Data Mining

Analyse et Fouille de données

Dr. Sana Hamdi sana.hamdi@fst.utm.tn

Similarité et dissimilarité

- Mesure de dissimilarité: plus la mesure est faible, plus les points sont similaires
- Mesure de similarité: plus la mesure est grande, plus les points sont similaires
- Métrique de similarité/dissimilarité : exprimée en termes d'une fonction de distance, typiquement d(i,j)
- La fonction de distance dépend du type des données : binaires, nominales, ordinales ou continues
- Difficulté de définir « *suffisamment similaires* » → la réponse est très subjective

Types de données

- Continue sur un intervalle (poids, taille, salaire)
- Binaire
- Nominale (couleur, profession)
- Ordinale (Mention, taille vestimentaire, stade d'une maladie)
- Mixte (différents types)

Comment procéder pour calculer la similarité entre ces observations à regrouper ???

Valeurs continues sur un intervalle: normalisation

- Transformer en valeurs entre 0 et 1: $x'_{if} = \frac{x_{if} - \min x_f}{\max x_f - \min x_f}$
- Exemple:
  - max distance du salaire : 11122-11000 = 122
  - max distance de l'âge : 70-50 = 20
  - $x'_{1\,salaire} = \frac{11000-11000}{122} = 0$
  - $x'_{3\,age} = \frac{60-50}{20} = 0{,}5$

	Age	Salaire
Personne 1	0	0
Personne 2	1	0,82
Personne 3	0,5	1
Personne 4	0,5	0,6

	Age	Salaire
Personne 1	50	11000
Personne 2	70	11100
Personne 3	60	11122
Personne 4	60	11074

Valeurs continues sur un intervalle: standardisation

- Standardiser les données: Égaliser le poids des variables pour assurer l'indépendance par rapport aux unités de mesures

1. Calculer l'écart absolu moyen:

$$S_f = \frac{1}{n}(|x_{1f} - mf| + |x_{2f} - mf| + \cdots + |x_{nf} - mf|)$$

Où

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \cdots + x_{nf})$$

2. Calculer la mesure standardisée (***z-score***):

$$z_{if} = \frac{x_{if} - mf}{S_f}$$

Valeurs continues sur un intervalle: standardisation

	Age	Salaire
Personne 1	50	11000
Personne 2	70	11100
Personne 3	60	11122
Personne 4	60	11074

$m_{Age} = 60$

$$S_f = \frac{1}{4}(|50 - 60| + |70 - 60| + |60 - 60| + |60 - 60|) = 5$$

$m_{Salaire} = 11074$ $S_{Salaire} = 148$ $/4=37$

	Age	Salaire
Personne 1	-2	-2
Personne 2	2	0,702
Personne 3	0	1,287
Personne 4	0	0

Valeurs continues sur un intervalle: Fonctions de distance

- Souvent, les distances sont utilisées
- Propriétés des distances :
  - Symétrie : Pour tout A et tout B, $d(A, B) \geq 0$, and $d(A, B) = d(B, A)$
  - Pour tout A, $d(A, A) = 0$
  - Inégalité triangulaire : $d(A, C) \leq d(A, B) + d(B, C)$

Valeurs continues sur un intervalle: Fonctions de distance

- Distance de Minkow:

$$d(i,j)=\sqrt[q]{|x_{i1}-x_{j1}|^q+|x_{i2}-x_{j2}|^q+\cdots+|x_{ip}-x_{jp}|^q}$$

avec $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ et $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ deux objets à p dimensions, et q un entier positif

- si q = 1 : distance de Manhattan

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\cdots+|x_{ip}-x_{jp}|$$

- si q = 2 : distance euclidienne

$$d(i,j)=\sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\cdots+|x_{ip}-x_{jp}|^2}$$

Valeurs binaires

- Il est nécessaire de déterminer d'abord la table de contingence

		Objet j	
		1	**0**
Objet i	**1**	a	b
	0	c	d

a= nombre de positions où
Objet i possède 1 et Objet j possède 1

- Exemple: Objet i=(1,1,0,1,0) et Objet j=(1,0,0,0,1)
- a=1, b=2, c=1, d=1

Valeurs binaires: Mesures de similarité

		Objet j	
		1(Présence)	**0 (Absence)**
Objet i	**1 (Présence)**	a	b
	0 (Absence)	c	d

