An improved Fuzzy Clustering methodology applied to the study of Protein Conformational Ensembles

Duhu Man, Isabel Timón-Pérez, Jesús Soto-Espinosa, Antonio Flores-Sintas, José M. Cecilia, and Horacio Pérez-Sánchez

Dept. of Information Engineering
Graduate School of Engineering, Hiroshima University, Japan
manduhu@cs.hiroshima-u.ac.jp
Bioinformatics and High Performance Computing Research Group (BIO-HPC)
Computer Science Department
Universidad Católica San Antonio de Murcia (UCAM), Guadalupe E30107, Spain
imtimon@alu.ucam.edu
{jsoto,aflores,jmcecilia,hperez}@ucam.edu

Abstract. Clustering is a technique that aims to group data objects. Various similarity measures such as Euclidean, city-block, Mahalanobis distances and coisine similarity [1] have been used for discovering the underlying structures in data. Formally, the problem of clustering may be described as follows: Given a set of data objects $X = \{x_1, x_2, \ldots, x_n\}$, a clustering algorithm determines a suitable number k of homogeneous groups, and maps the data points to the labels in the set $C = \{1, 2, \ldots, k\}$, where each label identifies a homogeneous group of objects.
A good clustering algorithm should have the following characteristics: it should be scalable, i.e., perform well on data sets having large number of objects and also large number of attributes, should be able to determine clusters of varying shape and size, should have least requirement of domain knowledge (e.g., number of clusters, thresholds, termination condition parameters), should work well in the presence of noise and outliers, and should be insensitive to order of objects [2].
We present in this work an improved fuzzy clustering algorithm [3] that fulfills those criteria and we show its application for the classification of protein conformational ensembles, problem that arises in many domains of structural bioinformatics.

Keywords: Fuzzy Clustering, Structural Bioinformatics, HPC, Soft Computing

References

1. S. M. K. V. Tan, P.-N., *Introduction to Data Mining.* Addison-Wesley, 2005.
2. K. M. Han, J., *Data Mining: Concepts and Techniques.* San Francisco: Morgan Kaufmann, 2006.
3. A. Flores-Sintas, J. Cadenas, and F. Martin, "A local geometrical application to fuzzy clustering," *Fuzzy Sets and Systems*, vol. 100, pp. 245–256, 1998.