S3FCM

Recently, semi-supervised clustering (SSC) has emerged as a valuable data mining tool for uncovering intrinsic data structures by leveraging prior knowledge. Generally, SSC posits that prior knowledge can enhance clustering performance. However, if erroneous information is used, such as incorrect labels, it may negatively impact performance. To tackle this issue, Gan proposed safe-semi-supervised fuzzy c-Means (S3FCM) clustering in 2019, which is extended from the traditional semi-supervised fuzzy c-Means (SSFCM). This algorithm analyzes the information from labeled samples through unsupervised methods to mitigate corresponding risks. Consequently, it formulates a regularization expression based on unsupervised outputs to enhance the accuracy of predictions regarding labeled data. Specifically, the algorithm first segments the dataset into k clusters utilizing the traditional fuzzy c-Means algorithm, disregarding the labels. Given that the cluster labels generated by FCM may not align with the provided labels, it employs a mapping algorithm to correlate predicted cluster labels with the actual labels. This process results in a permutation matrix reflecting the relationships between the given and predicted cluster labels.

The objective function of the S3FCM algorithm is defined as the following equation:

|  |  |
| --- | --- |
| Subject to: | (1) |

In Eq. (1), the concluding term denotes the unsupervised output-based regularizer, where and represent the regularization parameters. The terminal two components impose constraints on the predictions of SSC relative to the specified labels and the labels derived from FCM, respectively [H. Gan, 2019]. The matrices and serve as partition matrix, in which represents the fuzzy membership degree of sample to cluster , and elucidating the corresponding associations between the predicted labels and the provided labels. is a label-indicating vector of length , wherein is assigned a value of 1 if is categorized, and 0 otherwise. pertains to the membership degrees of the labeled samples, where is valued at 1 if sample is affiliated with class , and 0 in all other instances. Furthermore, denotes the Euclidean distance between sample and the centroid of cluster . By resolving the Lagrange equation, the updating equations (Eq. (2) and Eq. (3)) are derived for and , correspondingly:

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

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

where in Eq. (2)

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