Expert Systems with Applications 36 (2009) 5761–5767

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Constraint-based clustering and its applications in construction management

Ying-Mei Cheng [a,*], Sou-Sen Leu [b,1]

[a] Department of Civil Engineering, China University of Technology, 56 Hsing-Lung Road, Section 3, Taipei 116, Taiwan, ROC

[b] Department of Construction Engineering, National Taiwan University of Science and Technology, 43 Keelung Road, Section 4, Taipei 10672, Taiwan, ROC

ARTICLE INFO

Keywords:
Constraint-based clustering
Construction management
k-Means
k-Prototypes
Affinity diagram

ABSTRACT

Both mixed data types and cluster constraints are frequently encountered in the classification problems of construction management. For example, in a bridge let project, engineers generally group the bridges into several subgroups based on their proximities, structure type, material, etc. Moreover, constraints may be set for each cluster to ensure the project's overall effectiveness. In this study, an effective clustering algorithm – the constrained k-prototypes (CKP) algorithm – is proposed to resolve the abovementioned problems. Several tests and experimental results have shown that CKP cannot only handle mixed data types but also satisfy user-specified constraints. In order to demonstrate the applicability of CKP, it is also applied to real-world problems in construction management.

1. Introduction

During the life cycle of infrastructure construction, engineers sometimes need to group similar objects into a cluster, and the similarities among the object being grouped are defined by their nature. Typical examples of the application of such clustering include the determination of the appropriate construction contractors in a highway construction project and the division of maintenance contracts in a city bridge maintenance project. Clustering techniques are appropriate for resolving these grouping problems. There are numerous scientific and practical clustering applications, for example, anthropology, disease classification, pattern recognition, and document retrieval (Dunham, 2003).

Both mixed data types and cluster constraints are frequently encountered in the classification problems of construction management. For example, in a bridge let project, engineers generally group the bridges into several subgroups based on their proximities, structure type, material, maintenance costs, etc. The attribute of the bridge structure type is categorical but the attribute of the maintenance cost is numerical. Moreover, a typical constraint is determining ways to limit the budget of each let project to a certain range. Owing to the limited capacity of the contractor, it is reasonable to set a ceiling budget to ensure construction quality.

The primary objective of this research is to develop an effective clustering method (constrained k-prototypes, CKP). The CKP algorithm can simultaneously handle user constraints and mixed data types and is coded by using the MATLAB 6.5 programming language. Some data sets were simulated to test the performance of the CKP algorithm. We also discuss two applications used for sorting construction defects and the contract packaging of bridge maintenance inspection.

Section 2 discusses the classification problems in construction management. Section 3 reviews state-of-the-art clustering methods. In section 4, the k-prototypes algorithm is briefly reviewed. Section 5 discusses constraint-based data clustering. Section 6 evaluates the constraint-based clustering method proposed in this study. In section 7, the applications of the method in construction management are presented. Finally, section 8 concludes the paper.

2. Classification problems in construction management

Studies addressing the classification problems in construction management have been conducted since 1996. Holt (1996) applied cluster analysis to classify construction contractors. Elazouni (2006) used unsupervised-learning neural networks to classify construction contractors and emphasized the need to improve the efficiency of contractor prequalification processes. Ezekiel and et al. (1998) proposed a needs-based methodology for classifying construction clients. They modified the traditional classification of clients (private, public, and developer) for evaluating contractors who could optimally satisfy the clients. Dzeng (2006) presented an analytical model to identify design management packages. This model reduces the number of design interfaces between the participating design firms. Tsai and Yang (2004) employed the constrained fuzzy c-mean clustering algorithm to plan bridge let projects. The majority of the abovementioned investigations have adopted the clustering method to treat classification problems in construction management. All of them obtained good

* Corresponding author. Tel.: +886 2 2931 3416x2465; fax: +886 2 2934 6117.
E-mail addresses: yingmei.cheng@msa.hinet.net (Y.-M. Cheng), leuss@mail.ntust.edu.tw (S.-S. Leu).
[1] Tel.: +886 2 2733 3141x7511; fax: +886 2 2737 6606.

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.06.100

results; therefore, the clustering method has been adopted for the present study.

Most of the existing clustering algorithms can either handle mixed data types or user constraints (Bradley, Bennett, & Demiriz, 2000; Chan & Chung, 1999; He & Xiong, 2005; Huang, 1998; Li, Gao, & Jiao, 2003; Ng, 2000; Ralambondraint, 1995; Tung, Ng, Lakshmanan, & Han, 2001). Few algorithms can perform both the abovementioned functions well. For example, Bradley proposed the constrained k-means clustering (Bradley et al., 2000) that can avoid local solutions involving empty clusters; however, this algorithm cannot deal with categorical data. Huang (1998) presented the k-prototypes algorithm, which can cluster objects with mixed data types but cannot consider any user constraint. Current clustering algorithms mainly focus on the constraints at either the cluster level or instance level. Dan and Kamvar (2002), Wagstaff and Cardie (2001), and Davidson (2005) separately proposed constrained clustering with background knowledge. They emphasized the importance of pairwise constraints at the instance level on the clustering process, such as must link and cannot link. These constraints are not applicable in this study.

