Distâncias Heterogêneas

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

Bases de Dados Heterogêneas

- E se o vetor de atributos da base não for puramente categórico nem puramente numérico?

- **Solução 1**
  - Converter todos valores categórico para números ou todos os valores numéricos para categóricos

- **Solução 2**
  - **Distâncias Heterogêneas**
  - Fazer uso de distâncias que trabalhem com os dois tipos de dados ao mesmo tempo.

Distâncias Heterogêneas

Combina duas funções de distâncias: uma pra atributos numéricos e outra para atributos categóricos.

HEOM (Euclidiana + Hamming)

HVDM (Euclidiana + VDM)

DVDM ($VDM_{discretizado}$ + VDM)

IVDM ($VDM_{interpolado}$ + VDM)

HEOM
(*Heterogeneous Euclidian-Overlap Metric*)

- Distância Euclidiana para atributos numéricos
- Distâncias de Hamming para atributos categóricos

$$HEOM(x, y) = \sqrt{\sum_{a=1}^{n} heom_a(x_a, y_a)^2}$$

$$heom_a(x_a, y_a) = \begin{cases} 1, \text{ se } x_a \otimes y_a \text{ é desconhecido} \\ 0, \text{ se } x_a \wedge y_a \text{ são desconhecidos} \\ h_a(x_a, y_a), \text{ se } a \text{ é categórico} \\ dif_a(x_a, y_a), \text{ se } a \text{ é numérico} \end{cases}$$

$$h_a(x_a, y_a) = \begin{cases} 1, \text{ se } x_a \neq y_a \\ 0, \text{ se } x_a = y_a \end{cases}$$

$$dif_a(x_a, y_a) = \frac{|x_a - y_a|}{\max(a) - \min(a)}$$

HVDM
(*Heterogeneous Value Difference Metric*)

- Distância Euclidiana para atributos numéricos
- VDM para atributos categóricos

$$HVDM(x, y) = \sqrt{\sum_{a=1}^{n} hvdm_a(x_a, y_a)^2}$$

$$hvdm_a(x_a, y_a) = \begin{cases} 1, \text{ se } x_a \otimes y_a \text{ é desconhecido} \\ 0, \text{ se } x_a \wedge y_a \text{ são desconhecidos} \\ vdm_a(x_a, y_a), \text{ se } a \text{ é categórico} \\ dif_a(x_a, y_a), \text{ se } a \text{ é numérico} \end{cases}$$

$$vdm_a(x, y) = \sum_{c=1}^{C} \left| P_{a,x,c} - P_{a,y,c} \right|^q$$

$$dif_a(x_a, y_a) = \frac{\left| x_a - y_a \right|}{\max(a) - \min(a)}$$

Centro de Informática U·F·P·E

DVDM
(*Discretized Value Difference Metric*)

- $VDM_{discretizado}$ para atributos numéricos
- VDM para atributos categóricos

$$DVDM(x, y) = \sqrt{\sum_{a=1}^{n} dvdm_a(x_a, y_a)^2}$$

$$dvdm_a(x_a, y_a) = \begin{cases} 1, \text{ se } x_a \otimes y_a \text{ é desconhecido} \\ 0, \text{ se } x_a \wedge y_a \text{ são desconhecidos} \\ vdm_a(discretize(x_a), discrestize(y_a)), \text{ para os outros casos} \end{cases}$$

$$vdm_a(x, y) = \sum_{c=1}^{C} \left| P_{a,x,c} - P_{a,y,c} \right|^q$$

IVDM
(*Interpolated Value Difference Metric*)

- $VDM_{interpolado}$ para atributos numéricos
- VDM para atributos categóricos

$$IVDM(x, y) = \sqrt{\sum_{a=1}^{n} ivdm_a(x_a, y_a)^2}$$

$$ivdm_a(x_a, y_a) = \begin{cases} 1, \text{ se } x_a \otimes y_a \text{ é desconhecido} \\ 0, \text{ se } x_a \wedge y_a \text{ são desconhecidos} \\ vdm_a(x_a, y_a), \text{ se } a \text{ é categórico} \\ \sum_{c=1}^{C} \left| P_{a,c}(x_a) - P_{a,c}(y_a) \right|^2, \text{ se } a \text{ é numérico} \end{cases}$$

$$P_{a,c}(x_a) = P_{a,u,c} + \left(\frac{x - meio_{a,u}}{meio_{a,u+1} - meio_{a,u}} \right) \times \left(P_{a,u+1,c} - P_{a,u,c} \right)$$

$$u = discretize(x_a)$$

$$meio_{a,u} = \min(a) + \left(\max(a) - \min(a) \right) \times \left(u + 0.5 \right)$$

$$vdm_a(x, y) = \sum_{c=1}^{C} \left| P_{a,x,c} - P_{a,y,c} \right|^q$$

Centro de Informática U·F·P·E

Exemplo

- Hepatitis
  - 2 classes, 155 padrões (32 + 123)
  - 20 atributos numéricos e categóricos
  - Algumas instâncias possuem atributos com valor desconhecido

KNN - [sem peso]

	k = 1	k = 2	k = 3	k = 5	k = 6	k = 11	k = 16	k = 21	k = 31
HEOM	59,68	62,39	62,77	66,9	66	64,97	65,94	**66,97**	**68,77**
HVDM	**59,16**	**64,19**	**62,71**	**63,87**	**65,68**	**66,06**	**66,65**	**67,48**	**63,87**
DVDM	57,55	60,84	60,13	61,55	63,03	63,42	63,68	63,94	62,84
IVDM	**58,77**	**63,1**	**63,81**	**65,16**	**64**	**63,16**	**64,65**	**64,65**	**63,42**

KNN - [com peso]

Hepatt	k = 1	k = 2	k = 3	k = 5	k = 6	k = 11	k = 16	k = 21	k = 31
HEOM	59,68	59,68	59,68	60,97	61,48	64	**66,65**	65,74	**67,61**
HVDM	**59,16**	**59,16**	**59,16**	**62,84**	**63,94**	**63,42**	**65,29**	**65,74**	**67,23**
DVDM	57,55	57,29	57,29	58,84	59,03	61,87	62,65	63,1	63,23
IVDM	**58,77**	**58,77**	**58,77**	**61,42**	**62,19**	**63,81**	**64,32**	**64,39**	**64,06**

Referências

- D. R. Wilson and T. R. Martinez, “Improved Heterogeneous Distance Functions,” **Journal of Artificial Intelligence Research**, vol.6, pp.1-34, 1997.
- Hui Wang, “Nearest Neighbor by Neighborhood Counting”, **IEEE Transactions on Pattern Analysis and Machine Intelligence**, vol.28, no.6, pp. 942-953, 2006.
- UCI Machine Learning Repository
  - http://archive.ics.uci.edu/ml/
  - Repositório de base de dados amplamente utilizada para comparar resultados