Um estudo de Algoritmos Genéticos Aplicado ao Problema de Clusterização

1st Karine Pestana Ramos
Centro de Desenvolvimento Tecnológico
Universidade Federal de Pelotas
Pelotas, Brasil
kpramos@inf.ufpel.edu.br

***Abstract*—O trabalho desenvolvido é baseado em um estudo realizado anteriormente sobre o problema de implementar algoritmos genéticos para a clusterização de conjuntos. Algoritmos genéticos permitem otimizações de problemas, dado a forma que são implementados, buscam encontrar a melhor solução para o problema em questão. Neste estudo, o GA desenvolvido é aplicado especificamente ao problema de clusterização. Foram desenvolvidos *datasets* e conjuntos de testes para a avaliação do GA. Através dos testes realizados observou-se que o GA apresenta resultados satisfatórios para a clusterização, entretanto, o mesmo ainda pode ser otimizado.**

***Index Terms*—algoritmos genáticos, ga, clusterização**

I. Introdução

Este trabalho aborda o estudo de Algoritmos Genéticos (GA) para a resolução de problemas de Clusterização, técnica que permite a descoberta de padrões presente em um conjunto de objetos.

O objetivo é a implementação de um GA em busca de ótimas soluções para o problema de clusterização, juntamente com a elaboração de experimentos que sejam capazes de testar a usabilidade do GA desenvolvido.

A metodologia abordada e os experimentos realizados foram inspirados no estudo de [5].

O estudo está organizado da seguinte maneira: a Seção II é designada para apresentação de cada uma das respectivas áreas de estudo abordadas, subvidindo-se em duas subseções, II-A e II-B; a Seção III aborda a metodologia desenvolvida e a as especificações necessárias para a implementação de um GA são discutidas nas subseções III-A, III-B, III-C e III-D; a Seção IV discute sobre os procedimentos adotados para a realização de testes sobre o GA; na Seção V é comparado brevemente os principais pontos em comum e divergentes entre o trabalho implementado e o estudo de [5]; a Seção VI discute os resultados encontrados pelo estudo; e, por fim, na Seção VII são apresentadas as conclusões deste trabalho.

II. Referencial Teórico

Essa seção apresenta uma breve introdução sobre os tópicos de estudos relacionados ao trabalho apresentado nesse artigo. Abordando sobre Clusterização e Algoritmos Genéticos.

A. Clusterização

A Clusterização consiste em um agrupamento de elementos por semelhança de uma base de dados, onde objetos mais similares fiquem no mesmo *cluster* e objetos distantes sejam designados para *clusters* diferentes.

A clusterização pode acontecer de forma que o número de *clusters* é definido previamente ou não. São conhecidos como Problema de Clusterização (PC) e Problema de Clusterização Automática (PCA), respectivamente [1].

A Figura 1 [7] é capaz de representadar de forma básica a ideia de clusterização. Onde dados elementos de natureza diferente são agrupados conforme sua semelhança.

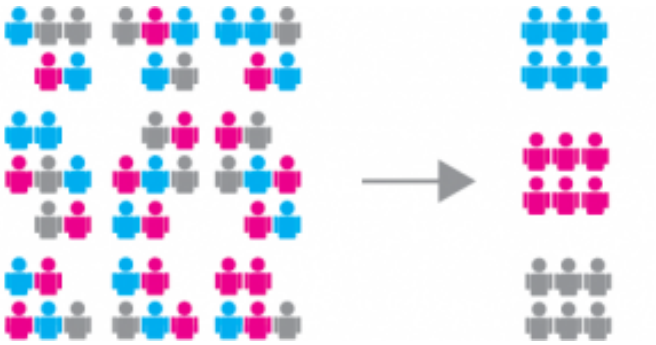

Fig. 1. Processo de clusterizar em grupos

Assim o objetivo da clusterização pode ser compreendido pelo agrupamento automatico através do uso de aprendizado não supervisionado, onde os n casos da base de dados são agrupados em k grupos, ou seja, em *clusters*. Na clusterização os padrões que estão agrupados em um mesmo *cluster* devem possuir mais semelhanças entre si do que os padrões que estão em outros *clusters*.

Há uma abundância de algoritmos de clusterização e suas variáveis atualmente, isto ocorre pois assim é possível escolher o melhor algoritmo para determinado conjunto de dados [2].

