IFNMG - Instituto Federal do Norte de Minas Gerais
Curso: Ciência da Computação
Disciplina: Técnicas de Busca Heurística

Trabalho 003

Artur Pereira Neto
Breno Vambáster Cardoso Lima

Introdução

Este trabalho apresenta os resultados da investigação com busca local e tabu em um problema de minimização de distâncias de pontos e seus respectivos centroides.

Contextualização do problema

Este trabalho utiliza a base de dados Wine disponibilizada pelo UCI Machine Learning Repository. Essa base de dados é amplamente empregada em tarefas de aprendizado de máquina, incluindo classificação, agrupamento e análise de características químicas. Ela contém informações sobre os componentes químicos de vinhos oriundos de três variedades distintas cultivadas na região italiana de Lombardia. O objetivo principal dessa base é possibilitar a classificação dos vinhos em uma das três categorias, com base nas características químicas fornecidas. A base de dados foi originalmente publicada por Forina et al., no estudo intitulado "Chemical and Sensory Analysis of Italian Wines", que explorou como os componentes químicos do vinho poderiam ser utilizados para diferenciá-los de acordo com suas variedades. Essa base é uma referência na literatura devido à sua qualidade e relevância para estudos em classificação e agrupamento de dados multidimensionais. A base de dados contém 3 classes distintas, cada uma correspondendo a uma variedade de vinho. Cada registro é descrito por 13 atributos (contínuos e discretos) que descrevem as características físico-químicas do vinho. Nesse estudo a fim de reduzir a complexidade do problema a dimensionalidade da base de dados foi reduzida para um domínio em R^2. Os parâmetros Flavonoides e Fenóis totais foram as características escolhidas para a simplificação da base. Semelhante à primeira etapa da investigação

Objetivos

Aplicar uma abordagem evolutiva (algoritmos genéticos) a fim de otimizar a definição do conjunto de 3 centróides que minimizasse o somatório das distâncias euclidianas dos registros às suas respectivas coordenadas.

Conceitos principais

Algoritmos genéticos

Os algoritmos evolutivos constituem uma classe de métodos de otimização inspirados nos processos biológicos de evolução e seleção natural. Eles buscam soluções eficientes para problemas complexos por meio da simulação de mecanismos evolutivos, como reprodução, mutação e seleção. Dentro dessa categoria, os algoritmos genéticos (AGs) destacam-se como uma das abordagens mais conhecidas e amplamente aplicadas. Baseados na teoria da evolução de Darwin, esses algoritmos operam sobre uma população de soluções candidatas, refinando-as ao longo de múltiplas gerações até alcançar um resultado satisfatório.

O funcionamento de um algoritmo genético se inicia com a definição de uma população inicial de soluções representadas alegoricamente por “cromossomos”. Cada cromossomo codifica uma possível solução para o problema em questão, sendo formado por “genes” que representam variáveis específicas do problema. Durante o processo evolutivo, a qualidade de cada solução é avaliada por meio de uma função de fitness, que quantifica sua adequação ao objetivo do problema.

A cada geração, os indivíduos mais aptos são selecionados para reprodução, garantindo a propagação das características mais vantajosas. O cruzamento (crossover) permite a combinação de informações de diferentes indivíduos, promovendo a exploração do espaço de soluções. A mutação, por sua vez, introduz pequenas alterações nos cromossomos, contribuindo para a diversidade da população e evitando a convergência prematura para soluções subótimas. O elitismo é uma estratégia complementar que assegura a preservação dos melhores indivíduos de cada geração, garantindo que a qualidade das soluções não diminua ao longo do processo evolutivo.

Para um funcionamento eficaz, os algoritmos genéticos dependem da configuração adequada de diversos parâmetros fundamentais:

- **Tamanho da População**: Refere-se ao número de indivíduos em cada geração. Populações maiores proporcionam maior diversidade genética, porém aumentam o custo computacional.
- **Taxa de Cruzamento**: Define a probabilidade de dois indivíduos trocarem informações genéticas. Uma taxa elevada favorece a recombinação, acelerando a convergência para soluções promissoras.
- **Taxa de Mutação**: Controla a frequência com que mutações ocorrem. Um valor muito baixo pode levar à estagnação do algoritmo, enquanto um valor excessivo pode comprometer a convergência.
- **Critério de Parada**: Determina quando o algoritmo deve ser interrompido, seja após um número máximo de gerações, seja quando a melhoria das soluções se torna insignificante.

