Aula 4 – Mineração de Dados
Medidas de Distância

Profa. Elaine Faria

UFU

Agradecimentos

Este material é baseado

- No livro Tan et al, 2006
- Nos slides do prof Andre C. P. L. F. Carvalho

- Agradecimentos
  - Ao professor André C. P. L. F. Carvalho que gentilmente cedeu seus slides

Medidas de Similaridade e Dissimilaridade

- Importância
  - São usadas em uma série de técnicas de MD e AM. Ex: agrupamento, KNN e detecção de novidade

- Pode ser visto com uma transformação dos dados para um espaço de similaridade (dissimilaridade)
  - Em muitos casos o conjunto de dados inicial não é necessário para executar a técnica de MD → apenas as medidas de similaridade ou dissimilaridade são suficientes

- Proximidade entre objetos refere-se à proximidade entre seus atributos

Medidas de Similaridade e Dissimilaridade

- Similaridade entre dois objetos
  - É uma medida numérica do quão parecido dois objetos são
    - Objetos parecidos → similaridade alta
  - É um número não negativo entre 0 (não similar) e 1 (completamente similares)

- Dissimilaridade entre dois objetos
  - É uma medida numérica do quão diferente dois objetos são
    - Objetos similares → dissimilaridade baixa
  - Está no intervalo [0,1] ou [0, ∞]
  - Distância é tipo especial de dissimilaridade

Medidas de Similaridade e Dissimilaridade

- Transformação
  - Converter similaridade para dissimilaridade ou vice-versa
  - Transformar uma medida de proximidade para um intervalo particular, ex: [0,1]

    Ex: medida de similaridade no intervalo [1,10], mas o algoritmo só trabalha com similaridade entre [0,1] → aplicar transformação

$$s' = (s - min_s)/ (max_s - min_s)$$

$$s' = (s - 1)/9$$

Medidas de Similaridade e Dissimilaridade

- Transformação: similaridade para dissimilaridade
  - Se está no intervalo [0,1]

    $d = 1 - s$ (ou $s = d - 1$)

  - Se não está no intervalo [0,1]

    $s = 1/(d+1)$, $s = e^{(-d)}$, $s = 1 - ((d - min)/(max - min))$

Similaridade e Dissimilaridade entre Atributos Simples

- Proximidade com 1 atributo

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Tabela retirada dos slides do prof. André C. P. L. F. Carvalho – disciplina Aprendizado de Máquina – ICMC-USP

Dissimilaridade entre Objetos

- Existem várias medidas de dissimilaridade
  - Diferentes medidas podem ser aplicadas a diferentes problemas
- Objetos (ou Instâncias) são descritos por *n* atributos
  - Calcular a medida de dissimilaridade usando os *n* atributos
  - Em geral, usa-se medidas de distância
- Distância
  - Medida de dissimilaridade que possui certas propriedades (ver slide 12)

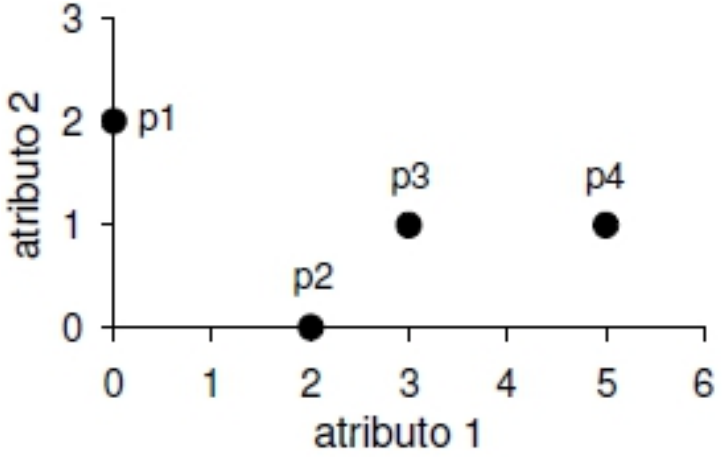
Medidas de Distância

- Distância Euclidiana
  - Distância *d* entre dois objetos x e y em um espaço *n* dimensional

$$d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

X_k e y_k são o k-ésimo atributo dos objetos x e y

Medidas de Distância



	atributo 1	atributo 2
p1	0	2
p2	2	0
p3	3	1
p4	5	1

matriz de distância

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Imagem retirada dos slides do prof. Ricardo J. G. B. Campello – disciplina Aprendizado de Máquina – ICMC-USP

