ABSTRACT

Clustering is an important tool for data exploration. Several clustering algorithms exist, and new algorithms are frequently proposed in the literature. These algorithms have been very successful in a large number of real-world problems. However, there is no clustering algorithm, optimizing only a single criterion, able to reveal all types of structures (homogeneous or heterogeneous) present in a dataset. In order to deal with this problem, several multi-objective clustering and cluster ensemble methods have been proposed in the literature, including our multi-objective clustering ensemble algorithm. In this chapter, we present an overview of these methods, which, to a great extent, are based on the combination of various aspects of traditional clustering algorithms.

INTRODUCTION

Clustering is an important tool for the exploration of datasets with no or very little prior information (Jain & Dubes, 1988). Several clustering algorithms have been proposed in the literature. They have been very successful to solve a large number of real problems in areas as diverse as biology, medicine, engineering, marketing, and remote sensing. A recent area in which cluster analysis has contributed with very important results is bioinformatics (Narayanan, 2005; Wang, Zaki, Toivonen & Shasha, 2003). In bioinformatics, researches related to functional genomics and gene
Cluster Ensemble and Multi-Objective Clustering Methods

expression data analysis are employing clustering techniques to find clusters of genes (e.g., for the identification of gene functions) and clusters of tissue samples (e.g., for discovering groups and subgroups of cancer) (Costa, Carvalho & Souto, 2004; D’haeseleer, 2005; Eisen, Spellman, Brown & Botstein, 1998; Golub et al., 1999; Hautaniemi et al., 2003; Nikkilä, Törönen, Kaski, Venna, Castrén & Wong, 2002; Pal, Aguan, Sharma & Amari, 2007; Spellman et al., 1998; Tamayo et al., 1999).

In spite of this success, traditional cluster analysis algorithms present several limitations that need to be addressed. For instance, there is no precise definition of cluster. As a consequence, the literature describes a large number of clustering algorithms, each one looking for clusters according to a given clustering criterion, or cluster definition (Law, Topchy, & Jain, 2004).

Another problem is the fact that there is no clustering algorithm, optimizing a single criterion and able to reveal all types of structures (homogeneous or heterogeneous) present in a dataset. In fact, it is unfeasible to establish, a priori, which clustering criterion is more appropriated to capture the true structure. It is also important to mention that clustering algorithms can find structures at various refinement levels, depending on their parameter settings (Jain & Dubes, 1988).

Moreover, the same data can have more than one relevant structure, each one in accordance with a different cluster definition and/or with a different refinement level (e.g., number of clusters). For example, in functional genomics and gene expression data analysis, data usually have multiple meaningful interpretations; genes can fit into more than one functional category, or a disease such as cancer may have different subtypes depending on the required level of investigation (Alizadeh et al., 2000; Golub et al., 1999; Yeoh et al., 2002).

By contrast, the usual application of cluster analysis for the exploration of a dataset focuses on the discovery of only a single structure that best fits the data (Handl & Knowles, 2005; Jain & Dubes, 1988; Xu & Wunsch, 2005). In these cases, several clustering algorithms are applied to the dataset, extracting different structures. Next, a validation method selects the structure that best fits the data. However, the search for only one best-fit structure limits the amount of knowledge that might be obtained. Besides, most validation measures are biased toward a given clustering criterion, which reduces the robustness of the analysis performed.

As an attempt to overcome the limitations of traditional clustering algorithms, finding a higher-quality solution and improving the robustness against different data conformations, recent works have proposed the use of multi-objective clustering algorithms (Handl & Knowles, 2007; Korkmaz, Du, Alhajj, & Barker, 2006; Liu, Ram, & Lusch, 2005; Naverniouk, 2005; Ripon, Tsang, & Kwong, 2006) and cluster ensemble methods (Ayad & Kamel, 2003; Boulis & Ostendorf, 2004; Fern & Brodley, 2004; Fred & Jain, 2005; Law et al., 2004; Monti, Tamayo, Mesirov, & Golub, 2003; Strehl & Ghosh, 2002; Topchy, Jain, & Punch, 2004; Topchy, Jain, & Punch, 2005). In fact, any area that benefits from using cluster analysis can take advantage of using cluster ensemble or multi-objective clustering methods. Applications of the first started in domains like network anomaly detection (Munson & Caruana, 2006), document clustering (Greene, Tsyymbal, Bolshakov, & Cunningham, 2006), image analysis (Lourenço & Fred, 2005), medical diagnostics (Greene, Tsyymbal, Bolshakov, & Cunningham, 2006), and bioinformatics (Asur, Parthasarathy, & Ucar, 2006; Hu, 2006; Hu, Yoo, Zhang, Nanavati, & Das, 2005; Monti et al., 2003; Souto, Araujo, & Silva, 2006). Multi-objective clustering applications started in the following applications: marketing (Liu et al., 2005), computer networks (Cheng, Cao, Wang, & Das, 2006), and bioinformatics (Handl, 2006; Mitra & Banka, 2006).

Despite their advantages when compared with traditional clustering methods, each previous ap-