Os algoritmos genéticos são amplamente empregados em diferentes áreas da matemática, engenharia, aprendizado de máquina, bioinformática, logística e inteligência artificial. Sua capacidade de explorar grandes espaços de busca e encontrar soluções eficientes os torna uma ferramenta valiosa para investigação de problemas onde métodos tradicionais de otimização são inviáveis ou insuficientes.

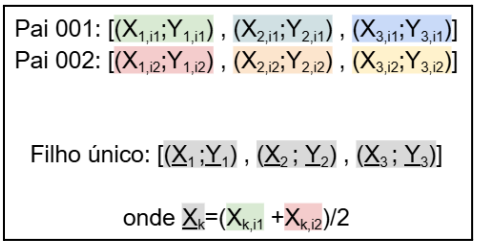
Proposta de abordagem

- **Definição da população**
  - Foi usada a solução ótima da etapa anterior junto com mais 19 centróides aleatórios dentro do espaço de possibilidades, definindo a população original, **P**.

- **Função de fitness**
  - Avaliação da distância euclidiana (métrica de avaliação) de cada um dos indivíduos da população P aos centróides de cada proposta de solução.

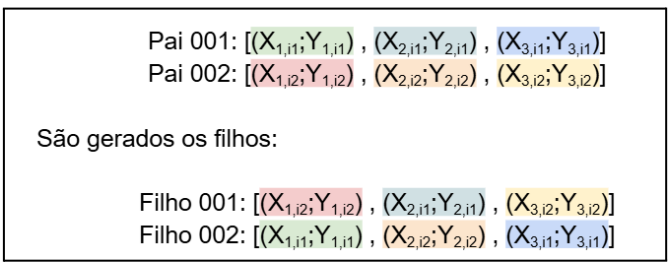
- **Cruzamento**
  - Os indivíduos da população **P** são agrupados em pares.
  - O Pareamento é baseado na métrica de avaliação, a fim de garantir que os “genes” de uma boa solução possam ser repassados para a próxima geração.
  - Os filhos compõem a população **F**.
  - Foram avaliadas duas propostas, P001 e P002, para a geração dos filhos, apresentadas nas figuras 1 e 2.

Figura 1: Proposta P001 de geração de filhos: baseada na média das coordenadas dos pais.



Fonte: autores

Figura 2: Proposta P002 de geração de filhos: baseada na troca de coordenadas de 2 dos 3 centróides.



Fonte: autores

Na proposta 001 são gerados 10 novos indivíduos na na população F enquanto que na proposta 002 são gerados 20 novos indivíduos. Os indivíduos da população **P** e **F** juntos formam a população **T**, que agrupa todos os indivíduos daquela geração.

- **Mutação**
    - Para cada indivíduo da população T é avaliada a possibilidade de ocorrência de mutação.
    - Define-se uma taxa de ocorrência de mutação $t_{mutação}$=0,03.
    - Para cada coordenada, de cada um dos centróides, de cada um dos indivíduos da população T é gerado um número um número aleatório $p_{mutação}$, entre 0 e 1, que indica a probabilidade daquela coordenada sofrer mutação. Caso $p_{mutação} \leq t_{mutação}$ a coordenada avaliada é alterada, por meio da adição/subtração de um valor d onde $d \in [-d_{max} , d_{max}]$

        - Se as três coordenadas do indivíduo $[(X_{1,i2};Y_{1,i2}) , (X_{2,i1};Y_{2,i1}) , (X_{3,i2};Y_{3,i2})]$ forem escolhidas para mutação, as novas coordenada serão do indivíduo serão $[(X_{1,i1} \pm d_{I} ; \ Y_{1,i1} \pm d_{II}) , (X_{3,i1} \pm d_{III} ; \ Y_{3,i1} \pm d_{IV}) , (X_{3,i1} \pm d_{V} ; Y_{3,i1} \pm d_{VI})]$.

- Seleção
    - É usada a abordagem de elitismo: escolher os 20 melhores indivíduos da população T, com base na métrica de avaliação, para formarem a nova geração de indivíduos da população P.

Resultados e Discussão

São apresentados a seguir os resultados para as duas abordagens de geração de filhos, P001 e P002, para 200 e 500 gerações respectivamente. Os gráficos das figuras 3 e 4 apresentam os resultados para a abordagem P001 enquanto os gráficos 5 e 6 apresentam os resultados para a abordagem P002. O quadro 1 reúne os resultados de todas as abordagens avaliadas ao longo do semestre para o problema.