Medidas de Distância

- Distância de Minkowski
  - Generalização da distância Euclidiana

$$d(x,y) = \left(\sum_{k=1}^{n} | x_k - y_k |^r \right)^{\frac{1}{r}}$$

  - r = 1 : distância *city block* (*Manhattan* ou L_1 norm)
  - r = 2 : distância Euclidiana (L_2 norm)
  - r = ∞ : distância Suprema (L_{max} ou L_∞ norm)

$$d(x,y) = \lim_{r \to \infty} \left(\sum_{k=1}^{n} | x_k - y_k |^r \right)^{\frac{1}{r}}$$

Propriedades das distâncias

- Positividade
  - d(x,y) >=0 para todo x e y
  - d(x,y)=0 somente se x = y
- Simetria
  - d(x,y) = d(y,x) para todo x e y
- Desigualdade triangular
  - d(x,z) <= d(x,y) + d(y,z) para todos os objetos x, y e z.

Propriedades das distâncias

- Medidas que satisfazem as 3 propriedades → métricas

- Ex. de medida de dissimilaridade que não é métrica

  Conjuntos A e B

  A – B: elementos que estão em A e não estão em B

  dist(A,B) = tamanho (A – B)

  Ex: A={1,2,3,4} B={2,3,4}

  Não atende a 2ª parte da propriedade da positividade, nem a simetria, nem a desigualdade triangular.

Similaridade entre Objetos

- Propriedades
  - s(x,y) = 1 somente se x = y (0 <= s <= 1)
  - s(x,y)=s(y,x) para todo x e y

  - Não há uma propriedade análoga à desigualdade triangular para medidas de similaridade

Medidas de Proximidade

- Medidas de similaridade para vetores binários
  - Chamadas de coeficiente de similaridade
  - Possuem valores entre 0 e 1 → 1: objetos completamente similares, 0: objetos não similares
  - Comparando objetos x e y que consistem de *n* atributos binários (vetores binários)
    - f_{00} = nro de atributos em que x=0 e y=0
    - f_{01} = nro de atributos em que x=0 e y=1
    - f_{10} = nro de atributos em que x=1 e y=0
    - f_{11} = nro de atributos em que x=1 e y=1

Medidas de Proximidade

- Medidas de similaridade para vetores binários: **Coeficiente de casamento simples**

$$SMC = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{00} + f_{11}}$$

  - Conta as presenças e ausências igualmente
  - Ex: encontrar os estudantes que responderam de forma similar a um teste que consiste de questões true/false.

Medidas de Proximidade

- Medidas de similaridade para dados binários: **Coeficiente de Jaccard**

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

  - Usado para atributos binários assimétricos
  - Não considera as coincidências de 0s

Medidas de Proximidade

- Ex:

x = (1,0,0,0,0,0,0,0,0,0)
y = (0,0,0,0,0,0,1,0,0,1)

f_{01} = 2 número de atributos em que x= 0 e y=1
f_{00} = 7 número de atributos em que x= 0 e y=0
f_{10} = 1 número de atributos em que x= 1 e y=0
f_{11} = 0 número de atributos em que x= 1 e y=1

SMC = 0 + 7/ (2+1+0+7) = 0.7
J = 0/ (2+1+0) = 0

Medidas de Proximidade

- Similaridade Cosseno
  - É uma medida do ângulo entre x e y. Se a similaridade é 1, o ângulo entre x e y é 0^0; se a similaridade é 0, o ângulo é 90^0

$$\cos(x, y) = \frac{x.y}{\| x \| \| y \|}$$

$$x.y = \sum_{k=1}^{n} x_k y_k$$

$$\| x \| = \sqrt{\sum_{k=1}^{n} x_k^{\,2}}$$

. → produto interno de dois vetores,

||x|| → é o tamanho (norma) do vetor x

Medidas de Proximidade

- Similaridade Cosseno

Ex. Sejam os vetores

x = (3,2,0,5,0,0,0,2,0,0)

y = (1,0,0,0,0,0,0,1,0,2)

x.y = 3*1+2*0+0*0+5*0+0*0+0*0+0*0+2*1+0*0+0*2 = 5

||x|| = sqrt(3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0) = 6.48

||y|| = sqrt(1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) = 2.24

cos(x,y) = **0.31**

Medidas de Proximidade

- Similaridade Cosseno
  - Muito usado em mineração de texto
    - Documentos são vetores, cada atributo representa a frequência de ocorrência de um termo (palavra) no documento
    - Cada documento é esparso (poucos atributos não zero)

Tarefa

- Leitura do Capítulo 2 (Seção 2.4) do livro Tan et al, 2006

Referências

- Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006.