3. Clustering and constraint-based clustering

In general, clustering analyses can be divided into various categories based on their principles and algorithms. The classification of clustering algorithms is shown in Fig. 1 (Berkhin, 2002; Han & Kamber, 2001). Traditional clustering methods include (1) partitioning methods such as the k-medoids, CLARANS (Ng & Han, 1994), and k-means methods (Hartigan, 1975; Hartigan & Wong, 1975); (2) hierarchical methods (Jain & Dubes, 1988; Kaufman & Rousseeuw, 1990); (3) density-based methods such as DBSCAN (Ankerst, Breunig, Kriegel, & Snader, 1999; Hinneburg & Keim, 1998); and (4) grid-based methods such as CLIQUE (Agrawal, Gehrke, Gunopulos, & Raghavan, 1998), STING (Wang, Yang, & Muntz, 1997) and MAFIA (Goil, Nagesh, & Choudhary, 1999). Traditional methods have several advantages. Generally, these methods involve simple algorithms that consume only a small amount of CPU time and they can easily be applied to large database systems. However, they have a common drawback: they fail to consider the constraint problems at either the instance level or cluster level. Several authors have developed constraint-clustering algorithms (Bradley et al., 2000; Dan & Kamvar, 2002; Davidson, 2005; He & Xiong, 2005; Ng, 2000; Qian, Zhang, & Lai, 2004; Quinlan, 1993; Tsai & Yang, 2004; Wagstaff & Cardie, 2001). Constraint-based clustering is the grouping of similar objects into several clusters while satisfying certain conditions such as maintaining a fixed number of objects in each cluster. Based on the nature of the constraints and applications, Tung et al. (2001) classified the constrained clustering problem into the following four categories: the constraints on individual objects, obstacle objects as constraints, clustering parameters as constraints, and constraints imposed on each individual cluster. This paper focuses on the fourth category. The constraint in this study is defined as follows:

Let each object O_i in data set D be associated with a set of m attributes $\{A_1, A_2, \ldots, A_m\}$. The value of an attribute A_j of an object O_i is denoted as O_i $[\{A_j]]$. Let ω be a comparator function, i.e., $\omega \in \{<, \leqslant, =, \geqslant, >\}$ and c represent a numeric limit set by the user. For a cluster K_l, the constraint on K_l is denoted as $\mathrm{SUM}(\{O_i[A_j] | O_i \in K_l\})\omega c$.

4. k-Prototypes algorithm

Huang (1998) presented the k-prototypes algorithm, which provides a straightforward method to integrate the k-means and k-modes algorithms to cluster mixed-data-type objects. The objective function is defined as follows:

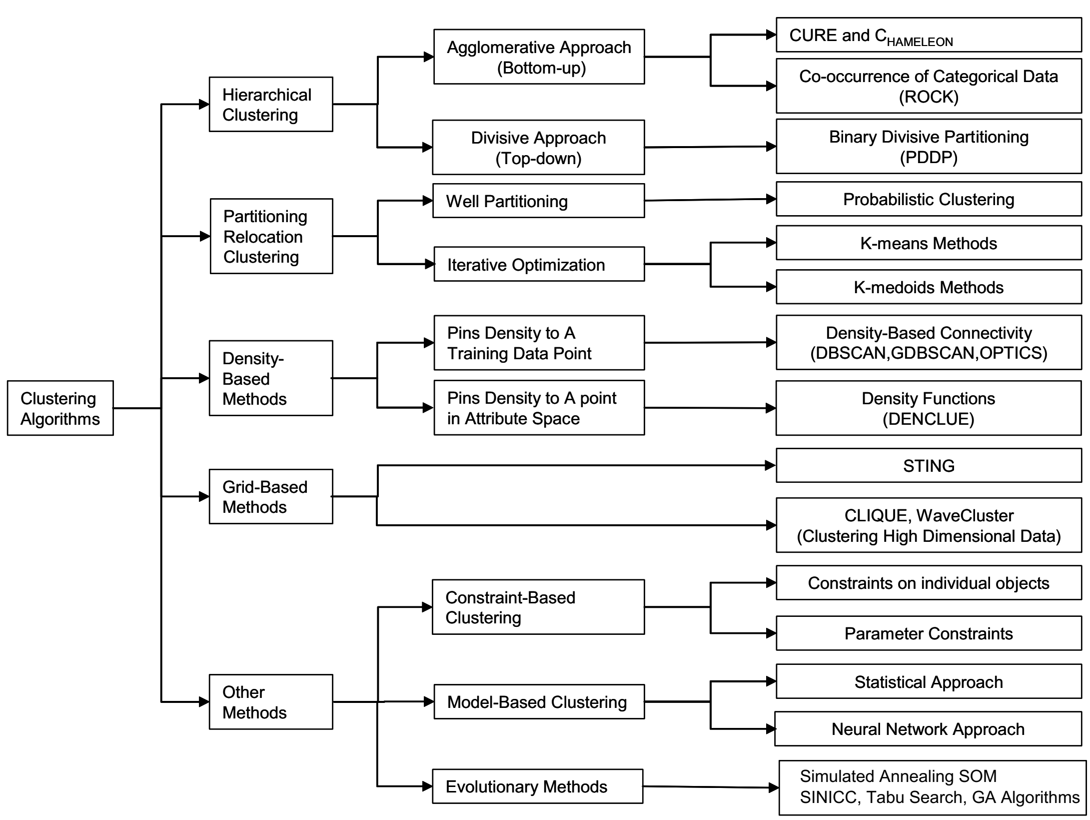

Fig. 1. Classification of clustering algorithms.