Figura 3: Evolução da função fitness para a abordagem P001 (um filho por pais).

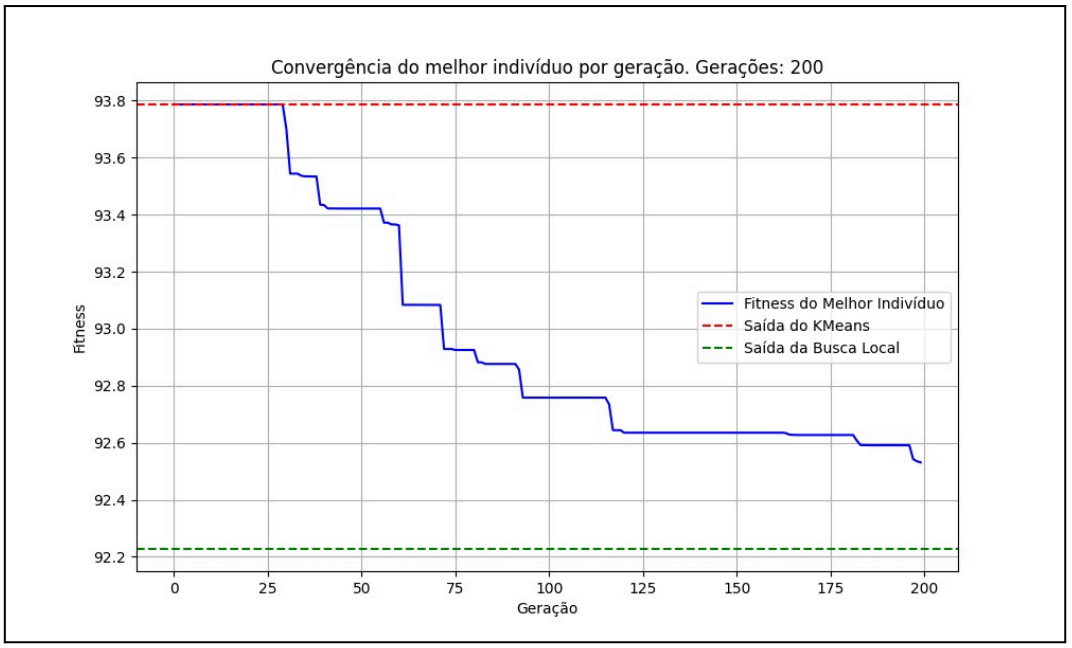
