Adaptive Dynamic Clustering Algorithm for Interval-valued Data based on Squared-Wasserstein Distance

Rong Guan*, Yves Lechevallier**
Huiwen Wang*

*School of Economics and Management, Beihang University, Beijing, 100191, China
rongguan77@gmail.com (R. Guan)
wanghw@vip.sina.com (H. Wang)
**INRIA-Institut National de Recherche en Informatique et en Automatique Domaine de Voluceau,
Rocquencourt B.P.105, 78153 Le Chesnay Cedex, France
Yves.Lechevallier@inria.fr

Abstract. Wide applications of interval-valued data in various domains have triggered the call for more powerful analytical tools. In light of this, this paper has presented an adaptive dynamic clustering algorithm for interval-valued data, using squared-Wasserstein distance. Experiments on both synthetic data and real data have unveiled the merits of the proposed algorithm.

1 Introduction

The technique of clustering deals with finding a structure in a collection of objects, which groups objects of similar kind into respective categories. As a main task of explorative statistical analysis, clustering has been widely used in machine learning, pattern recognition, image analysis and other fields of data mining.

The development of computer science in recent decades has enabled us to record immense amount of data. Data sets with a large number of objects are commonly seen in clustering. In some cases, however, analysts may prefer to concentrate on higher level conceptual objects rather than massive and too-specific individual objects. For example, it makes more sense to perform clustering on consumer groups, say male and female, or the young and the old, in order to analyze their buying behaviors. Potential applications also exist in complex-structured database or privacy-preserved census data, where conceptual observations should be employed to prevent identification of specific individuals. Symbolic Data Analysis (Diday, 1989; Bock and Diday, 2000; Billard and Diday, 2003; Diday and Noirhomme-Fraiture, 2008) has directed an innovative way for solving this problem. The technique aims to generalize large-scale individuals to conceptual objects described by symbolic data, such as categorical multi-valued data, interval-valued data, modal data, etc., and to extend classical statistical methods or develop new approaches for multivariate analysis on symbolic data. As a main topic in symbolic data analysis, clustering methods on symbolic data, especially on interval-valued data, has aroused much attention in recent years (Diday and Brito, 1989; Bock, 2002; Chavent et al., 2006; De Carvalho, 2007; Costa et al., 2010).