B. Algoritmos Genéticos

Algoritmos Genéticos (GAs, do inglês, *Genetic Algorithms*) tem sua origem inspirada na própria evolução das espécies, seguindo o princípio Darwiniano de reprodução e sobrevivência. Assim sua execução é análoga ao processo biológico de evolução natural. Além disso, GAs também são algoritmos probabilísticos que fazem uso de técnicas heurísticas para otimizações globais [3]. Suas buscas ocorrem de forma paralela e adaptativa sempre em busca da melhor solução.

Os algoritmos genéticos podem ser classificados através dos seguintes componentes [4]:

1) Problema a ser otimizado
2) Representação do problema
3) Decodificação do cromossomo
4) Avaliação
5) Seleção
6) Operadores Genéticos
7) Inicialização da População

Os componetes serão discutidos juntamente com o método proposto pelo estudo.

III. Método Proposto

Para a implementação de um algoritmo genético (GA) para o problema de clusterização foi usado como inspiração o estudo de [5]. Neste estudo é desenvolvido um algoritmo genético para o problema clusterização que é aplicado em experimentos onde seja possível comparar a eficiencia do GA em clusterizar.

No trabalho desenvolvido também é implementado um GA utilizando a linguagem de programação Python com auxílio da biblioteca DEAP [6]. Esta biblioteca é utilizada para facilitar o desenvolvimento de algoritmos genéticos possibilitando aos programadores o desenvolvimento mais rápido e simples de prototipagem e testes de ideias. Dessa forma, os algoritmos desenvolvidos são transparentes e bem estruturados.

Como já abordado anteriormente existem alguns critérios a serem definidos quando trata-se do desenvolvimento de um GA. O problema a ser otimizado é o problema de clusterização como explicado na seção II-A.

Todo o trabalho aqui proposto pode ser acessado via GitHub (https://github.com/karinepestana/SE-ClusterGA/blob/master/GA.ipynb) caso haja o interesse em analisar de forma mais completa a implementação elaborada.

A. Representação do problema

O problema de clusterização pode ser representado em um GA através da busca pelos pontos centrais de cada *cluster*, assim, dado um conjunto de dados e a quantidade de *clusters* em que se deseja que o conjunto seja dividido o GA busca encontrar esses pontos centrais de clusterização, chamados de centroides.

De acordo o número de *clusters* é definido o tamanho do indivíduo. Essa representação é feita em binário.

É importante ressaltar que o tamanho do indivíduo varia em cada execução dos experimentos, porém sempre de acordo com o número mínimo de bits em binário necessário para a representação de pares numéricos para cada *cluster*. Logo, para cada *cluster* há uma representação de um par ordenado. No caso de uma clusterização com 2 *clusters* com um dataset de pares ordenados variados de 0 a 127 é possível usar "0111101 1001001 1110100 0000101" como exemplo de *string* de representação. Neste caso, os espaços são utilizados para uma melhor vizualização de cada valor armazenado, podendo-se observar que para cada valor é necessária uma representação de 7 bits (a menor possível). A cada 14 bits é encontrado um par ordenado que representa o centroide do *cluster*. Para a representação geral deste exemplo em específico (a clusterização com k=2) é preciso minimamente uma representação de 28 bits para cada indivíduo da população.

B. Decodificação do cromossomo

Como descrito na Seção III-A os indivíduos são representados com uso de bits (com sua quantidade variando de acordo com o número de *clusters* de cada execução).

Para decodificação do indivíduo com a codificação "0111101 1001001 1110100 0000101" é feita a conversão de cada valor (ou seja, a cada 7 bits) para decimal. Tornando possível o cálculo da função de avaliação.

C. Função de avaliação

Para a função de avaliação é usado o cálculo da distância Euclidiana. Assim, para cada ponto do *dataset* analisado é calculada a distância até os centroides (indivíduo da população), com uso da distância Euclidiana. Ocorrendo a distribuição em n *clusters* destes pontos conforme pré determinado. Após a separação dos *clusters*, os centroides são recalculados e para a saída da função da avaliação é recalculado os valores das distâncias com uso dos novos centroides.

D. Outros parâmetros

Além dos pontos já abordados anteriormente mais alguns critérios são importantes na definição de um GA.

