Curso de Data Mining

Sandra de Amo

Aula 13 - Análise de Clusters - Introdução

Análise de Clusters é o processo de agrupar um conjunto de objetos físicos ou abstratos em classes de objetos similares. Um *cluster* é uma coleção de objetos que são similares uns aos outros (de acordo com algum critério de similaridade pré-fixado) e dissimilares a objetos pertencentes a outros clusters.

As diferenças básicas entre as tarefas de Classificação e Análise de Clusters: Análise de Clusters é uma tarefa que *Aprendizado não-supervisionado*, pelo fato de que os clusters representam classes que não estão definidas no início do processo de aprendizagem, como é o caso das tarefas de Classificação (*Aprendizado Supervisionado*), onde o banco de dados de treinamento é composto de tuplas classificadas. Clusterização constitui uma tarefa de aprendizado *por observação* ao contrário da tarefa de Classificação que é um aprendizado *por exemplo*.

Clusterização Conceitual versus Clusterização Convencional: Na clusterização Conceitual o critério que determina a formação dos clusters é um determinado *conceito*. Assim, uma classe de objetos é determinada por este conceito. Por exemplo, suponhamos que os objetos são indivíduos de uma população e que o critério determinante para se agrupar indivíduos seja o risco de se contrair uma determinada doença. Já na clusterização convencional, o que determina a pertinência dos objetos a um mesmo grupo é a distância geométrica entre eles.

1 Tipos de dados em Análise de Clusters

Alguns algoritmos de análise de clusters operam com os dados organizados numa *matriz de dados* $n \times p$, conforme ilustrado na tabela abaixo:

$$\begin{vmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{vmatrix}$$

Esta matriz é simplesmente a tabela dos dados de treinamento. Cada linha desta tabela representa as coordenadas de um objeto i. Cada coluna representa os valores de um atributo assumidos por cada um dos n objetos.

Por outro lado, muitos algoritmos de clusterização se aplicam em dados organizados numa *matriz de dissimilaridade*, onde o elemento da coluna j e linha i da matriz é o número $d(i, j)$ representando a distância entre os objetos i e j.

$$\begin{vmatrix} 0 & \dots & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d(n,1) & d(n,2) & d(n,3) & \dots & 0 \end{vmatrix}$$

Para que uma função d seja uma *distância* é necessário e suficiente que as seguintes condições sejam satisfeitas, para quaisquer objetos i, j, k:

1. $d(i,j) \geq 0$
2. $d(i,i) = 0$.
3. $d(i,j) = d(j,i)$ (simetria)
4. $d(i,j) \leq d(i,k) + d(k,j)$ (desigualdade triangular)

A propriedade (1) implica que todos os elementos da matriz de dissimilaridade são não-negativos, a propriedade (2) implica que a diagonal da matriz de dissimilaridade é formada por zeros. A propriedade (3), por sua vez, implica que a matriz de dissimilaridade é simétrica com relação à diagonal e por isso, só registramos nela os elementos abaixo da diagonal.

Exercício : O que implica a propriedade (4) da distância, com relação à matriz de dissimilaridade ?

Assim, qualquer função que satisfaz às quatro propriedades acima é chamada de *distância*. As mais importantes funções nesta categoria são:

- Distância Euclidiana :

$$d(i,j) = \sqrt{\mid x_{i_1} - x_{j_1} \mid^2 + \mid x_{i_2} - x_{j_2} \mid^2 + \ldots + \mid x_{i_n} - x_{j_n} \mid^2}$$

- Distância de Manhattan :

$$d(i,j) = \mid x_{i_1} - x_{j_1} \mid + \mid x_{i_2} - x_{j_2} \mid + \ldots + \mid x_{i_n} - x_{j_n} \mid$$

- Distância de Minkowski :

$$d(i,j) = \sqrt[q]{\mid x_{i_1} - x_{j_1} \mid^q + \mid x_{i_2} - x_{j_2} \mid^q + \ldots + \mid x_{i_n} - x_{j_n} \mid^q}$$

  onde $q \geq 1$. Logo, a distância de Minkowski generaliza tanto a distância euclidiana (caso especial onde $q = 2$) quanto a distância de Manhattan (caso especial onde $q = 1$).