- Coefficient d'appariement simple: **S(i,j) = (a+d)/(a+b+c+d)**
- Coefficient de Russel et Rao: **S(i,j) = a/(a+b+c+d)**
- Coefficient de Jaccard: **S(i,j) = a/(a+b+c)**
- Coefficient de Dice**: S(i,j) = 2a/(2a+b+c)**

Valeurs binaires: Mesures de dissimilarité

		Objet j	
		1(Présence)	**0 (Absence)**
Objet i	**1 (Présence)**	a	b
	0 (Absence)	c	d

- coefficient d'appariement simple (invariant si la variable est symétrique)

$$d(i,j) = b+c/(a+b+c+d)$$

- Coefficient de Jaccard

$$d(i,j) = b+c/(a+b+c)$$

Valeurs binaires

- Une variable binaire peut être soit:

1. Variable symétrique: Ex. le sexe d'une personne, i.e coder masculin par 1 et féminin par 0 c'est pareil que le codage inverse (coder masculin par 0 et féminin par 1)
2. Variable asymétrique: Ex. Test B. Le test peut être positif ou négatif (0 ou 1) mais il y a une valeur qui sera plus présente que l'autre. Généralement, on code par 1 la modalité la moins fréquente
   - 2 personnes ayant la valeur 1 pour le test sont *plus similaires* que 2 personnes ayant 0 pour le test

Valeurs binaires: Exemple

	Sexe	Fièvre	Tousse	Test-1	Test-2	Test-3	Test-4
Jaques	M	O	N	P	N	N	N
Marie	F	O	N	P	N	P	N
Jean	M	O	P	N	N	N	N

- Sexe est symétrique
- les autres sont asymétriques, soit O et P = 1, et N = 0, la distance n'est mesurée que sur les asymétriques
  - d (Jaques, Marie) = 0+1/2+0+1 = 0,33
  - d (Jaques, Jean) = 1+1/1+1+1 = 0,67
  - d (Jean, Marie) = 1+2/1+1+2 =0,75
- Les plus similaires sont Jaques et Marie : atteints de la même maladie

Variables nominales

❖ Généralisation des valeurs binaires : plus de deux états possibles

- **Méthode 1:** Appariement (matching) simple
  - m : nombre d'appariements
  - p : nombre total de variables

$$d(i, j) = (p - m)/p$$

- **Méthode 2:** utiliser un grand nombre de variables binaires
  - Créer une variable binaire pour chaque modalité (ex: variable rouge qui prend les valeurs vrai ou faux)

Variables ordinales

- L'ordre est important : rang
- Peut être traitée comme une variable continue sur un intervalle
- remplace x_{if} par son rang: $r_f \in \{1, \ldots, M_f\}$
- transforme chaque variable sur [0,1] en remplaçant le *i*-ème objet de la *f*-ème variable

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

- calcule la dissimilarité en utilisant les méthodes de valeurs continues sur un intervalle

Variables mixtes

- Lorsque les observations ont des variables de différents types

→ On utilise une formule pondérée pour faire la combinaison

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

- $\boldsymbol{\delta_{ij}^{f}}$ **= 0 ou 1**
- $\boldsymbol{\delta_{ij}^{f}}$ **= 0** si * i ou j admet des données manquantes

  * $x_i^f = x_j^f = 0$ et $\boldsymbol{f}$ une variable binaire asymétrique
- $\boldsymbol{\delta_{ij}^{f}}$ **= 1** sinon

Variables mixtes

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

- Si f est binaire ou nominale: $d_{ij}^{f} = 0$ si $x_i^f = x_j^f$,

  Sinon $d_{ij}^{f} = 1$
- Si f est de type intervalle: utiliser une distance normalisée
- Si f est ordinale
  - calculer les rangs r_{if} et
  - Ensuite traiter z_{if} comme une variable de type intervalle: $z_{if} = \frac{r_{if} - 1}{M_f - 1}$

Le clustering

- Deux méthodologies générales: Algorithmes de partitionnement et Algorithmes hiérarchiques
- Partitionnement
  - Diviser un ensemble de N instances en K clusters
- Hiérarchique
  - Par agglomérations : les paires d' instances ou de clusters sont successivement liés pour produire des clusters plus grands (bottom-up)
  - Par divisions : commencer par l'ensemble entier comme cluster et successivement diviser en de plus petites partitions (top-down)