Para *crossover* e mutação são usados *crossover* de dois pontos. Na mutação é permitido que ocorra a inversão de um bit com uma probabilidade de 0.5. O processo de seleção de cromossomos ocorre através do método de torneio.

Para execuções é considerada uma população de tamanho 8. O GA tem como seu critério de parada um número fixo de iterações.

IV. Experimentos

Para a realização dos testes do GA desenvolvido foram ebalorados dois experimentos com base nos experimentos realizados em [5].

A seguir cada experimento será descrito conforme sua implementação.

A. Experimento 1

Para esse experimento são gerados 3 conjuntos de 10 pontos aleatórios pertencentes a R^2, resultando em 3 *datasets* de 10 pontos cada.

O experimento serve para verificar se o GA é capaz de clusterizar cada dataset para k=2, onde k é o número de *clusters*.

Os valores gerados estão na faixa de 0 a 127. Para a execução do GA em cada dataset foi considerado como critério de parada até 40 gerações.

B. Experimento 2

São criados 4 grupos de valores aleatórios pertencentes a R^2. Cada grupo é gerado de maneira que a distância de qualquer ponto a partir de seu valor médio é menor que a distância desse ponto em relação a qualquer ponto pertencente a um outro grupo. Esses grupos compõem o dataset que é usado no experimento. Ou seja, os pontos já estão clusterizados em 4 grupos. Para se ter uma ideia é possível olhar a Figura 2 retirada do trabalho de [5].

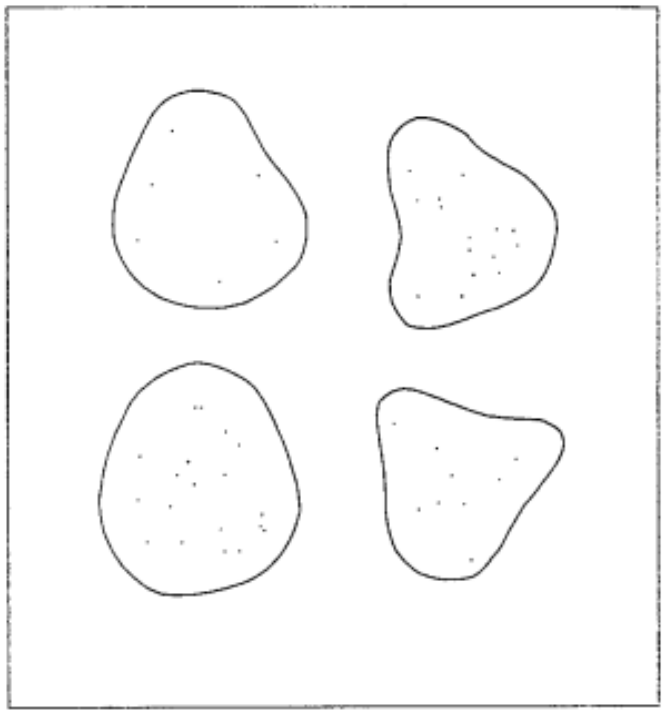

Fig. 2. Pontos e configuração de clusterização para o Experimento 2.

Dado que os pontos já estão agrupados, a intenção do experimento é verificar se o GA é capaz de encontrar os mesmos grupos de *clusters*.

V. Comparação da implementação feita com o estudo de [5]

Apesar do GA desenolvido ser baseado no GA implementado por [5] ainda há diferenças a serem ressaltadas e discutidas.

Não é específicado no estudo a linguagem em que o GA é implementado, ademais se há algum uso de bibliotecas de algoritmos genéticos. O objetivo principal de ambos os trabalhos é o mesmo: desenvolver um GA para clusterização. Contudo algumas específicações deixam a desejar na publicação do estudo. Muitas vezes as informações são discutidas no estudo, porém não de maneira clara como ocorre a implementação em si. Como alguns exemplos, pode-se citar a *string* de representação, como o GA implementado funciona e faz funcionar a clusterização e até mesmo a descrição de experimentos realizados. Como forma de transpassar e seguir com a implementação semelhante ao estudo a representação do problema, decodificação do cromossomo e a função de avaliação (Seções III-A, III-B e III-C, respectivamente) são apenas levemente inspiradas no artigo, pois havia falta de informações. Outros parâmetros como tamanho da população, tipo de *crossover*, tipo de mutação, probabilidade de mutação e critério de parada são mantidos ou apenas levemente modificados já que tratam-se de parâmetros numéricos que foram especificados com precisão no estudo. Para um comparação de forma geral é possível analisar a Tabela I

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

Sobre a Tabela I é importante ressaltar que a taxa de mutação é alta pois segundo o estudo de [5] é necessária uma alta variabilidade genética para alcançar a melhor solução. Porém, neste estudo a taxa de mutação varia conforme o GA converge para a melhor solução. Assim, a taxa de mutação inicia em 0.5 e decresce conforme a convergência. No trabalho aqui implementado, a taxa de mutação se mantem constante durante toda a execução.