Exercício: Sejam $X_1 = (1,2)$ e $X_2 = (4,6)$. Calcule cada uma das 3 distâncias acima entre os objetos X_1 e X_2 (para a de Minkowski considere $q = 3$) e ilustre no plano xy os segmentos representando cada distância e comente as diferenças entre elas.

Às vezes deseja-se ressaltar a importância de certos atributos no cálculo da distância. Para isto, considera-se uma distância *ponderada*, que consiste em se associar pesos a cada uma das coordenadas do objeto. Por exemplo, a distância euclidiana ponderada é dada por :

$$d(i,j) = \sqrt{w_1 \mid x_{i_1} - x_{j_1} \mid^2 + w_2 \mid x_{i_2} - x_{j_2} \mid^2 + \ldots + w_n \mid x_{i_n} - x_{j_n} \mid^2}$$

onde $w_1, ..., w_n$ são os pesos de cada um dos atributos envolvidos na descrição dos objetos.

2 Preparação dos Dados para Análise de Clusters

Como dissemos acima, muitos algoritmos se aplicam à matriz de dissimilaridade dos objetos (só interessam as distâncias relativas entre os objetos e não os valores dos atributos dos objetos). Assim, antes de aplicar o algoritmo é preciso transformar a matriz de dados em uma matriz de dissimilaridade. Os métodos de transformação dependem do tipo de valores que assumem os atributos dos objetos.

2.1 Atributos Contínuos em intervalos

É o caso quando todos os atributos possuem valores que ficam num intervalo contínuo $[a, b]$, como por exemplo, peso, altura, latitude, longitude, temperatura. Os valores são ditos contínuos quando não forem discretizados, isto é, o número de valores assumidos é grande. As unidades de medida utilizadas para medir estes valores (kg, g, metro, cm, ...) podem afetar a análise de clusters. Se a unidade for muito grande (muito grosseira), teremos poucos clusters, se for pequena (muito refinada), teremos muitos clusters. Assim, antes de calcular a distância entre os objetos é preciso *padronizar* os dados. O processo de *padronização* tem como objetivo dar um peso igual a cada um dos atributos. O procedimento para padronizar os dados de uma matriz de dados é o seguinte:

1. Calcula-se o *desvio médio absoluto* para cada atributo A_f :

   $$s_f = \frac{1}{n}(\mid x_{1f} - m_f \mid + \mid x_{2f} - m_f \mid + \ldots + \mid x_{kf} - m_f \mid)$$

   onde m_f = valor médio do atributo A_f. Veja que s_f é um valor associado à *coluna* f da matriz de dados, onde estamos operando com os valores x_{if} da coordenada f de cada objeto X_i.

2. Calcula-se a *medida padrão* ou z-score para o atributo f de cada objeto i:

   $$z_{if} = \frac{x_{if} - m_f}{s_f}$$

   Este é o valor padronizado do elemento x_{if}.

Observamos que o desvio médio absoluto s_f é mais robusto no que diz respeito a ruídos (outliers) do que o desvio médio padrão σ_f:

$$\sigma_f = \frac{1}{n}(\mid x_{1f} - m_f \mid^2 + \mid x_{2f} - m_f \mid^2 + \ldots + \mid x_{kf} - m_f \mid^2)$$

Isto é, se um dos valores aparecendo na coluna f está bem longe da média dos valores (tratando-se portanto de um outlier) seu efeito é amenizado no cálculo do desvio padrão (muito mais do que no cálculo do desvio absoluto).

2.2 Atributos binários

Atributos de tipo binário ou booleano só têm dois valores : 1 ou 0, sim ou não. Tratar valores binários como valores numéricos pode levar a análises de clusters errôneas. Para determinar a matriz de dissimilaridade para valores binários, isto é, determinar $d(i, j)$ para cada par de objetos i, j, vamos considerar primeiramente a *tabela de contingência* para i, j. Nesta tabela:

- q é o número de atributos com valor 1 para i e j
- r é o número de atributos com valor 1 para i e 0 para j
- s é o número de atributos com valor 0 para i e 1 para j
- t é o número de atributos com valor 0 para i e 0 para j
- p é o número total de atributos. Portanto $p = q + r + s + t$.

