A new grouping genetic algorithm for clustering problems

ARIA KHOSH SIRAT

FATEMEH NIKMEHR

MEHRNOOSH MASHAYEKHI ZADEH

WINTER 2017

دانشگاه شیراز
Shiraz University

Introduction

- Clustering is an important subgroup of unsupervised learning techniques consisting in grouping data objects into disjoint groups of clusters.

- Uses include pattern recognition, bio-engineering, image quantization, renewable energy prediction, etc.

- Evolutionary computing algorithms (EAs) have been widely applied to clustering problems due to their capacity to be applied to very different problems with very few changes.

Clustering evaluation

- Validation or evaluation of the resulting clustering allows analyzing the result in terms of objective measures.
- Two groups of evaluation methods
  - Supervised measures :
    - Rand index (R)
    - Jaccardindex(J)
  - Unsupervised measures :
    - Davis-Bouldin Index (DB)
    - Silhouette coefficient(S)

Proposed grouping genetic algorithm

- GGA is a class of evolutionary algorithm especially modified to tackle grouping problems, i.e. problems in which a number of items must be assigned to a set of predefined groups. (by Falkenauer)
- Problem encoding:
  - Separating each individual in the algorithm into two parts : $c = [l|g]$

$$l1, l2, \ldots, lN | g1, g2, \ldots, gk$$

example :

1 3 2 1 4 1 1 2 3 2 1 3 4 2 1 | 1 2 3 4

Fitness Function

- Two different fitness evaluations
  - Davis-Bouldin Index (DB):

$$DB(U) = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} \left\{ \frac{\sum_{x \in C_i} d^2(\mathbf{x}, \boldsymbol{\mu}_i) + \sum_{x \in C_j} d^2(\mathbf{x}, \boldsymbol{\mu}_j)}{d^2(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)} \right\}$$

  - Silhouette coefficient (S):

$$s_j = \frac{a_j - b_j}{\max(a_j, b_j)},$$

$$S_j = \sum_{\mathbf{x}_j \in C_j} s_j,$$

$$S_j = \sum_{\mathbf{x}_j \in C_j} s_j,$$

Selection operator

- Rank-based wheel selection mechanism

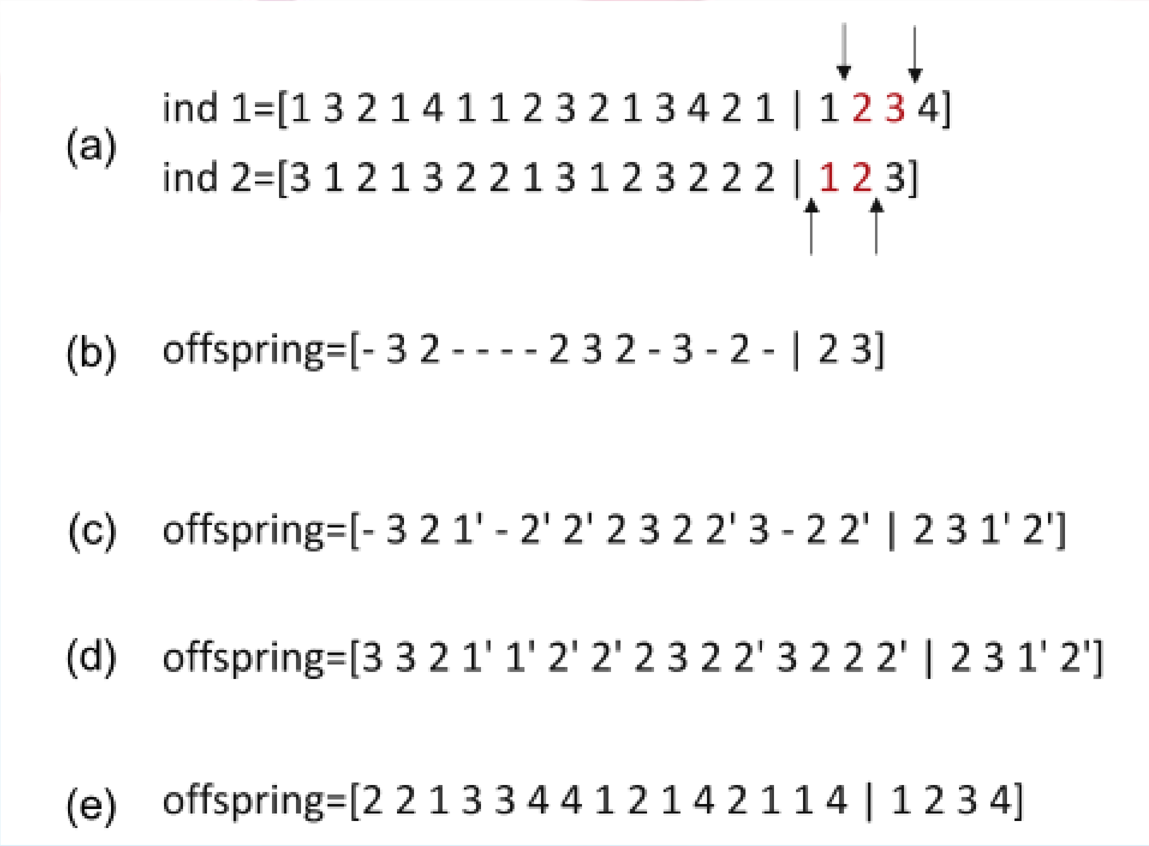
- $f = \frac{2\,.\,R}{\varepsilon\,.\,(\varepsilon+1)}$