Os experimentos realizados são fortementes baseados nos experimentos 1 e 2 do estudo. Os experimentos 3 e 4 tinham seu objetivo principal a comparação do GA com o algoritmo K-Means não foram abordados neste estudo.

VI. Resultados

Nesta seção são discutidos os resultados a respeito do GA implementado por esse estudo.

A. Experimento 1

O experimento 1 tinha como objetivo clusterizar 3 diferentes dataset em no máximo 40 gerações. Em média a população começa a demonstrar alguma convergência entre a 17° e a 18° geração.

Os resultados da clusterização realizada pelo GA podem ser vistos em forma de gráfico de pontos atráves das Figuras 3, 4 e 5.

Através das Figuras 3, 4 e 5 é possível observar que o GA pode não apresentar a melhor clusterização, porém realiza com sucesso o processo de clusterizar.

B. Experimento 2

O experimento 2 já tinha como base um dataset clusterizado, já existiam grupos de pontos visivelmente distantes uns dos outros. O objetivo desse teste era avaliar se o GA seria capaz de agrupar os pontos da mesma maneira, encontrando os centroides ideias. Neste caso, a população começou a demonstrar convergência entre a 15° geração e a 16°.

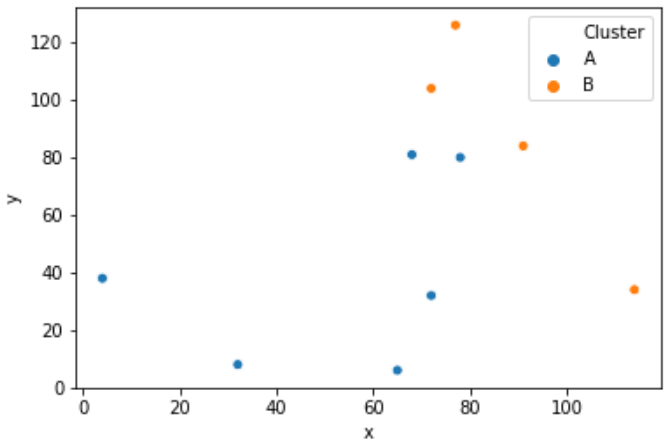

Fig. 3. Resultado da clusterização do dataset1 (Experimento 1)

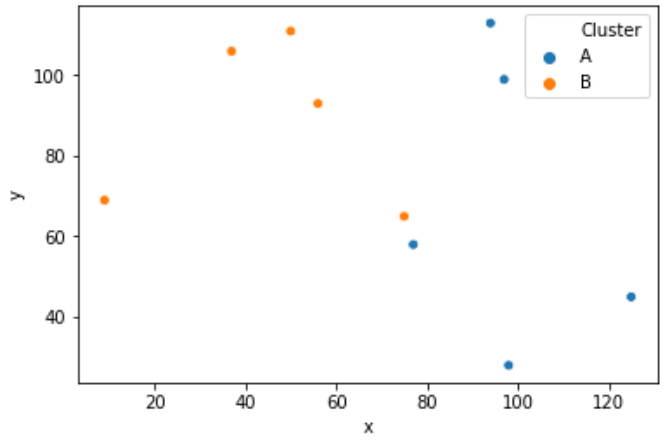

Fig. 4. Resultado da clusterização do dataset2 (Experimento 1)

