Rough-Fuzzy Clustering and Its Applications

Pradipta Maji

Machine Intelligence Unit
Indian Statistical Institute, Kolkata, India
E-mail: pmaji@isical.ac.in

Cluster analysis is one of the important problems related to a wide range of engineering and scientific disciplines such as pattern recognition, machine learning, psychology, biology, medicine, computer vision, web intelligence, communications, and remote sensing. It finds natural groups present in a data set by dividing the data set into a set of clusters in such a way that two objects from the same cluster are as similar as possible and the objects from different clusters are as dissimilar as possible. Hence, it tries to mimic the human ability to group similar objects into classes and categories. One of the most widely used prototype based partitional clustering algorithms is *k*-means or hard *c*-means (HCM). The hard clustering algorithms generate crisp clusters by assigning each object to exactly one cluster. When the clusters are not well defined, that is, when they are overlapping, one may desire fuzzy clusters. The fuzzy *c*-means (FCM) [1] relaxes the requirement of the HCM by allowing gradual memberships. In effect, it offers the opportunity to deal with the data that belong to more than one cluster at the same time. It assigns memberships to an object those are inversely related to the relative distance of the object to cluster prototypes. Also, it can deal with the uncertainties arising from overlapping cluster boundaries. Although the FCM is a very useful clustering method, the resulting membership values do not always correspond well to the degrees of belonging of the data, and it may be inaccurate in a noisy environment [2]. However, in real data analysis, noise and outliers are unavoidable. To reduce this weakness of the FCM, and to produce memberships that have a good explanation of the degrees of belonging for the data, the possibilistic *c*-means (PCM) has been proposed [2], which uses a possibilistic type of membership function to describe the degree of belonging. However, the PCM sometimes generates coincident clusters. Some clustering algorithms have also been proposed [8, 9], integrating both probabilistic and possibilistic fuzzy memberships.

On the other hand, the theory of rough sets deals with uncertainty, vagueness, and incompleteness. It is proposed for indiscernibility in classification or clustering according to some similarity [10]. Combining fuzzy sets and rough sets provides an important direction in reasoning with uncertainty. Both fuzzy sets and rough sets provide a mathematical framework to capture uncertainties associated with the data. They are complementary in some aspects [5].

Combining both rough sets and fuzzy sets, several rough-fuzzy clustering algorithms, namely, rough-fuzzy *c*-means (RFCM) [3], rough-possibilistic *c*-means (RPCM), rough-fuzzy-possibilistic *c*-means (RFPCM) [4], and robust rough-fuzzy *c*-means (rRFCM) [6, 7] have been proposed, where each cluster is represented by a lower approximation and a boundary region. In this regard, the talk will cover the basic notions in the theory of RFCM algorithm [3, 4], as each of the above mentioned rough-fuzzy clustering algorithms can be devised as a special case of it. The RFCM is based on both rough sets and fuzzy sets, where each cluster consists of two regions, namely, lower approximation and boundary region. While the membership function of fuzzy sets enables efficient handling of overlapping partitions, the concept of lower and upper approximations of rough sets deals with uncertainty, vagueness, and incompleteness in class definition. Each partition is represented by a set of three parameters, namely, a cluster prototype or centroid, a crisp lower approximation, and a fuzzy boundary. The lower approximation influences the fuzziness of the final partition. The cluster prototype or centroid depends on the weighting average of the crisp lower approximation and fuzzy boundary. The effectiveness of the RFCM algorithm, along with a comparison with other clustering algorithms, is demonstrated on grouping functionally similar genes from microarray data, identification of co-expressed microRNAs, and segmentation of brain magnetic resonance images using some standard validity indices.

References:

[1] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithm*. New York: Plenum, 1981.

[2] R. Krishnapuram and J. M. Keller. A Possibilistic Approach to Clustering. *IEEE Transactions on Fuzzy Systems*, 1(2):98–110, 1993.

[3] P. Maji and S. K. Pal. RFCM: A Hybrid Clustering Algorithm Using Rough and Fuzzy Sets. *Fundamenta Informaticae*, 80(4):475–496, 2007.

[4] P. Maji and S. K. Pal. Rough Set Based Generalized Fuzzy C-Means Algorithm and Quantitative Indices. *IEEE Transactions on System, Man, and Cybernetics, Part B: Cybernetics*, 37(6):1529–1540, 2007.

[5] P. Maji and S. K. Pal. *Rough-Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging*. Wiley-IEEE Computer Society Press, New Jersey, 2012.

[6] P. Maji and S. Paul. Robust Rough-Fuzzy C-Means Algorithm: Design and Applications in Coding and Non-Coding RNA Expression Data Clustering. *Fundamenta Informaticae*, 124:153–174, 2013.

[7] P. Maji and S. Paul. Rough-Fuzzy Clustering for Grouping Functionally Similar Genes from Microarray Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(2):286–299, 2013.

[8] F. Masulli and S. Rovetta. Soft Transition from Probabilistic to Possibilistic Fuzzy Clustering. *IEEE Transactions on Fuzzy Systems*, 14(4):516–527, 2006.

[9] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek. A Possibilistic Fuzzy C-Means Clustering Algorithm. *IEEE Transactions on Fuzzy Systems*, 13(4):517–530, 2005.

[10] Z. Pawlak. *Rough Sets: Theoretical Aspects of Reasoning About Data*. Dordrecht, The Netherlands: Kluwer, 1991.