Programa de Pós Graduação em Computação
Universidade Federal de Pelotas

Um estudo de Algoritmos Genéticos Aplicado ao Problema de Clusterização

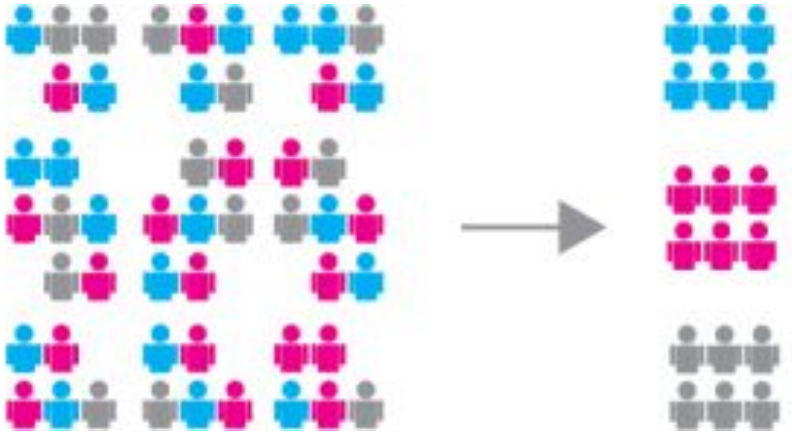
Karine Pestana Ramos
kpramos@inf.ufpel.edu.br

Introdução

- Desenvolver um GA para o problema de clusterização
- Busca por uma ótima solução
- Baseado no estudo de [3]
- Implementação de experimentos para avaliar o GA

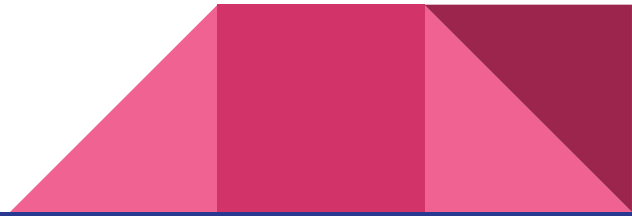
Clusterização

- Agrupamento por semelhança
- Uso de aprendizado não supervisionado



Algoritmos Genéticos

- Inspiração na evolução das espécies
- Buscas adaptativas em busca da melhor solução
- Podem ser classificados através dos seguintes componentes [1]:
  a. Problema a ser otimizado
  b. Representação do problema
  c. Decodificação do cromossomo
  d. Avaliação
  e. Seleção
  f. Operadores Genéticos
  g. Inicialização da População



Metodologia

- Uso de Python e DEAP [2]
- Disponibilidade do código online

Representação do problema

- Busca pelos centroides dos clusters
- Tamanho do indivíduo depende do número de clusters
- Representação em binário
- Uso do mínimo de bits necessários
- 1 cluster = um par ordenado (centroide)
- Exemplo:
  - “0111101 1001001 1110100 0000101”
  - pares ordenados de 0 a 127 (7 bits)

Outros componentes

- Decodificação do cromossomo
  - Conversão de binário para decimal
- Função de avaliação
  - Cálculo da distância Euclidiana
  - Distribuição do dataset em clusters
  - Novos centroides
  - Novas distâncias calculadas com os novos centroides

Comparação entre os estudos

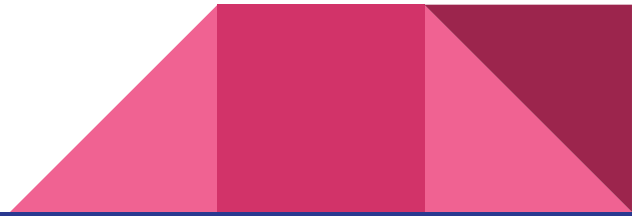
- Informações discutidas no estudo porém não de maneira detalhada
- Representação do problema, decodificação do cromossomo e função de avaliação apenas levemente inspiradas no estudo

TABLE I
COMPARAÇÃO DE PARAMÊTROS ENTRE OS ESTUDOS

Parâmetro	Estudo de [5]	Trabalho inspirado
Tamanho da população	6	8
Tipo de *crossover*	Em um único ponto	Em dois pontos
Tipo de mutação	Inversão de um bit	Inversão de um bit
Taxa de mutação	0.5	0.5
Critério de parada	Número de gerações	Número de gerações