Tabela de contingência para os objetos i e j

		Objeto j		
		1	0	Soma
Objeto i	1	q	r	q + r
	0	s	t	s + t
	Soma	q + s	r + t	p

Atributos simétricos

Um atributo de tipo booleano é dito *simétrico* se ambos os valores 0 ou 1 são igualmente importantes para a análise de clusters. Por exemplo, o atributo Gênero é simétrico, pois os dois valores M e F são igualmente importantes e além disto, estes dois valores têm a mesma probabilidade de ocorrência. Neste caso, a distância entre i e j é definida como o *coeficiente de simples concordância* :

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

isto é, $d(i, j)$ é a porcentagem de atributos que discordam entre os dois objetos.

Atributos assimétricos

Um atributo de tipo booleano é dito *assimétrico* se existe uma predominância de algum dos valores. Por exemplo, os resultados de um teste para detectar uma doença. Neste caso, o valor mais importante é o mais raro, isto é, teste positivo. Este será o valor 1. Logo, a concordância entre dois 1's é muito mais importante do que a concordância entre dois 0's. Neste caso, a distância entre i e j é definida como sendo o *coeficiente de Jacquard* :

$$d(i,j) = \frac{r+s}{q+r+s}$$

isto é, $d(i,j)$ é a porcentagem de atributos que discordam entre os dois objetos, onde no total de atributos foi desconsiderado aqueles atributos cujos valores concordam e são ambos iguais a 0.

Para ilustrar este cálculo, consideremos o seguinte banco de dados de treinamento:

Nome	Gênero	Febre	Tosse	Teste1	Teste2	Teste3	Teste4
Jack	M	Sim	Não	Pos	Neg	Neg	Neg
Mary	F	Sim	Não	Pos	Neg	Pos	Neg
Jim	M	Sim	Sim	Neg	Neg	Neg	Neg
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Queremos construir clusters de objetos neste banco de dados tais que objetos de um mesmo cluster correspondem a indivíduos sofrendo de uma mesma doença.

O único atributo simétrico é Gênero. Os restantes são assimétricos, pois somente o resultado positivo é importante. Para os objetivos da análise de clusters pretendida, vamos supor que a distância entre objetos é calculada tendo como base somente os atributos assimétricos referentes a resultados de testes e ocorrência de sintomas (febre, tosse). O atributo Gênero não é importante para esta análise. Neste caso, a distância é calculada utilizando o coeficiente de Jacquard. Assim,

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Estas medidas sugerem que Jack e Mary estão mais próximos, portanto, provavelmente serão diagnosticados como sendo portadores de uma mesma doença. Por outro lado, Jim e Mary estão bem distantes. Logo, é bem provável que estejam em clusters distintos, isto é, serão diagnosticados como portadores de doenças distintas.

2.3 Atributos Nominais, Ordinais e Escalonados

Atributos Nominais

Atributo nominal é um atributo discreto, assumindo um número pequeno de valores possíveis. Trata-se de uma generalização dos atributos booleanos, onde o número de valores assumidos é 2. Um exemplo de atributo nominal é Cor, podendo assumir cinco valores: vermelho, amarelo, verde, azul e rosa. Em geral, seja M o número de valores que pode assumir um atributo nominal. Ao invés de denotar os valores por strings, podemos associar a eles números inteiros 1, 2, ... M[1] A distância entre os objetos i e j é medida de maneira análoga como foi feito no caso de atributos booleanos: considerando o coeficiente de coincidência simples:

$$d(i,j) = \frac{p-m}{p}$$

onde p é o número total de atributos e m é o número de atributos onde há coincidências. Assim, $d(i,j)$ é a porcentagem de atributos cujos valores não coincidem. Também podem ser atribuídos pesos a atributos dependendo do número de valores que pode assumir. Por exemplo, suponhamos que tenhamos dois atributos A e B, e que A assuma 5 valores e B 2 valores. Associamos a A o peso 1.5 e a B o peso 0.5. Suponhamos que somente os valores do atributo A coincidam. Então:

$$d(i,j) = \frac{2-1.5}{2} = \frac{0.5}{2} = 0.25$$

Um atributo nominal A assumindo M valores pode ser codificado por um conjunto de atributos binários (booleanos) assimétricos. Cria-se um novo atributo binário para cada um dos M valores que assume A. Se i é um objeto e o valor de A para i é n então os valores dos M atributos $(B_1, B_2, ..., B_M)$ correspondentes a A são : $B_1 = 0, B_2 = 0, ..., B_n = 1, B_{n+1} = 0, ..., B_M = 0$. Com esta transformação de um atributo nominal num atributo binário, a distância entre dois objetos i, j pode ser calculada utilizando o método para atributos binários discutido acima. Repare que esta transformação que fizemos de um atributo nominal em atributo binário é o mesmo que utilizamos nas aulas 10 e 11 para transformar o input de uma rede neural (correspondendo a uma tupla do banco de amostras) num vetor de 0's e 1's.