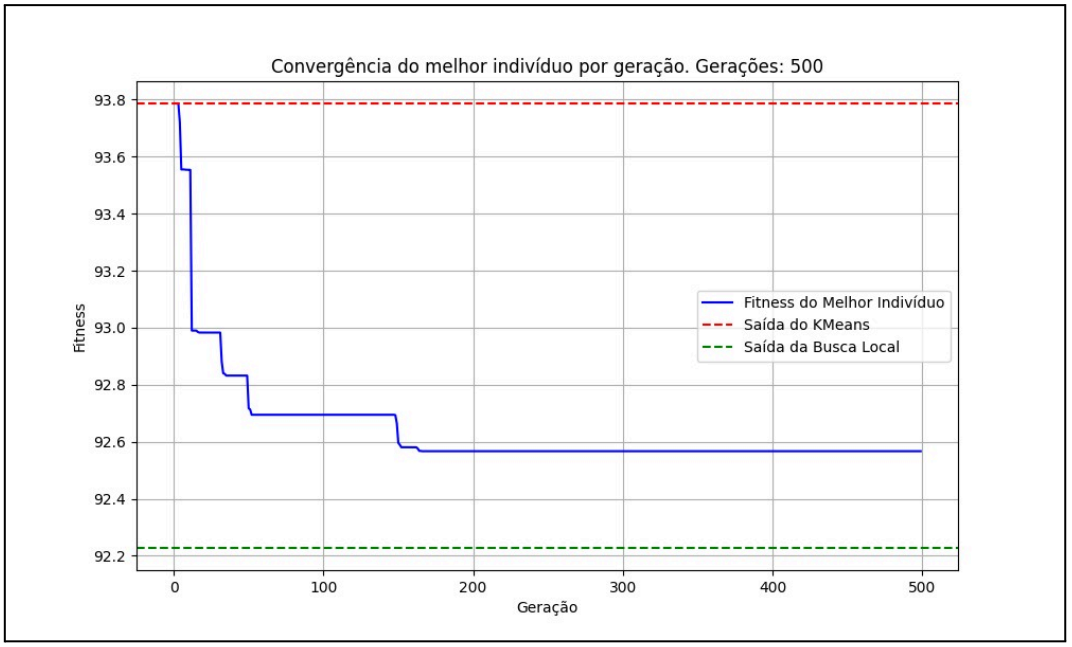
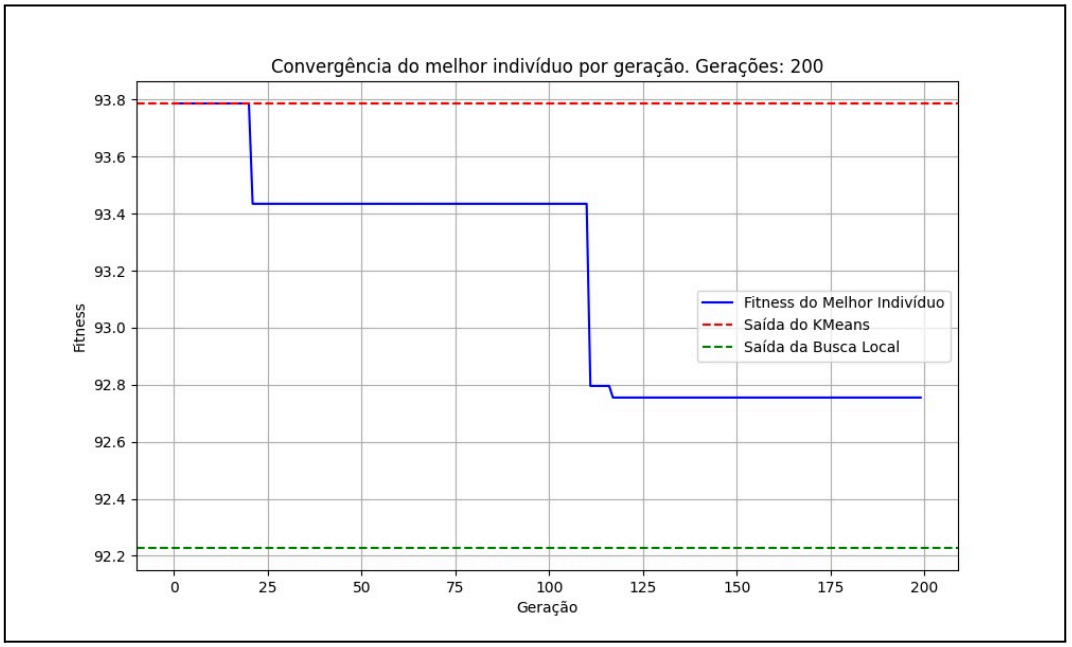


