SSFCM-FWCW

The Semi-supervised Fuzzy C-means (SSFCM) represents a refined variant of the traditional fuzzy c-means algorithm, leveraging the principles of fuzzy logic alongside supplementary class distribution knowledge. Despite the enhancements in performance derived from the integration of this additional information into the clustering mechanism, semi-supervised fuzzy methodologies continue to encounter certain challenges. In the context of clustering, it is presupposed that all data features within the feature space possess equal significance; however, it is evident that certain features may yield more informative insights than others. Consequently, the critical issue of feature importance remains unaddressed within the semi-supervised framework. To mitigate this challenge, in 2024, Oskouei et al. [1] proposed, an advanced semi-supervised C-means fuzzy algorithm that incorporates a feature weighting and cluster weighting strategy, designated as SSFCM-FWCW. This algorithm employs a feature weighting methodology throughout the clustering iterations, applying a dynamic weight to each feature contingent upon its relevance in the formation of clusters. Moreover, in conjunction with feature weighting, a weighting scheme for clusters is also implemented, which is adaptively adjusted for each individual cluster, thereby rendering the process less sensitive to the initial center selections. By simultaneously considering both feature weighting and cluster weighting, the approach facilitates an examination of the significance of each feature across every cluster, thus promoting the establishment of an optimal clustering structure.

The objective function of the SSFCM-FWCW algorithm is as follows:

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

Subject to

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

In Equation (1), N, M, and K represent the quantity of data samples, the dimensionality of the feature space, and the number of clusters, respectively.  constitutes the partition matrix, wherein each element reflects the degree of membership of the data point to the cluster , and indicates the set of centers associated with the clusters indexed by . Furthermore, in the matrix , the entry represents the weight assigned to the *m*th feature in the *k*th cluster, is a vector, where refers to *k*th cluster weight. The parameter , constrained by , serves as a regularization factor pertinent to the second segment of the objective function that incorporates class information. The vector is a label indicator vector, where if the sample is labeled and 0 otherwise; denotes the degrees of fuzziness of the labeled samples such that when sample is classified as belonging to class , and 0 otherwise. The parameter is constrained within the range and , while the parameter is limited to the range . The specific values of the parameters and are determined by the user. The term indicates a non- Euclidean distance metric and is defined as follows:

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

where indicates inverse of the variance of *m*th feature of dataset:

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

Leveraging the closed-form solution for Eq. (1), the following update equations are obtained for , , and , respectively:

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

|  |  |
| --- | --- |
|  | (6) |
|  | (7) |
|  | (8) |

This methodology has developed a semi-supervised fuzzy clustering objective function by leveraging the principles of fuzzy logic in conjunction with existing class distribution information. However, given that the prior knowledge utilized may encompass inaccuracies, such as erroneous labels, it adversely impacts the algorithm's overall performance. On the other hand, this algorithm is not suitable for separable nonlinear data.

[1] A. Golzari Oskouei, N. Samadi, and J. Tanha, "Feature-weight and cluster-weight learning in fuzzy c-means method for semi-supervised clustering▪," 2024.