Trabalho prático de Mineração de Dados

André Henrique Santos Silva, Frederico Augusto, Guilherme Cota Soares, Gustavo Alves, Leonardo Flavio Santos, Rafael Alves Conrado

1. Introdução

O que diferencia os diversos algoritmos é o método ou estratégia utilizada para identificar os pares de clusters mais semelhantes. Para isso, diversas estratégias foram propostas, tais como: Single Link, Complete Link, Average Link, Centroid-based e etc, que serão descritas ao longo do texto.

2. Conceitos

2.1 Agrupamento (Cluster)

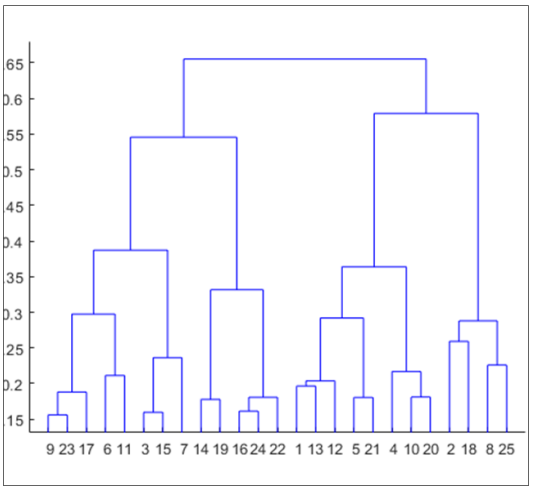
A técnica de mineração por agrupamentos busca ligar grupos de objetos ou elementos mais parecidos entre si. O agrupamento é feito com base nas semelhanças entre os elementos, cabendo ao analisador das classes resultantes avaliar se estas significam algo útil. Por exemplo, agrupar sintomas pode gerar classes que não representem nenhuma doença explicitamente, uma vez que doenças diferentes podem possuir os mesmos sintomas.

2.2 Agrupamento Hierárquico

Neste método de agrupamento, a quantidade de grupos a serem formados não é um parâmetro, isto acontece porque o resultado final é só um grupo contendo todos os objetos. A idéia é colocar as uniões numa hierarquia decrescente de similaridade. Primeiro são agrupados os dois objetos que possuem maior similaridade, depois este grupo de dois objetos é unido com o terceiro objeto mais similar e assim sucessivamente até formar um único grupo Essa representação facilita a visualização sobre a formação dos agrupamentos em cada estágio onde ela ocorreu e com que grau de semelhança entre eles.

2.3 Dendograma

Dendrograma é a representação gráfica, em forma de árvore, da estrutura dos agrupamentos. Nos métodos aglomerativos, o dendrograma representa a ordem em que os dados foram agrupados



3. Metodologia de Aglomeração

O método para se obter os agrupamentos pode ser feito de acordo com os passos abaixo:

1- Cada agrupamento contém um único padrão ou um único objeto.
2- Calcular a matriz de similaridades entre os grupos (pode ser aplicado o método single link, complete link entre outro)
3- Forma-se um novo agrupamento pela união dos agrupamentos com maior grau de similaridade.
4- Os passos 2 e 3 são executados (N-1) vezes, até que todos os objetos estejam em um único agrupamento.

A matriz de similaridade contém a distância entre os agrupamentos em cada estágio do algoritmo. Dessa forma, imaginando um estágio do algoritmo onde o número de agrupamentos corrente é três (G1, G2, G3), pode-se supor a seguinte matriz de similaridades entre os agrupamentos:

	G1	G2	G3
G1	X	0,1	0,3
G2	0,1	X	0,4
G3	0,3	0,4	X

Pode observar que G1 e G2 são os agrupamentos mais similares, enquanto que G2 e G3 são os menos similares.

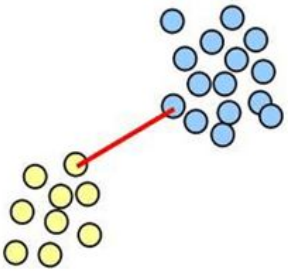
4. Cálculos de Distância

4.1 Single Link

A distância entre dois agrupamentos é dada pela distância entre os seus padrões mais similares (próximos).