- Static : Probabilities of survival (given by f) do not depend on the generation, but on the position of the individual in the list.

Crossover operator

- The probability of crossover must be high in the first stages of the algorithm, and moderate in the last ones in order to properly explore the search space.

$$P(j) = Pi - \frac{j}{TG}(Pi - Pf)$$



Mutation operator

- Mutation operator includes small modifications in each individual of the population with a low probability, in order to explore new regions of the search space and also scape from local optima.
- Two different mutation operators :
    - Mutation by cluster splitting :

      [2 2 1 3 3 4 4 1 2 1 4 2 1 1 4 | 1 2 3 4] $\rightarrow$ 2 2 1 3 3 4 4 5 2 1 4 2 5 1 4 | 1 2 3 4 5
    - Mutation by clusters merging :

      [2 2 1 3 3 4 4 1 2 1 4 2 1 1 4 | 1 2 3 4] $\rightarrow$ 2 2 1 3 3 2 2 1 2 1 2 2 1 1 2 | 1 2 3

$$P(j) = Pi + \frac{j}{TG}(Pf - Pi)$$

Replacement and elitism

- Elitist schema is also applied , the best individual in generation j isautomatically passed onto the population of generation $j + 1$.

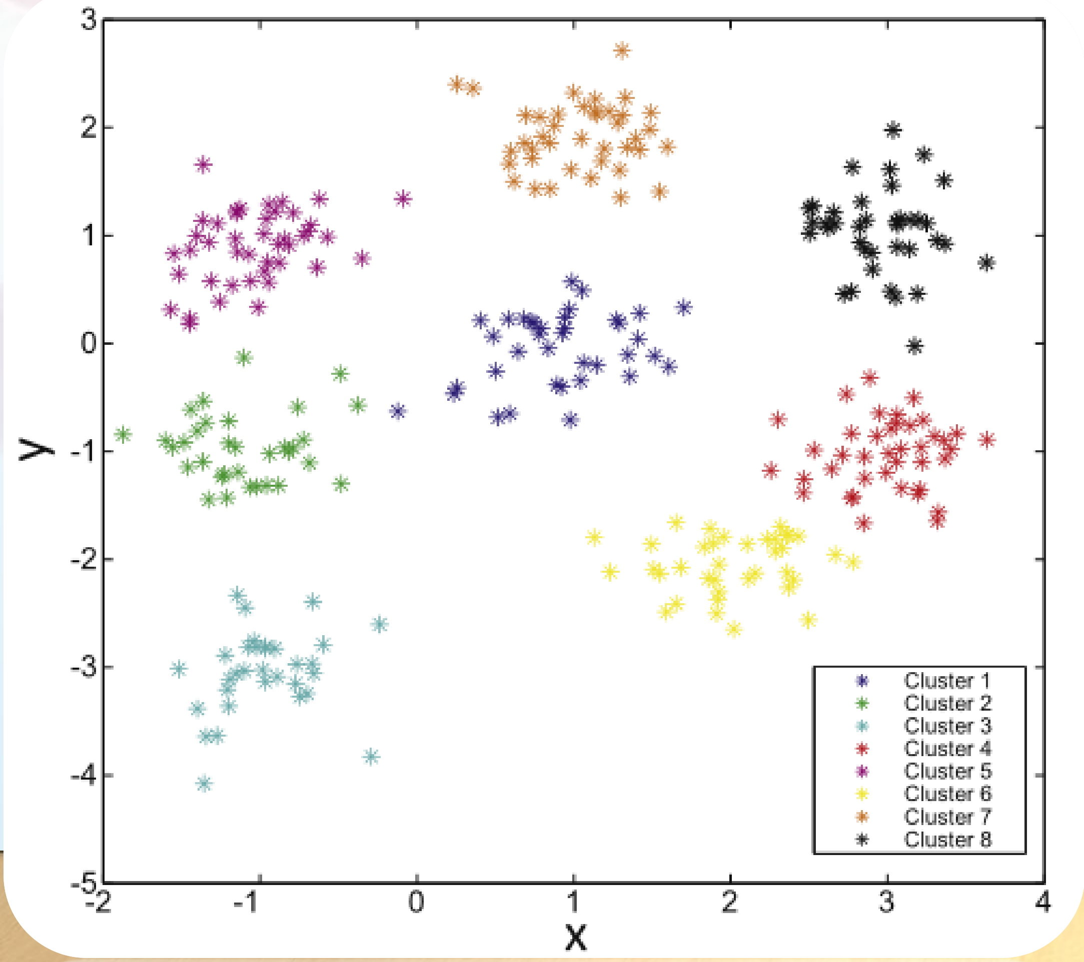
- Best solution encountered so far in the evolution is always kept by the algorithm.

