Knowledge-Based Systems 32 (2012) 101–115

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

Consensus clustering based on constrained self-organizing map and improved Cop-Kmeans ensemble in intelligent decision support systems

Yan Yang [a,b,*], Wei Tan [a,b], Tianrui Li [a,b], Da Ruan [c]

[a] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, PR China
[b] Key Lab of Cloud Computing and Intelligent Technology, Sichuan Province, Chengdu 610031, PR China
[c] Belgian Nuclear Research Centre (SCK•CEN), Mol & Ghent University, Gent, Belgium

ARTICLE INFO

Article history:
Available online 30 August 2011

Keywords:
Clustering ensemble
Semi-supervised clustering
Cop-Kmeans
Self-organizing map (SOM)
Decision support systems (DSS)

ABSTRACT

Data mining processes data from different perspectives into useful knowledge, and becomes an important component in designing intelligent decision support systems (IDSS). Clustering is an effective method to discover natural structures of data objects in data mining. Both clustering ensemble and semi-supervised clustering techniques have been emerged to improve the clustering performance of unsupervised clustering algorithms. Cop-Kmeans is a K-means variant that incorporates background knowledge in the form of pairwise constraints. However, there exists a constraint violation in Cop-Kmeans. This paper proposes an improved Cop-Kmeans (ICop-Kmeans) algorithm to solve the constraint violation of Cop-Kmeans. The certainty of objects is computed to obtain a better assignment order of objects by the weighted co-association. The paper proposes a new constrained self-organizing map (SOM) to combine multiple semi-supervised clustering solutions for further enhancing the performance of ICop-Kmeans. The proposed methods effectively improve the clustering results from the validated experiments and the quality of complex decisions in IDSS.

1. Introduction

A decision support system (DSS) is a computer-based information system that supports business or organizational decision-making activities [1]. The term of intelligent decision support systems (IDSS) describes DSS that make extensive use of artificial intelligence (AI) techniques. Along with knowledge-based decision analysis models and methods, IDSS incorporate well databases, model bases and intellectual resources of individuals or groups to support effective decision making [2,3]. Some research in AI, focused on enabling systems to respond to novelty and uncertainty in more flexible ways has been successfully used in IDSS. For example, data mining in AI that searches for hidden patterns in a database has been used in a range of decision support applications. The data mining process involves identifying an appropriate data set to mine or sift through to identify relations and rules for IDSS. Data mining tools include techniques like case-based reasoning, clustering analysis, classification, association rule mining, and data visualization. Data mining increases the "intelligence" of DSS and becomes an important component in designing IDSS.

Decision making becomes more sophisticated and difficult in today's rapid changed decision environments. Decision makers often require increasing technical support to high quality decisions in a timely manner. Among major types of IDSS, data-driven DSS [4] emphasizes access to and manipulation of a time-series of internal company data and sometimes external data. The more advanced data-driven DSS is combined with online analytical processing (OLAP) and data mining techniques (such as, spatial data mining, correlation mining, clustering, classification, and Web mining). For massive and time-variant data, *e.g.*, data from railroad sensor, data mining techniques are suitable to solve railway DSS problems in a series of datasets, which includes attributes and decisions. The calculation on these datasets clusters them and digs out relevant knowledge rules and worn-out or defective rails to avoid the derailments.

Clustering is an effective method to discover natural structures of data objects in data mining [5] and pattern recognition [6]. It refers to all the data objects that are divided into several disjunctive groups such that the similarity of objects from the same group is larger than that of objects from the different groups according to a given measure of the similarity. However, traditional clustering algorithms are defined as a kind of unsupervised learning and perform without considering any prior knowledge provided by real world users. These algorithms usually tend to classify the data objects by different ways of optimization and criteria. Many improved clustering algorithms have been proposed, but they are

* Corresponding author at: School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, PR China.
E-mail addresses: yyang@swjtu.edu.cn (Y. Yang), tanwei1103@126.com (W. Tan), trli@swjtu.edu.cn (T. Li), druan@sckcen.be, da.ruan@ugent.be (D. Ruan).

0950-7051/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.08.011

not easy to be used as a single algorithm to explore variety of structures of data objects. If the algorithm is not well suited for the dataset, it will result in a worse clustering. In recent years, semi-supervised clustering and clustering ensemble have been emerged as powerful tools to solve the above-mentioned problems.

Inspired by multiple classifiers ensemble, clustering ensemble [7–13] has been proved to improve the performance of traditional clustering algorithms. It integrates multiple clustering components generated by different algorithms, the same algorithm with different initialization parameters and so on. The final consensus clustering with its higher stability and robustness is obtained after such a combination. Establishing consensus functions is the key issue for clustering ensemble. Fred and Jain [8] explored an evidence accumulation clustering approach with the single-link and average-link hierarchical agglomerative algorithms. Their approach maps the clustering ensemble into a new similarity measure between patterns by accumulating pairwise pattern co-associations. Strehl and Ghosh [9] used three graph-based partitioning algorithms such as CSPA, HGPA and MCLA to generate the combined clustering. Zhou and Tang [10] employed four voting methods to combine the aligned clusters through the selective mutual information weights. Ayad and Kamel [11] sought cumulative vote weighting schemes and corresponding algorithms to compute an empirical probability distribution summarizing the ensemble. Yang et al. [12] improved an ant-based clustering algorithm to produce multiple clustering components as the input of an Adaptive Resonance Theory (ART) network and obtained the final partition. Wang et al. [13] proposed Bayesian cluster ensembles, a mixed-membership generative model to obtain a consensus clustering by combining multiple base clustering results.