Onde i e j são respectivamente os padrões dos agrupamentos C1 e C2 e d(i, j) é a distância entre os objetos i e j.

$$D(C_1, C_2) = \min_{\substack{i \in C_1 \\ j \in C_2}} (d(i, j))$$

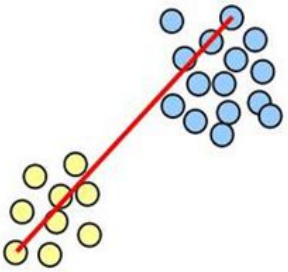


single-link

4.2 Complete Link

A distância entre dois agrupamentos é dado pela distância entre os seus padrões menos similares (próximos). Onde i e j são respectivamente os padrões dos agrupamentos C1 e C2 e d(i, j) é a distância entre os objetos i e j.

$$D(C_1, C_2) = \max_{\substack{i \in C_1 \\ j \in C_2}} (d(i, j))$$



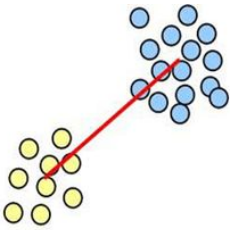
complete-link

4.3 Average Link

A distância entre dois agrupamentos é dada pela média das distâncias entre cada padrão de um agrupamento em relação aos padrões do outro agrupamento.

Onde N1 e N2 são respectivamente os números de objetos dos agrupamentos C1 e C2 e i e j são respectivamente os padrões das classes C1 e C2.

$$D(C_1, C_2) = \frac{N_1 \underset{i \in C_1}{d}(i, C_1) + N_2 \underset{j \in C_2}{d}(j, C_2)}{N_1 + N_2}$$



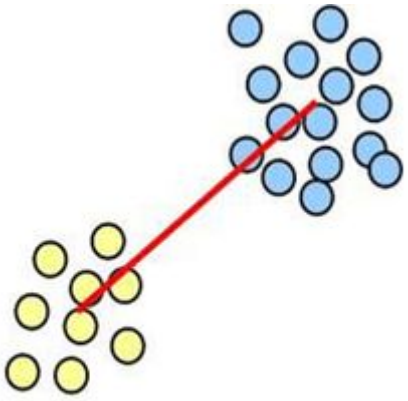
average-link

4.4 Centroid-based

A distância entre dois agrupamentos é dado pela distância entre os centróides. O centróide é a média dos padrões do agrupamento.

Onde μ1 e μ2 são respectivamente os centróides dos agrupamentos C1 e C2 e d(μ1, μ2) é a distância entre eles.

CD(C ,C) = d μ ,μ



average-link

5. Algoritmos

5.1 Algoritmo Single-link

- Defina cada padrão como sendo um cluster;
- Construa uma lista das distâncias entre padrões, para todos os pares de padrões;
- Ordene essa lista em ordem crescente;
- Percorra essa lista de distâncias do seguinte modo: para cada valor dk da lista de distâncias, construa um grafo onde, os pares de padrões cujos valores são mais próximos de dk são conectados por uma aresta;
- Se todos os padrões são membros de um grafo conexo, então, pare. Caso contrário repita todos esses passos

5.2 Algoritmo Complete-link

- Defina cada padrão como sendo um cluster;
- Construa uma lista das distâncias entre padrões, para todos os pares de padrões;
- Ordene essa lista em ordem crescente;
- Percorra essa lista de distâncias do seguinte modo: para cada valor dk da lista de distâncias, construa um grafo onde, os pares de padrões cujos valores são mais próximos de dk são conectados por uma aresta;
- Se todos os padrões são membros de um grafo conexo, então, pare. Caso contrário repita todos esses passos.

6. Referências

VALE, Marcos Neves do . Agrupamentos de Dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos. 2006. Dissertação de Mestrado em Engenharia Elétrica. Pontifícia Universidade Católica- Rio de Janeiro -RJ. Disponivel em: <https://www.maxwell.vrac.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=7975@1>

METZ, Jean; **MONARD**, Maria Carolina. Estudo e Análise das Diversas Representações e Estruturas de Dados Utilizadas nos Algoritmos de Clustering Hierárquico.Universidade de São Paulo Instituto de Ciências Matemáticas e de Computação. São Carlos -SP. 2006. Disponivel em <http://conteudo.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_269.pdf>

SOBCZACK, G., **PIKULA**, M., **SYDOW**, M. (2012). Agnes: a novel algorithm for visualising diversified graphical entity summarisations on knowledge graphs. In *Foundations of intelligent systems, 20º International symposium, ISMIS 2012, Macau, China, Dezembro 4–7, 2012* (vol. 7661, pp. 182–191). LNCS/Springer. Disponivel em: <http://users.pja.edu.pl/~msyd/papers/agnes-ismis12.pdf>