Distâncias para dados categóricos

George Darmiton da Cunha Cavalcanti
Tiago Buarque Assunção de Carvalho
{gdcc, tbac}@cin.ufpe.br
CIn - UFPE

Tipos de Dados

- Dados Numéricos (Escala Racional)
  - Permite operações matemáticas, tais como: soma, subtração, multiplicação, divisão, igualdade, maior que e menor que
  - Exemplos: peso e altura
  - Pode-se medir a dissimilaridade usando distância Euclidiana
- Dados Categóricos (Escala Nominal)
  - Permite apenas operações de igualdade e diferença
  - Exemplos: sexo, estado civil e cor do olhos

Distância de Hamming

- Como calcular a distância entre dados categóricos?
- Uma alternativa é a Distância de Hamming – DH(a, b)

$$DH(a,b) = \sum_{i=1}^{n} dh(a_i, b_i)$$

$$dh(a_i, b_i) = \begin{cases} 0, a_i = b_i \\ 1, a_i \neq b_i \end{cases}$$

Um exemplo

- Um atributo que define a cor possui três valores: vermelho, verde e azul
- O objetivo é identificar se um objeto é uma maçã ou não
- Assim, vermelho e verde devem ser considerados mais próximos do que vermelho e azul
- Pois vermelho e verde possuem correlação com a classe desejada: maçã.



VDM – *Value Difference Metric*

- Distância entre dados categóricos
- Descobre quando dois valores tem distribuição igual entre as classes
- Geralmente mais preciso do que Hamming

$$VDM(a,b) = \sqrt{\sum_{i=1}^{n} vdm_i(a_i, b_i)}$$

VDM – *Value Difference Metric*

$$VDM(a,b) = \sqrt{\sum_{i=1}^{n} vdm_i(a_i, b_i)}$$

$$vdm_i(a_i, b_i) = \sum_{c=1}^{C} \left| \frac{N_{i,a,c}}{N_{i,a}} - \frac{N_{i,b,c}}{N_{i,b}} \right|^q = \sum_{c=1}^{C} \left| P_{i,a,c} - P_{i,b,c} \right|^q$$

- $\mathbf{N_{i,a}}$ é o número de instâncias no conjunto de treinamento que tem o valor $\mathbf{a_i}$ para o atributo **i**
- $\mathbf{N_{i,a,c}}$ é o número de instâncias no conjunto de treinamento que tem o valor $\mathbf{a_i}$ para o atributo **i** e pertence à classe **c**
- **C** é o número de classes
- **q** é uma constante, geralmente 1 ou 2
- $\mathbf{P_{i,a,c}}$ é probabilidade condicional do padrão pertencer à classe c dado que possui o atributo $\mathbf{a_i}$ na posição **i**: $\mathbf{P(c|a_i)}$

Usando $vdm_a(x,y)$, dois valores serão considerados próximos se eles possuem classificações similares (i.e., correlações similares entre as classes), não importando a ordem dos valores.

Exemplo

- Audiology
  - 24 classes
  - 69 atributos categóricos por padrão
  - Base desbalanceada: algumas classes têm muitos padrões e outras têm poucos padrões
  - Fonte: UCI

	KNN - [sem peso]		KNN - [com peso]	
	Hamming	VDM	Hamming	VDM
k = 1	72,80	77,45	72,80	77,45
k = 2	64,60	69,55	73,35	77,45
k = 3	66,70	71,35	74,00	77,45
k = 5	64,40	65,30	75,40	**77,60**
k = 6	61,75	61,25	74,95	76,70
k = 11	59,00	54,25	72,60	74,90
k = 16	53,70	53,85	72,35	70,50
k = 21	50,40	48,40	69,50	68,05
k = 31	47,05	46,30	65,70	63,70

VDM aplicado a dados numéricos

- Seria possível aplicar VDM a dados numéricos?
- Como fazer isso?
  - **Discretização**: transformar os números em dados categóricos
  - **Interpolação**: definir alguns valores discretizados e interpolar $\mathbf{P_{i,a,c}}$ com base na probabilidades dos valores discretizados

Discretização

- Transformando números em categorias
- Problema:
  - Número de categorias pode ser muito grande, o que não dá qualquer informação
  - e.g., todas as instâncias do conjunto de treino possuem atributos Reais e cada atributo gera uma nova categoria
- Solução:
  - Gerar um número de categorias **S** que dê informação útil sobre a distribuição

Discretização

- Define-se um número de categorias **S**
- Os valores numéricos serão convertido em **S** categorias de **0** a **S-1**
- **S=10** apresentou bons resultados em alguns testes, mas depende da base de dados

$$discretize_i(a_i) = \begin{cases} s, \; se \; a_i = \max_i \\ \left\lfloor \dfrac{(a_i - \min_i)}{\omega_i} \right\rfloor, \; \text{se} \;\; a_i \neq \max_i \end{cases}$$

$$\omega_i = \frac{\left|\max_i - \min_i\right|}{s}$$

VDM discretizado

- Para dados numéricos
- Discretiza os dados numéricos e aplica VDM

$$VDM_{\text{discretizado}}(a,b) = \sqrt{\sum_{i=1}^{n} \left| vdm_i(discretize_i(a_i), discretize_i(b_i)) \right|^2}$$

$$discretize_i(a_i) = \begin{cases} s, \text{ se } a_i = \max_i \\ \left\lfloor \dfrac{(a_i - \min_i)}{\omega_i} \right\rfloor, \text{se } a_i \neq \max_i \end{cases}$$

$$\omega_i = \frac{\left| \max_i - \min_i \right|}{s}$$

VDM interpolado

- Para dados numéricos
- Interpola as probabilidades de cada valor numérico baseado nas probabilidades discretizadas

$$VDM_{\text{interpolado}}(a,b) = \sqrt{\sum_{i=1}^{n} \left| p_{i,c}(a_i) - p_{i,c}(b_i)) \right|^2}$$

$$p_{i,c}(a_i) = P_{i,u,c} + \left(\frac{u - mid_{i,u}}{mid_{i,u+1} - mid_{i,u}} \right) \times \left(P_{i,u+1,c} - P_{i,u,c} \right)$$

$$u = discretize(a_i)$$

$$mid_{i,u} \leq a_i < mid_{i,u+1}$$

$$mid_{i,u} = discretize^{-1}(u)$$