Atributos Ordinais

Um atributo ordinal é semelhante a um atributo nominal, exceto que os valores assumidos são ordenados, o que não acontece com os atributos nominais. Por exemplo, o atributo TipoMedalha pode assumir os valores nominais Bronze, Prata e Ouro. A estes valores são associados os números 0, 1, 2 respectivamente. A ordem entre os números estabelece uma ordem entre os valores Bronze, Prata, Ouro.

[1] O fato de se ter associado números inteiros aos valores do atributo, não significa que uma ordem entre estes valores foi determinada. O objetivo desta associação é simplesmente de poder tratar valores nominais como sendo números inteiros. A ordem não é considerada. Esta é a diferença fundamental entre atributos nominais e atributos ordinais que veremos mais adiante.

1. Seja x_{if} o valor do atributo A_f do i-ésimo objeto e suponha que estes valores podem ser mapeados numa escala crescente $0, 1, ..., M_f - 1$, onde M_f é o total de valores que pode assumir o atributo A_f. Substitua cada x_{if} pela sua correspondente posição r_{if} na escala $0, 1, ..., M_f - 1$. Por exemplo, se o atributo A_f é TipoMedalha e seus valores são {Bronze, Prata, Ouro} então a escala é : Bronze $\rightarrow$ 0, Prata $\rightarrow$ 1, Ouro $\rightarrow$ 2. Aqui, $M_f = 3$.

2. Como cada atributo ordinal tem um número distinto de valores possíveis, é frequentemente necessário mapear estes valores para o intervalo $[0, 1]$ de tal maneira que cada atributo tenha um peso igual no cálculo da distância. Isto pode ser conseguido, substituindo-se o número inteiro r_{if} por:

$$z_{if} = \frac{r_{if}}{M_f - 1}$$

3. Uma vez feita esta transformação de cada valor inicial x_{if} em z_{if} procede-se ao cálculo da distância entre os objetos i e j utilizando uma das funções distâncias discutidas anteriormente (Euclidiana, Manhattan, Minkowski). Por exemplo, a distância euclidiana entre os objetos i e j é dada por:

$$d(i, j) = \sqrt{\mid z_{i1} - z_{j1} \mid^2 + \mid z_{i2} - z_{j2} \mid^2 + \ldots + \mid z_{in} - z_{jn} \mid^2}$$

Atributos escalonados não-lineares

Atributos escalonados não-lineares são como os atributos contínuos em intervalos. A diferença entre eles é que um atributo contínuo em intervalo representa uma medida segundo uma escala *linear* (temperatura, longitude, peso, altura, etc). Já um atributo escalonado não-linear representa uma medida segundo uma escala não-linear, na maioria das vezes uma escala exponencial, segundo a fórmula Ae^{Bt} ou Ae^{-Bt}, onde A e B são constantes positivas. Por exemplo, o crescimento de uma população de bactérias ou a desintegração de um elemento radioativo são conceitos medidos de acordo com uma escala exponencial.

Existem três maneiras para se calcular a dissimilaridade $d(i, j)$ entre dois objetos i e j onde todos os atributos são escalonados não-lineares:

1. Trata-se os atributos escalonados não-lineares da mesma forma como se tratou os atributos contínuos em intervalos. Esta não é uma boa maneira pois nos atributos contínuos em intervalos, a escala é linear. Logo, é bem possível que, tratando-se atributos escalonados não-lineares como se fossem lineares, a escala seja distorcida.