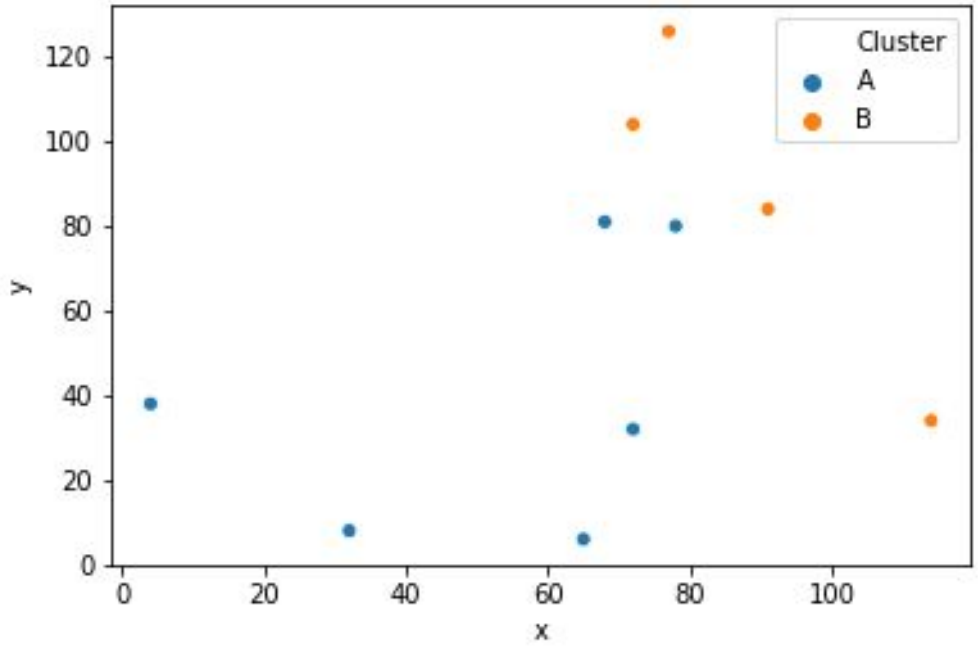
Experimento 1

- Dataset: 3 conjuntos de 10 pontos aleatórios pertencentes a R^2
- Clusterização de cada dataset para k=2
- Valores de 0 a 127
- Critério de parada: 40 gerações