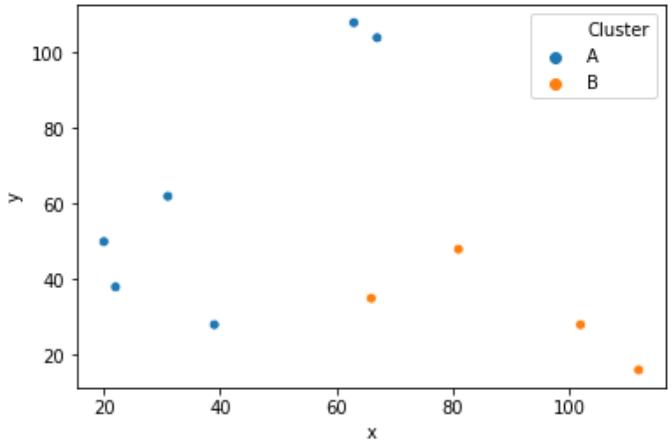

Fig. 5. Resultado da clusterização do dataset3 (Experimento 1)

Os resultados da clusterização realizada pelo GA podem ser vistos em forma de gráfico de pontos atráves da Figura 6.

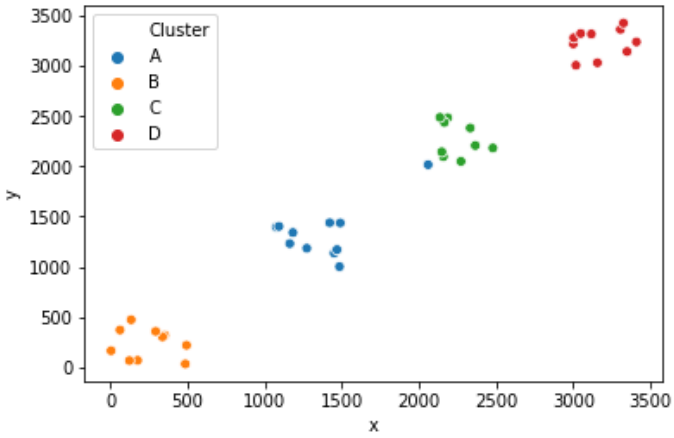

Fig. 6. Resultado da clusterização do dataset (Experimento 2)

VII. Conclusão

Este trabalho tinha como propósito a implementação (ao menos de forma similar) ao estudo de [5]. Mesmo com as diversidades na replicação do estudo, o algoritmo genético foi reimplementado fazendo uso de tecnologias e ferramentas atuais. O trabalho desenvolvido foi feito com a compreensão do estudo de [5]. É importante ressaltar que esta compreensão foi limitada as informações contidas no estudo e, por causa das diversas lacunas, levarem a busca por outras abordagens para a implementação do trabalho. Assim, o trabalho feito foi desenvolvido atingindo seu objetivo em pelo menos parcialmente replicar o estudo de [5] porém com buscas por informações adicionais sobre o processo de clusterização com algoritmos genéticos.

Assim como no estudo de [5] o tamanho da população é o mesmo para todos os experimentos, mas o espaço de busca varia de acordo com o tamanho do problema, ajustando-o conforme o tamanho do dataset a ser clusterizado.

Os experimentos realizados demonstram que o GA apesar de as vezes inicar alguma convergência até de maneira prematura é capaz de clusterizar. A alta taxa de mutação pode ter influenciado nesse processo.

Um trabalho futuro poderia ser a otimização desse estudo através da utilização de uma taxa de mutação ajustável, ela seria um adicional que poderia modificar positivamente as execuções além de auxiliar a convergência do GA, garantindo, dessa forma, que ao estar próximo do valor ótimo buscado não sofrerá mutações bruscas e estes valores não serão perdidos. Através do elitismo – incluso no estudo de [5] – é possível garantir que a sequência genética que possui valores altos na função de avaliação não seja perdida no decorrer nas populações.

References

[1] O., Luiz Satoru, C. Rodrigo Dias, and S. S. Furtado Soares. ”Clusterização em mineração de dados.” Instituto de Computação-Universidade Federal Fluminense-Niterói, 2004.

[2] Moscato, P., and Fernando J. Von Zuben. ”Uma visão geral de clusterização de dados.” DCA/FEEC/Unicamp, 2002.
[3] L. Ricardo. ”Algoritmos genéticos.” Brasport, 2008.
[4] 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.
[5] Murthy, Chivukula A., and Nirmalya Chowdhury. ”In search of optimal clusters using genetic algorithms.” Pattern Recognition Letters, 1996.
[6] 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.
[7] IDD - Inteligência de Dados. http://www.inteligenciadedados.com.br/clusterizacao/, 2019.