CLUSTER VALIDITY ANALYSIS FOR FLOOD CLUSTERING

ZHAOCHENG HOU

Post-doctorate researcher, Water Resources Department, China Institute of Water Resources and Hydropower Research,
100044 No. 20 West CheGongZhuang Road, Beijing, P. R. China
(Tel: +86-10-68785505, e-mail: houzc@iwhr.com)

Flood clustering is the process of determining flood patterns through clustering the selected typical floods. In clustering analysis, the number of clusters is a very important parameter. An ill-defined optimal number of clusters may lead to the false clustering results, which is not consistent with the real structure of the dataset. Thus, determination of the optimal cluster number is one of the research focuses in clustering analysis.

In fuzzy clustering analysis, the most widely used algorithm is the Fuzzy c-Mean (FCM) algorithm, which was proposed by Bezdek (1981) through extending the classical *within-group sum of squares* (*WGSS*) objective function. Many criteria have been proposed for clustering validity analysis. Validity criteria associated with, but not specifically designed for, the FCM algorithm began with the partition entropy (PE) (1981) and the partition coefficient (PC) (1974), which are used to measure the amount of "overlap" between clusters, with a viewpoint that a good clustering should be as clear as possible. Similar criteria may be found in the literature such as the proportion exponent (1981), the uniform data functional (1982). In spite of their simple implementation, these criteria suffer from the lack of a direct connection to the geometrical properties of data since they use only the fuzzy memberships. Taking the geometrical properties into consideration, another kind of criteria has been proposed, such as the Xie_Beni index (1991), the Pal_Bezdek index (1995), the Bensaid index (1996) and the Zahid index (1999). In their perspective, a good clustering should be made up of compact and well-separated clusters. Their performance is improved, to some extent, by the integration of fuzzy compactness within clusters and fuzzy separation between pairs of clusters. However, these criteria are highly dependent on the weighting parameter m, which is introduced by Bezdek to measure the fuzziness of the clustering. Chen (1994) established the Fuzzy Clustering Iterative Algorithm (FCIA), which tries to avoid the high dependence of the clustering results on m. Nevertheless, users of the FCIA algorithm are required to specify the number of clusters before clustering.

With a brief introduction of the FCIA algorithm, a new cluster validity criterion, which integrally considers two kinds of outputs of fuzzy clustering, is proposed in this paper. The followed experimental analysis testifies the validness of the proposed criterion. Taken 13 typical floods of the Biliuhe reservoir as an example, the validity of the proposed criterion is further testified. The genetic analysis and the clustering analysis are integrated in the flood clustering analysis of the Biliuhe reservoir, which aim to specifying the typical floods of different weather systems and then determining the relevant operating schedules toward those typical floods.

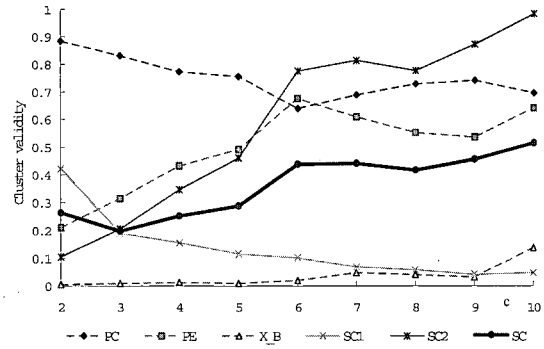

Fig. 1 Clustering result of IRIS

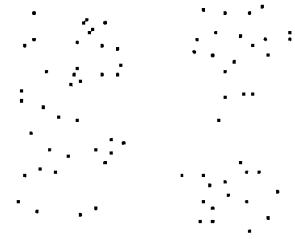

Fig. 2 Artificial dataset

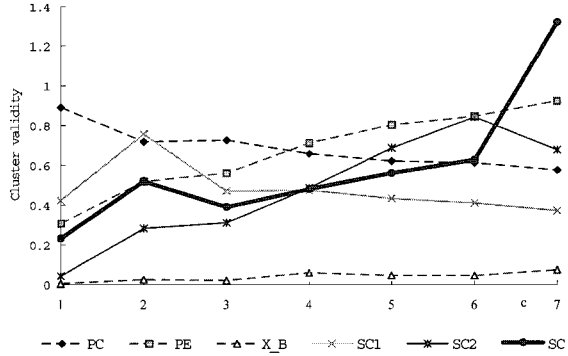

Fig. 3 Clustering result of artificial dataset

REFERENCES

J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981.

J.C. Bezdek, Cluster Validity with Fuzzy Sets. J. Cybernet., vol. 3, no. 3, pp. 58-72, 1974.

M.P. Windham, Cluster Validity for Fuzzy Clustering Algorithms. Fuzzy Set and Systems, vol. 5, pp. 177-185 1981.