Inteligência Artificial

Relatório - Trabalho Prático

Executando algoritmos de agrupamento

Prof.: Katti Faceli

Aluno: Breno Vinicius Viana de Oliveira	**RA:** 726498
Aluno: Bruno Morii Borges	**RA:** 726500
Aluno: Isabella Soares de Lima	**RA:** 726541
Aluno: Mariane Lamas Malheiros	**RA:** 726569

25 de Janeiro de 2019

1 Introdução

Neste documento está registrado um resumo sobre o trabalho prático e análise dos resultados obtidos pelos integrantes do grupo (citados acima) na implementação e execução dos algoritmos K-médias, Single-link e Average-link nos conjuntos de dados passados pela professora. Além disso, seguem descritos os resultados com a avaliação dos melhores obtidos através do cálculo do índice Rand ajustado ao comparar com os resultados reais esperados, também fornecidos pela professora.

Os conjuntos de dados recebidos contêm as seguintes características:

1. Nome: c2ds1-2sp; Número de Dados: 1000
2. Nome: c2ds3-2g; Número de Dados: 1000
3. Nome: monkey; Número de Dados: 4000

2 Implementação e Execução

Os algoritmos foram implementados de forma que:

- Seguissem as regras de implementação como visto em aula
- Sejam capaz de ler e escrever arquivos
- A implementação está em um loop com o número de clusters requisitado de cada conjunto de dados
  - 5 a 12 clusters para monkey
  - 2 a 5 para os outros
- K-médias possui número fixo de 1000 iterações

Houve tentativas de melhorar a eficiência dos algoritmos a fim de produzir resultados mais rapidamente sendo o K-medias o mais fácil de realizar as adaptações. No entanto, apesar dos esforços, os algoritmos single e average link continuaram com certa demora. E desta forma, o grupo encontrou um problema com o conjunto monkey na execução do single link e average link.

O problema consiste no tempo de execução que o algoritmo leva para executar. Com os conjuntos de 1000 dados, cada execução levou de 10 a 30 minutos, mas ainda foi possível obter os resultados. Porém, com o conjunto monkey, o algoritmo foi executado nos notebooks dos integrantes e levou horas sem sucesso. Também foi usado um computador de mesa com média performance e ainda assim, após 8 horas de execução, não havia alcançado metade do Conjunto. Ainda assim conseguimos ao menos o resultado para 5 clusters com single link.

Assim, para o conjunto monkey, os resultados estão incompletos quando se trata do algoritmo Single link, ou seja, este não será considerado na análise dos resultados.

Para cada Conjunto de Dados, há uma pasta de resultados com o mesmo nome e cada pasta contém uma subpasta "kx"em que x é o número de clusters usado no algoritmo para gerar aqueles resultados. Em cada subpasta há 3 arquivos, contendo resultados de cada algoritmo respectivamente (com exceção do conjunto monkey que não possui os resultados do Single Link).

3 Resultados

De forma direta, os melhores resultados comparando com os resultados reais foram:

- Nome: c2ds1-2sp; Rand = 1 (2 clusters), usando o algoritmo Single-Link
- Nome: c2ds3-2g; Rand = 0,91 (2 clusters), usando o algoritmo K-médias
- Nome: monkey; Rand = 0,62 (12 clusters), K-médias obrigatoriamente

Abaixo seguem algumas comparações e conclusões pelos resultados esperados.