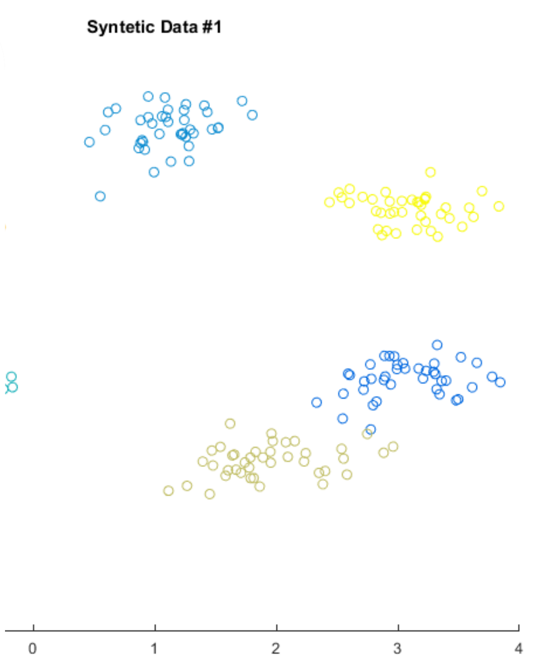
- Other individuals of population is formed by children.

Local search

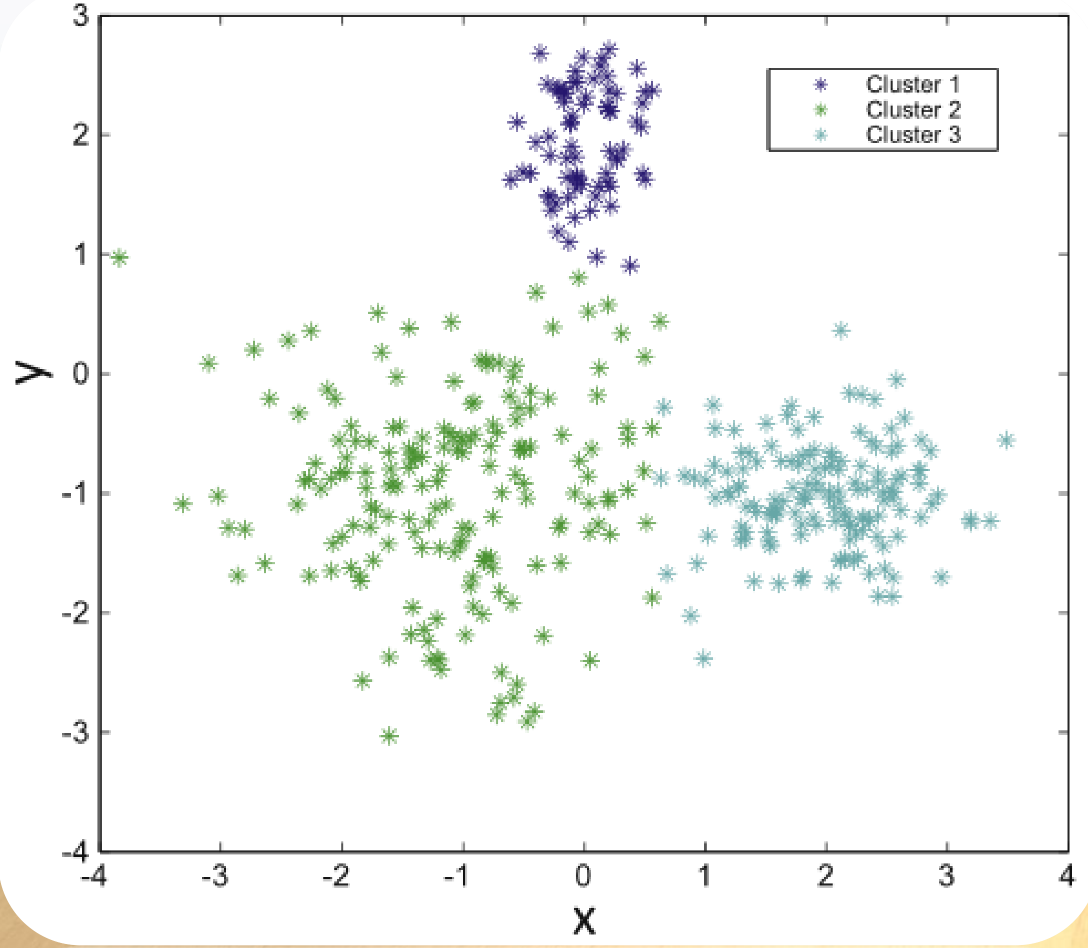
- Local search procedure tries to find local optimums in a close neighborhood of an individual.
- The implemented local search works over the element section of the individuals.
- For each observation, this operator determines the objective function variation obtained when the observation is assigned to the other clusters in the solution.
- keep the assignment with the largest objective function.
- Time consuming , so small probability.