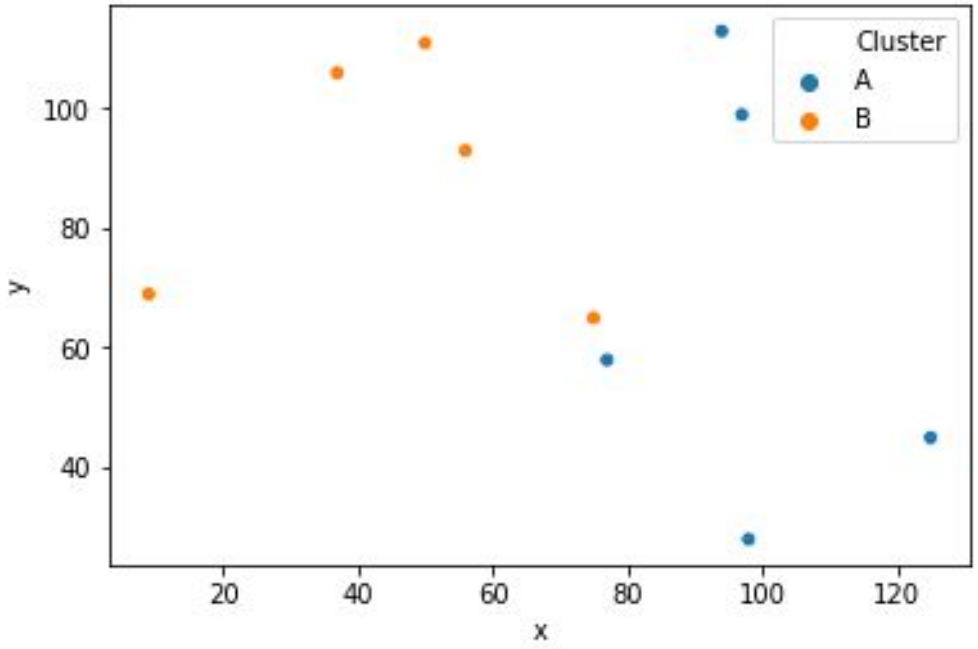
Experimento 1 - Resultados

Dataset1



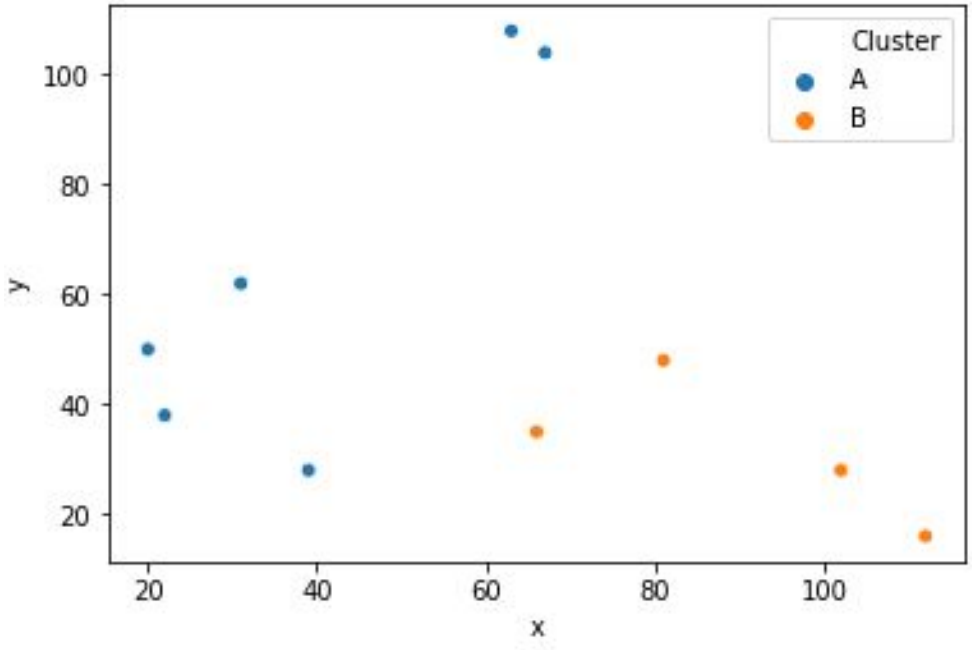
Experimento 1 - Resultados

Dataset2



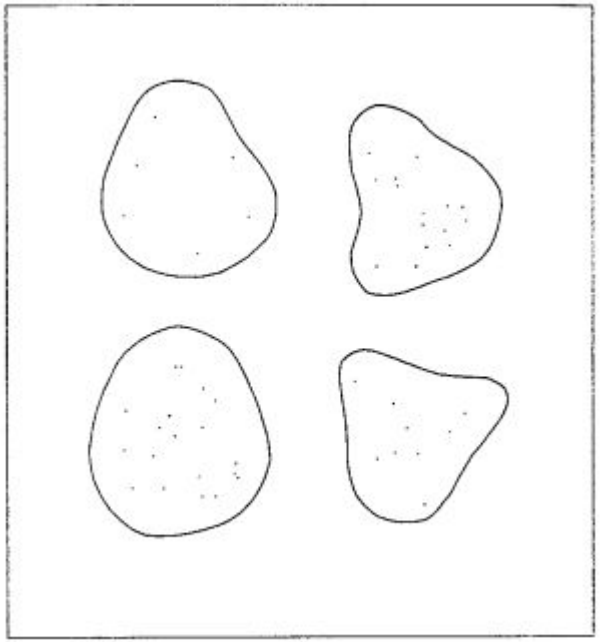
Experimento 1 - Resultados

Dataset3

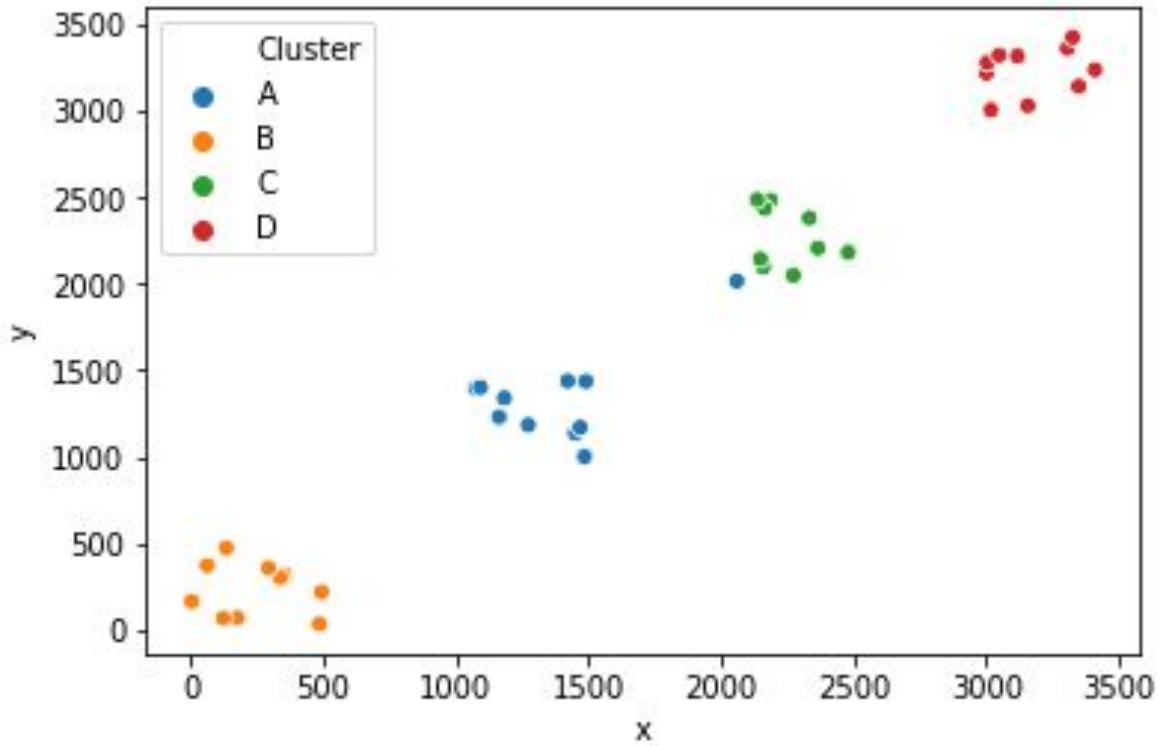


Experimento 2

- 1 único dataset
- 4 grupos de valores aleatórios pertencentes a R^2
- Pontos previamente agrupados
- Verificar se o GA é capaz de clusterizar da mesma maneira
- Critério de parada: 10000 gerações



Experimento 2 - Resultado



Considerações Finais

- Objetivo do estudo
- Compreensão limitada
- Tamanho da população x Espaço de busca
- Taxa de mutação variável e elitismo

Referências

1. P. Marco Aurélio Cavalcanti. "Algoritmos genéticos: princípios e aplicações." ICA: Laboratório de Inteligência Computacional Aplicada. Departamento de Engenharia Elétrica. Pontifícia Universidade Católica do Rio de Janeiro, 1999.
2. Felix-Antoine Fortin, Francois-Michel De Rainville, Marc-Andre Gardner, Marc Parizeau and Christian Gagne. "{DEAP}: Evolutionary Algorithms Made Easy". Journal of Machine Learning Research, 2012.
3. Murthy, Chivukula A., and Nirmalya Chowdhury. "In search of optimal clusters using genetic algorithms." Pattern Recognition Letters, 1996.

Programa de Pós Graduação em Computação
Universidade Federal de Pelotas

Um estudo de Algoritmos Genéticos Aplicado ao Problema de Clusterização

Dúvidas?

Karine Pestana Ramos
kpramos@inf.ufpel.edu.br