Entropic-Genetic Clustering

Mihaela Breaban*, Henri Luchian*, Dan A. Simovici**

*Univerity of Jassy, Dept. of Computer Science, Jassy, Romania,
e-mail:pmihaela,hluchian@infoiasi.ro
**Univ. of Massachusetts Boston, Massachusetts 02125, USA,
e-mail:dsim@cs.umb.edu

Abstract. This paper addresses the clustering problem given the similarity matrix of a dataset. We define two distinct criteria with the aim of simultaneously minimizing the cut size and obtaining balanced clusters. The first criterion minimizes the similarity between objects belonging to different clusters and is an objective generally met in clustering. The second criterion is formulated with the aid of generalized entropy. The trade-off between these two objectives is explored using a multi-objective genetic algorithm with enhanced operators.

1 Introduction

This paper addresses the clustering problem given the similarity matrix of a dataset. A straightforward representation of the problem instance in this case is a weighted graph, having the objects as vertices and weighted edges expressing the similarity between objects. This leads to a graph clustering/partitioning problem which aims at identifying groups of strongly inter-connected vertices. A survey of graph clustering is presented in Schaeffer (2007).

A *similarity space* is a pair (S, w), where $w : S \times S \longrightarrow \mathbb{R}$ is a function such that $w(s,t) \geq 0$, $w(s,t) = w(t,s)$, and $w(s,s) = 1$. for every $s, t \in S$. A similarity space (S, w) can be regarded as a labelled graph $G = (S, E, w)$, referred to as the *similarity graph*, where the set of edges E is defined as $E = \{(s_i, s_j) \mid s_i, s_j \in S \text{ and } w(s_i, s_j) > 0\}$. If S is a finite set $S = \{s_1, \ldots, s_n\}$, the dissimilarity w is described by a symmetric matrix $W \in \mathbb{R}^{n \times n}$, where $w_{ij} = w(s_i, s_j)$ for $1 \leq i, j \leq n$.

A *k-way clustering of a finite similarity space* (S, w) is a partition $\kappa = \{C_1, \ldots, C_k\}$ of S. The sets $C_1, \ldots, C_k$ are the clusters of κ. We seek a k-way partition of S, κ such that the cut size (i.e. the sum of weights of edges between clusters in the similarity graph) is minimal, and $|C_p| \approx |C_q|$, for $1 \leq p, q \leq k$, which means that the sizes of the clusters are as equal as possible. Presentations of the state-of-the-art of graph clustering can be found in Fjällström (1998), Karypis and Kumar (1998).

The paper is structured as follows. Section 2 examines the two objectives which have to be optimized as stated in the problem definition. Section 3 provides a brief survey on the genetic algorithms for clustering with an emphasis on the multi-objective formulation; the representation and the operators we used are detailed. Section 4 presents experimental results.