A GRASP Algorithm for Clustering

J.R. Cano¹, O. Cordón², F. Herrera², and L. Sánchez³

¹Dept. of Software Engineering
University of Huelva, 21071 La Rabida (Huelva), Spain
Jose.cano@diesia.uhu.es

²Dept. of Computer Science and A.I.
University of Granada, 18071-Granada, Spain
{ocordon,herrera}@decsai.ugr.es

³Dept. of Computer Science
University of Oviedo, Oviedo, Spain
luciano@lsi.uniovi.es

Abstract. We present a new approach for Cluster Analysis based on a Greedy Randomized Adaptive Search Procedure (GRASP), with the objective of overcoming the convergence to a local solution. It uses a probabilistic greedy Kaufman initialization for getting initial solutions and K-Means algorithm as a local search algorithm. We compare it with some typical initialization methods: Random, Forgy, Macqueen and Kaufman. The new approach obtains high quality solutions for the benchmark problems.

1 Introduction

Clustering is a basic process to human understanding. The grouping of related objects can be found such diverse fields as statistics, economics, physics, psychology, biology, pattern recognition, engineering, and marketing [3,5].

The Clustering problem involves partitioning a set of entities into a given number of subsets and finding the location of a centre for each subset in such a way that a dissimilarity measure between entities and centres is minimized. K-Means is one of the most known clustering algorithms. Although it is known for its robustness, it can fall in local optimal solutions easily. It is widely reported that the K-Means algorithm suffers from initial starting conditions effects: initial clustering and instance order as shown in [4].

In this contribution, we propose a Greedy Randomized Adaptive Search Procedure (GRASP) [2] applied to the Clustering problem, using K-Means as local search procedure. Our algorithm tries to eliminate the classical problem of the K-Means algorithm, hold up solutions in local optima, by permitting a higher exploration and exploitation of the search space, with a medium computational cost.

In order to do that, the paper is organized as follows. Section 2 introduces a background on clustering, K-means and initialization approaches. Section 3 presents the GRASP approach to clustering. Section 4 shows the experiments and their analysis, and finally, Section 5 points out some concluding remarks.
2 Background

A common problem in cluster analysis is partitioning objects into a fixed number of groups to optimize an objective function-based clustering. These objects are measured along several features that characterize them. Patterns can be viewed as vectors in a high dimensional space, where each dimension corresponds to one feature.

In this section we introduce the formalization of clustering, the K-Means algorithms, and four initialization approaches.

2.1 Clustering Problem

The clustering problem can be formalized as follows [1]: Considering $N$ entities $e_i$, each with an associated weight $w_i$ ($i=1,..,N$), we search for $k$ centres $c_j$ ($j=1,..,k$) minimizing:

$$ f(c_1, ..., c_k) = \sum_{i=1}^{N} \min_j (w_i d(e_i, c_j)) $$

where $d(e_i, c_j)$ measures the dissimilarity between $e_i$ and $c_j$. In our case, where the entities are described by their co-ordinates in $\mathbb{R}^m$, $d(e_i, c_j)$ is the Euclidean distance. Basically, clustering is a combinatorial optimization problem.

Let

- $Q$ be set containing all objects to be clustered,
- $C$ be the set of all feasible clustering of $Q$,
- $J: C \rightarrow \mathbb{R}$ be the internal clustering criterion,

then the problem involves

$$ \text{Minimize } J(c) \text{ subject to } c \in C. $$

The complexity of the clustering problem is given by different factors:

1. The clustering is an NP-HARD problem. Therefore, an exhaustive approach is not practicable due to the exponential number of the potential partitions of the input data. The number of possible partitions of $N$ elements into $k$ clusters is given by

$$ \prod(k, N) = \frac{1}{k!} \sum_{j=1}^{k} (-1)^{k-j} \binom{k}{j} (j)^N. $$

2. The clustering complexity grows if the number of groups is unknown. In such a case the number of solutions becomes:

$$ \prod(k, N) = \sum_{i=1}^{k} \frac{1}{i!} \sum_{j=1}^{i} (-1)^{i-j} \binom{i}{j} (j)^N. $$

3. It is very difficult to translate the concept of 'similarity' into a unique mathematical model, but this depends on the clustering goal.