Figura 4:Evolução da função fitness para a abordagem P001 (um filho por pais) com 500 gerações.



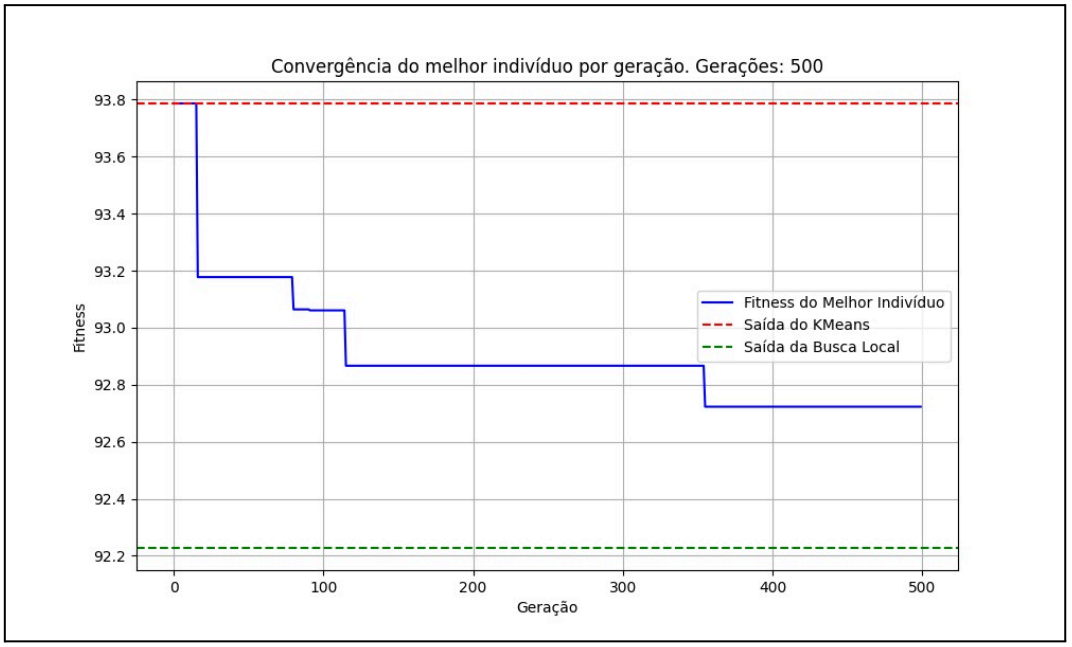
Fonte: autores

Figura 5: Evolução da função fitness para a abordagem P002 (dois filhos por pais) com 200 gerações.



Fonte: autores

Figura 6: Evolução da função fitness para a abordagem P002 (dois filhos por pais) com 500 gerações.



Fonte: autores

Quadro 1: Comparativo dos resultados das diferentes abordagens.

Abordagem construtiva (referência)		
Método	**Custo**	
k means (Referência)	93.7869	
Abordagem com vizinhança local		
Método	**Custo VPS**	**Custo VPD**
Primeira Melhora	93.308 *[Decréscimo de 0.0486%]*	93.2405 *[Decréscimo de 0.1209%]*
Melhor Melhora	92.3988 *[Decréscimo de 1.0226%]*	92.2274 *[Decréscimo de 1.2062%]*
Tabu	92.274 *[Decréscimo de 1.2062%]*	92.2274 *[Decréscimo de 1.2062%]*
Abordagem evolutiva [Algoritmo genético]		
Método	**Custo**	
Algoritmo genético com proposta P001 de geração de filhos	92.5308 (200 gerações) 92.5435 (500 gerações)	
Algoritmo genético com proposta P002 de geração de filhos	92.7550 (200 gerações) 92.6674 (500 gerações)	

Fonte: autores

Em ambas as abordagens o algoritmo genético foi capaz de otimizar a solução de referência gerada pelo KNN. Em comparação com as abordagens de busca local, a abordagem evolutiva não foi capaz de alcançar a performance das primeiras. Dentre as abordagens de gerações de filho investigadas a abordagem P001 que gerava filhos pela média das coordenadas dos pais mostrou uma melhor performance que a P002, que realizava a troca de coordenadas dos centroides originais. Tal resultado vai ao encontro das expectativas visto que a geração de filhos através da média das coordenadas dos pais assegura que a informação da posição de um centroide melhor ajustado, não seja perdida entre as diferentes gerações. Por outro lado, a simples troca de coordenadas pode, potencialmente, inserir desnecessariamente aleatoriedade na posição dos centroides investigados.

Considerações finais

Este trabalho reúne os resultados de três abordagens de busca aplicadas a uma base de dados de dimensionalidade reduzida gerada a partir da base "Wine Quality". As técnicas de busca aplicadas ao problema de agrupamento de vinhos da base *Wine Quality* demonstraram diferenças significativas em desempenho, destacando-se a busca tabu como a abordagem mais eficaz. A busca local com duas vizinhanças trouxe melhorias pontuais, mas sua limitação a ótimos locais restringiu o desempenho. Já a busca tabu, com sua capacidade de escapar de mínimos locais e explorar melhor o espaço de soluções, obteve os melhores resultados.O algoritmo genético não conseguiu superar a busca tabu devido a dificuldades na convergência para soluções de alta qualidade. A necessidade de ajuste dos parâmetros de taxa de mutação e cruzamento pode ter influenciado a eficácia da busca, levando a uma convergência prematura ou a uma exploração ineficiente do espaço de soluções. Além disso, a natureza do problema de agrupamento pode não ter favorecido a recombinação genética como mecanismo de otimização, dificultando a evolução para soluções superiores. Assim, embora os algoritmos genéticos sejam eficientes em diversos contextos, neste estudo, sua performance foi inferior à da busca tabu.