2. Aplica-se uma transformação logarítmica ao valor x_{if} de cada atributo A_f de um objeto i, obtendo $y_{if} = log(x_{if})$. Agora, os valores y_{if} podem ser tratados como se fossem valores contínuos em intervalos (medidos segundo uma escala linear). Repare que, dependendo de como foi escalonado o valor, outras transformações poderão ser empregadas. Nestes exemplo, utilizamos a função *log* já que é a inversa da função exponencial[2].

[2]No exemplo, foi esta a função utilizada no escalonamento dos valores do atributo.

3. Trata-se os valores x_{if} como se fossem valores ordinais contínuos. Associa-se a cada valor um número entre 0 e $M_f - 1$, onde M_f é o número total de valores assumidos pelo atributo A_f (este número M_f, ao contrário do que acontece com os atributos ordinais, pode ser muito grande). Uma vez feita esta associação, trata-se os valores associados da mesma maneira como tratamos os atributos contínuos em intervalos.

Os dois últimos métodos são os mais eficazes. A escolha de um ou outro método depende da aplicação em questão.

2.4 Atributos Mistos

Nas seções anteriores, foi discutido como calcular a matriz de dissimilaridade entre objetos, considerando que todos os atributos são do mesmo tipo. Na realidade, um objeto possui atributos cujos tipos variam, podendo assumir tipos dos mais variados, entre todos os tipos que consideramos anteriormente.

Como calcular a dissimilaridade $d(i, j)$ entre objetos i, j, onde onde atributos são de tipos distintos ?

Enfoque de agrupamento: Neste enfoque, agrupa-se os atributos de mesmo tipo em grupos. Para cada grupo de atributos, faz-se a análise de clusters dos objetos (somente considerando os atributos do grupo, os outros são desconsiderados). Assim, teremos tantas análises de clusters quanto for o número de tipos de atributos do banco de dados de objetos. Se os resultados de cada análise são compatíveis, isto é, se objetos que são similares numa análise, continuam similares em outra análise e objetos dissimilares segundo uma análise são dissimilares segundo outra análise, então este método é factível. Entretanto, em aplicações reais, é muito improvável que os resultados de análises separadas sejam compatíveis.

Enfoque da uniformização da escala dos valores: Uma técnica que é muito utilizada consiste em transformar todos os valores dos diferentes atributos em valores de uma escala comum no intervalo [0,1]. Suponha que o banco de dados contém p atributos de tipos mistos. A dissimilaridade $d(i, j)$ entre os objetos i, j é definida por:

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{f} d_{ij}^{f}}{\sum_{f=1}^{p} \delta_{ij}^{f}}$$

onde:

- $\delta_{ij}^{f} = 0$ se uma das possibilidades abaixo ocorre:
  1. os valores x_{if} ou x_{jf} são incompletos (do tipo NULL, isto é, não foram fornecidos),
  2. o atributo A_f é booleano assimétrico e $x_{if} = x_{jf} = 0$.
- $\delta_{ij}^{f} = 1$, caso nenhuma das condições acima ocorrem.

Os números d_{ij}^{f} representam a contribuição do atributo A_f no cálculo da dissimilaridade entre os objetos i e j. O cálculo deste número depende do tipo do atributo A_f:

1. Se o atributo A_f é booleano ou nominal : $d_{ij}^f = 0$ se $x_{if} = x_{jf}$. Caso contrário, $d_{ij}^f = 1$.

2. Se o atributo A_f é contínuo em intervalo:

$$d_{ij}^f = \frac{\mid x_{if} - x_{jf} \mid}{max_h\{x_{hf}\} - min_h\{x_{hf}\}}$$

   onde h varia entre todos os objetos onde o atributo f não é incompleto, isto é, seu valor não é NULL.

3. Se o atributo A_f é ordinal ou escalonado não-linear: calcula-se os inteiros r_{if} associados ao valor x_{if} e considera-se $z_{if} = \frac{r_{if}}{M_f - 1}$. A partir daí, trata-se os valores z_{if} como se fossem contínuos em intervalos, e calcula-se a dissimilaridade d_{ij}^f de acordo.

Exercício: Dê uma maneira mais refinada de se calcular a contribuição d_{ij}^f do atributo A_f, no caso deste atributo ser do tipo contínuo em intervalo, utilizando a padronização dos dados do atributo A_f (ver seção 2.1). Compare esta maneira com a maneira acima. Qual em sua opinião é a melhor ?