial Intelligence,* vol. 81, pp. 107-116, 2019/05/01/ 2019, doi: <https://doi.org/10.1016/j.engappai.2019.02.007>.