3.1 Conjunto c2ds1-2sp

De acordo com a análise do gráfico esperado, percebe-se que a forma é de duas espirais bem separadas entre si. Essas espirais formam dois clusters em que os dados são próximos um do outro formando um encadeamento, característica interessante para o melhor resultados obtido. Abaixo, tem-se o gráfico do melhor resultado :

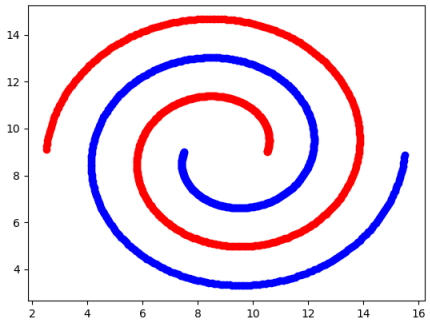

Figura 1: (a) Gráfico obtido [Single, 2 clusters]

Como esperado, o melhor resultado vem do algoritmo single-link pois ele agrupa os dados que estão ligados por encadeamentos (são vizinhos próximos). Além disso, o single-link tem problemas com dados em que há pouca separação espacial o que não ocorre nesse exemplo, já que as duas espirais estão bem separadas.

3.2 Conjunto c2ds3-2g

De acordo com a análise do gráfico esperado, percebe-se que a forma é de dois conjuntos circulares levemente se encostando. Essa forma de conjunto de dados entra na categoria de compactação em que há variações pequenas dentro de um cluster e os clusters normalmente são bem separados. Abaixo, tem-se o gráfico do melhor resultado:

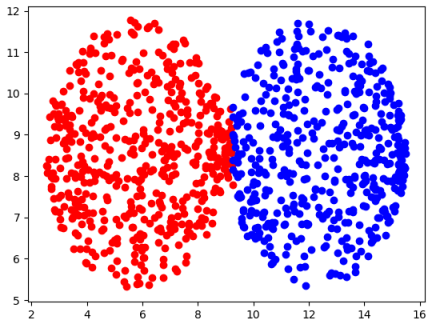

Figura 2: (a) Gráfico obtido [K-médias, 2 clusters]

Ainda que as duas esferas do resultado estejam próximas, o melhor resultado foi obtido com o k-medias que aborda dados compactos para agrupá-los.

3.3 Conjunto monkey

De acordo com a análise do gráfico esperado, percebe-se que a forma é de um rosto de macaco sendo cada cluster uma componente (orelhas esquerda, orelha direita, olhos, boca, contorno do rosto, sobrancelha esquerda, sobrancelha direita). Neste caso há agrupamento de dados encadeados e compactos o que pode gerar dúvida no melhor algoritmo. Abaixo, tem-se o gráfico do melhor resultado:

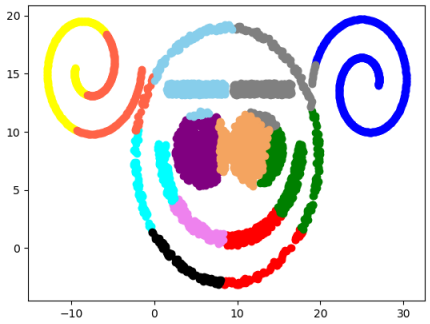

Figura 3: (a) Gráfico obtido [K-médias, 12 clusters]

Infelizmente, o melhor resultado é o maior RAND do K-médias que ainda é aceitável, porém, pela visualização dos dados reais, o single-link teria melhores resultados.

4 Conclusões

O trabalho prático foi de extrema utilidade para relembrar os conceitos de agrupamento estudados anteriormente na disciplina e complementar no aprendizado de forma que os integrantes tiveram o dever de implementar de fato os algoritmos e testar com os conjuntos fornecidos. Analisar os resultados e ponderar sobre as diferenças entre eles e os algoritmos também é de grande importância.

Além disso, ajudou a entender bem as diferenças e como os parâmetros dos algoritmos (número de clusters e iterações) afetam nos resultados. Também pode-se perceber que usar um números fixo, heurística ou aleatoriedade para definição de centroides no K-médias faz grande diferença.

Por fim, é reforçado como os algoritmos apesar de testados largamente e bem conhecidos, são lentos devido a quantidades de testes e dados que recebe. Graças a isso, o conjunto monkey ficou incompleto por ter problemas na demora de escutar os algoritmos single link e average link.