Semi-supervised clustering [14–20] algorithms obtain better results using some prior knowledge, which is often represented by seeds or pairwise constraints. The seeds give directly the class labels of data objects. The pairwise constraints indicate whether a pair of objects is classified into the same group (must-link, ML) or different groups (cannot-link, CL). The recent semi-supervised clustering algorithms consist of two types: the constraint-based methods and distance-based methods. Seeded-Kmeans and Constrained-Kmeans proposed by Basu et al. [14] both utilize seeds information to guide the clustering process. For Seeded-Kmeans, the seeds information is only used to initialize cluster centers. For Constrained-Kmeans, the seeds labels are kept unchanged during the iteration step besides the initialization of cluster centers. Wagstaff et al. [15] proposed the Cop-Kmeans algorithm, where Must-Link and Cannot-Link constraints are incorporated into the assignment step and cannot be violated. Basu et al. [16] proposed the PCK-Kmeans to impose the violation penalty on the objective function of K-means [21], and two kinds of constraints were violated by adding the penalty. Xing et al. [17] employed metric learning techniques to get an adaptive distance measure based on the given pairwise constraints. Zhu et al. [18] extended balancing constraints to size constraints, *i.e.*, based on the prior knowledge of the distribution of the data, and assigned the size of each cluster to find a partition that satisfies the size constraints. Zhang and Lu [19] proposed a kernel-based fuzzy algorithm to learn a cluster from both the labeled and unlabeled data. Abdala and Jiang [20] pre-

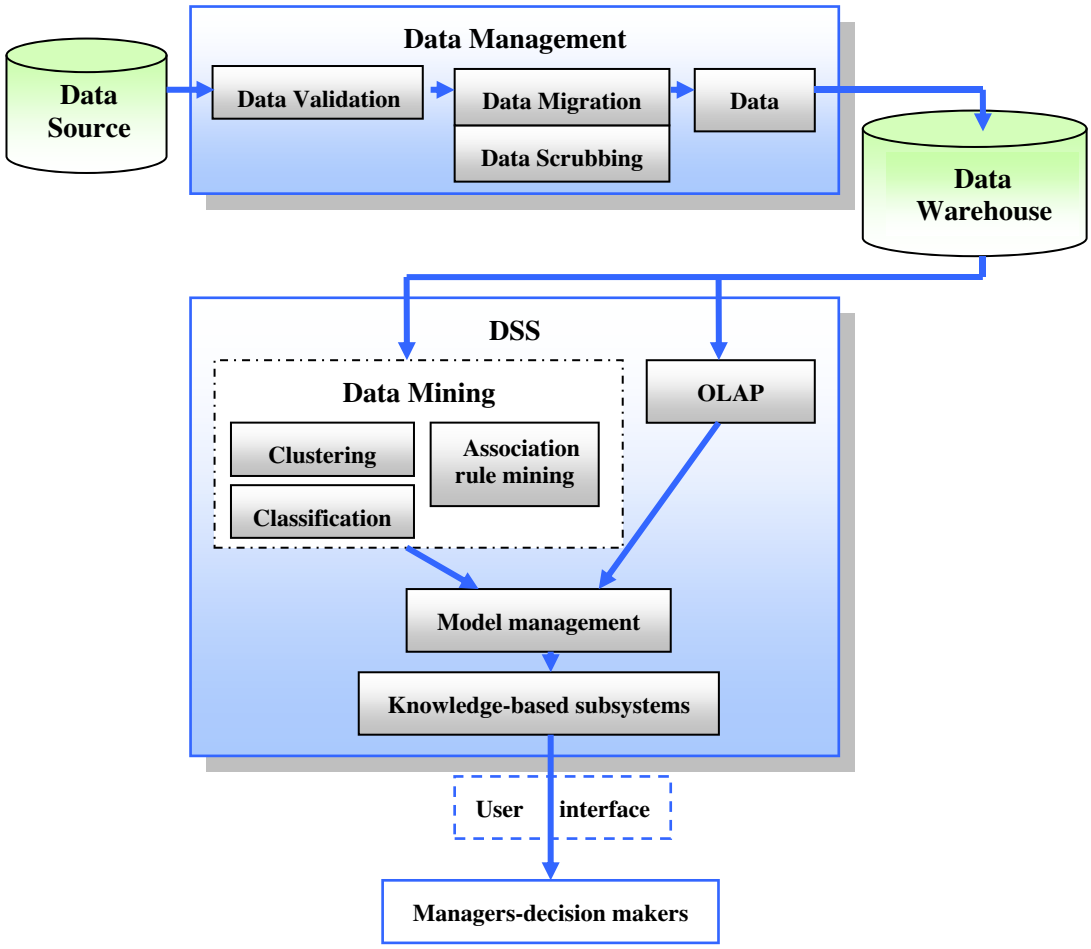

Fig. 1. A framework of data-driven DSS.