A New Approach for Semi-Supervised Clustering Based on Fuzzy C-Means

The domain of unsupervised learning is characterized by notable constraints. In contrast to the supervised learning, the outcomes of unsupervised learning merely consist of clusters or partitions, devoid of any descriptive insights regarding the generated segments. Nonetheless, users frequently seek not only these partitions but also an elucidation of the significance of those segments. To address this challenge, it becomes imperative to interpret the identified partitions, a process that poses considerable difficulty, as it necessitates the execution of intricate inferences to derive such insights. Consequently, the application of a supervised algorithm may be regarded as an alternative to the manual delineation of a data set. However, this task can be profoundly intricate, costly, labor-intensive, and reliant on the expertise of human specialists across various real-world applications. On the other hand, due to the frequent scarcity of labeled data available for training purposes, and given that supervised algorithms necessitate a substantial amount of labeled data for effective training, the application of these algorithms becomes impractical. These considerations indicate a methodology that leverages unlabeled samples to enhance the classifier's accuracy. Macario et al. (2010) [1] introduced a semi-supervised algorithm that capitalizes on the inherent knowledge within the data to guide the clustering procedure. This algorithm demonstrates promising performance in real-world applications characterized by a predominance of limited labeled data. The proposed algorithm undertakes a minimization of the defined objective function. The similarity metric employed in this objective function is the original Mahalanobis distance, which has been adapted to Euclidean distance. The formulation of the objective function for the proposed semi-supervised algorithm is articulated as follows:

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

In Eq. (1), represents the magnitude of the data set, denotes the quantity of clusters generated by the algorithm, and signifies the number of pre-existing classes. It is important to acknowledge that within the proposed algorithm, akin to the algorithm presented by Bouchachia and Pedrycz [2], a single class may be characterized by multiple clusters.  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 . The parameter represents the metric of distance between the data point and another data point . This is applicable for all indices, and . The information is utilized by the supervised binary variable , where . The value is designated as 1 if the sample is classified within cluster , and 0 in all other instances.

By solving the Lagrange equation, the following update equations for and are obtained:

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

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

[1] V. Macario and F. d. A. de Carvalho, "A new approach for semi-supervised clustering based on fuzzy c-means," in *International Conference on Fuzzy Systems*, 2010: IEEE, pp. 1-8.

[2] A. Bouchachia and W. Pedrycz, "Data clustering with partial supervision," *Data Mining and Knowledge Discovery,* vol. 12, pp. 47-78, 2006.