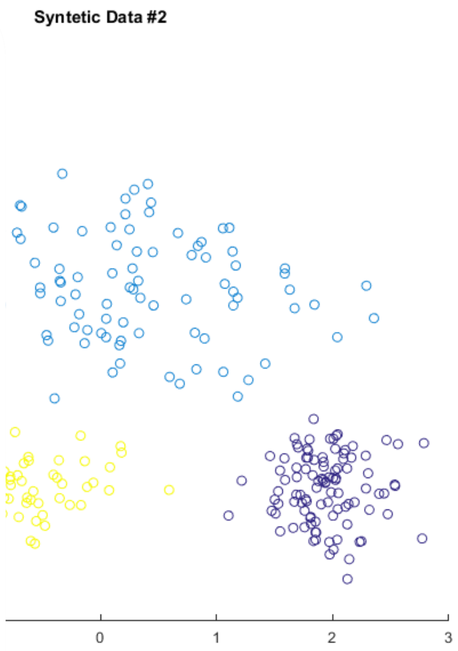
Experiments and results



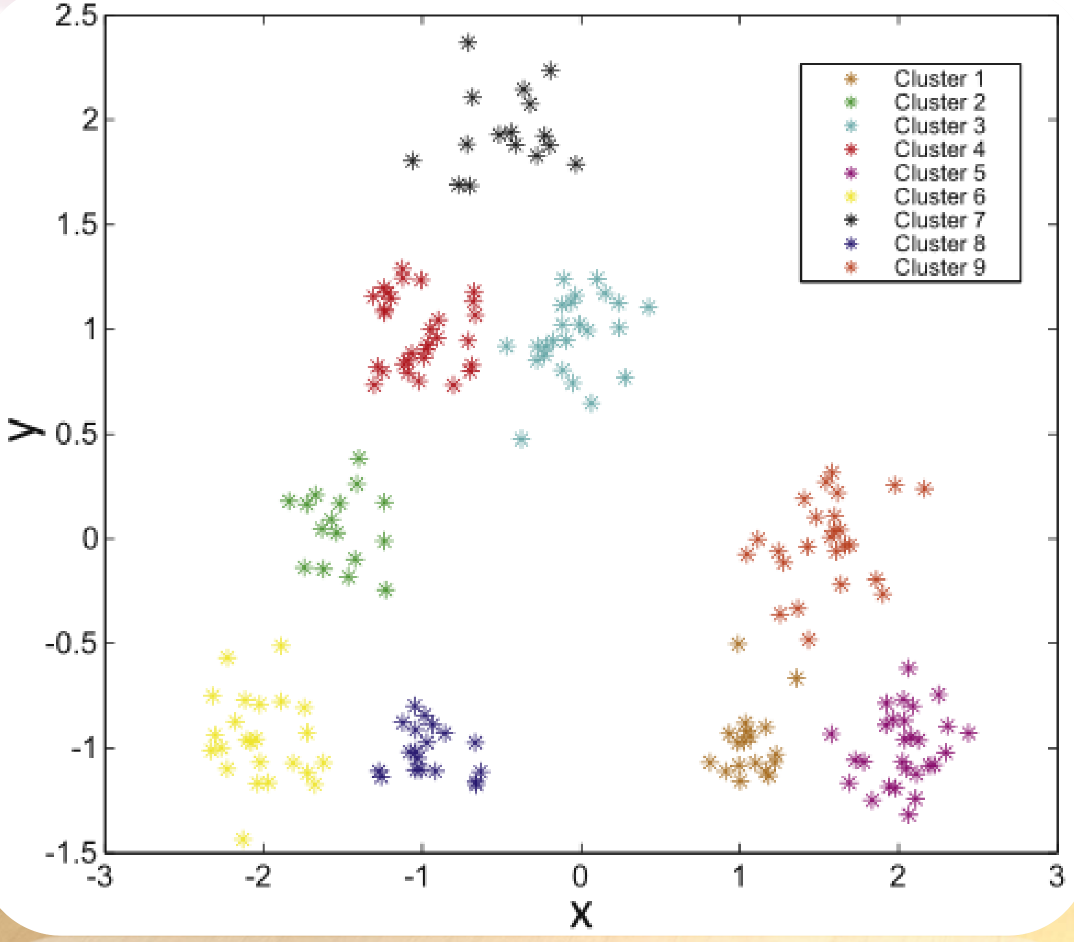


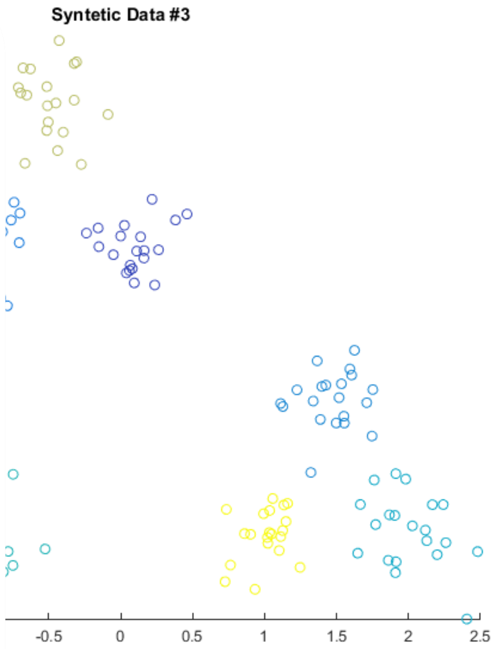
Experiments and results





Experiments and results





Experiments and results : Iris Dataset

Our Results

Fitness function	# Clusters	Best Fitness	RI
DB index	3	1.192004	0.87964
S index	3	-0.654043	0.874809

Paper Results

Fitness function	# Clusters	RI
DB index	3	0.8731
S index	3	0.8995

Experiments and results : Wine Dataset

Our Results

Fitness function	# Clusters	Best Fitness	RI
DB index	3	1.6248441	0.72424
S index	3	-0.656109	0.744239

Paper Results

Fitness function	# Clusters	RI
DB index	3	0. 7310
S index	3	0. 7220

References :

- Brown, E. C., & Sumichrast, R. T. (2005). Evaluating performance advantages of grouping genetic algorithms. Engineering Applications of Artificial Intelligence, 18(1), 1–12.
- Mclachlan, G. J., & Basford, K. E. (1988). Mixture models: inference and applications to clustering. New York: Marcel Dekker.
- L.E. Agustı´n-Blasa, S. Salcedo-Sanza,S. Jiménez-Fernándeza, L. Carro-Calvoa, J. Del Serb, J.A. Portilla-Figuerasa (2012). A new grouping genetic algorithm for clustering problems. Expert Systems with Applications, Elsevier