

Истраживање података

Ненад Митић

Математички факултет
`nenad.mitic@matf.bg.ac.rs`

Асиметрични атрибути

Једино се присуство не-нула вредности сматра значајним

- На пример, нека је објекат студент чији су атрибути информација да ли је слушао неки од курсева који се држе на факултету (1-слушао, 0-није слушао)
- Када су два студента слична по курсевима које су слушали?

Бинарни атрибути код којих су битне не-нула вредности се зову асиметрични бинарни атрибути

Мултидимензионални подаци

- Најједноставнији облик независних података
- Састоје се од слогова (инстанци, трансакција, ентитета, торки, објекта, вектора-особина,...)
- Слог се састоји од поља (атрибути, димензије, ...)
- Примери
 - Квантитативни (температура, висина, тежина, притисак, брзина)
 - Категорички (ЈМБГ, боја очију, поштански број, радно место)
 - Бинарни
 - Текстуални

Мултидимензионални подаци

x_1^1	x_1^2	x_1^3	x_1^4	...	x_1^d
x_2^1	x_2^2	x_2^3	x_2^4	...	x_2^d
x_3^1	x_3^2	x_3^3	x_3^4	...	x_3^d
x_4^1	x_4^2	x_4^3	x_4^4	...	x_4^d
x_5^1	x_5^2	x_5^3	x_5^4	...	x_5^d
x_6^1	x_6^2	x_6^3	x_6^4	...	x_6^d
...
x_n^1	x_n^2	x_n^3	x_n^4	...	x_n^d

Дефиниција 1. Скуп мултидимензионалних података \mathcal{D} представља скуп од n слогова $\overline{X}_1, \dots, \overline{X}_n$ таквих да сваки од слогова \overline{X}_i садржи скуп од d особина означених са (x_i^1, \dots, x_i^d)

Ретки подаци

x_1^1		x_1^3	x_1^4	...	
x_2^1				...	
	x_3^2			...	x_3^d
			x_4^4	...	
x_5^1		x_5^3		...	
	x_6^2			...	
...
	x_n^2			...	x_n^d

Међусобно зависни подаци

Међусобно зависни подаци - имплицитна или експлицитна зависност између података

Индекс	Име	Презиме	Датум уписа	Датум рођења	Место рођења
20140021	Милош	Перић	06.07.2014	20.01.1995	Београд
20140022	Маријана	Савковић	05.07.2014	11.03.1995	Краљево
20130023	Сања	Терзић	04.07.2013	09.11.1994	Београд
20130024	Никола	Вуковић	04.07.2013	17.09.1994	
20140025	Маријана	Савковић	06.07.2014	04.02.1995	Краљево
20140026	Зорица	Миладиновић	06.07.2014	08.10.1995	Врање
20130027	Милена	Станковић			

Међусобно зависни подаци

Отежано добијање резултата

- Различити формати
 - временске серије, дискретни подаци, просторни подаци, графовски и мрежни подаци, ...
- Имплицитна зависност
 - нпр. сви сензори се понашају слично и ако се у неком тренутку уочи велико одступање оно је од интереса за истраживање
- Експлицитна зависност
 - нпр. граф посета неком веб сајту, утицај лекова на други лек/болест, ...

Подаци везани за временске серије

- Садрже вредности добијене непрекидним мерењем кроз време
- Имплицитна зависност од претходних мерења (нпр. ЕКГ, мерење температуре, ...)

Подаци везани за временске серије

Издавајање зависности од временских атрибута

- Контекстуална зависност - зависност дефинисана преко садржаја
 - нпр. временски тренутак или позиција при добијању податка
- Зависност понашања - вредности које су измерене у одговарајућем контексту
 - може да постоји више оваквих атрибута за једну инстанцу

Подаци везани за временске серије

Дефиниција 2. Временска серија дужине n и димензије d садржи d нумеричких карактеристика за сваку од n временских тачака t_1, t_2, \dots, t_n . Свака временска тачка садржи компоненту за сваку од d серија, тако да су вредности добијене у временском тренутку t_i једнаке $\bar{Y}_i = (y_i^1, \dots, y_i^d)$. Вредност j -те серије у временском тренутку t_i је y_j^i

Дискретне секвенце и ниске

- Могу да се посматрају као категорички еквивалент временских серија
- Контекстуални атрибут је временска тачка или позиција у уређењу
- Атрибут који одређује понашање је категоричког типа

Stavka/Događaj



(A B)	(D)	(C E)
(B D)	(C)	(E)
(C D)	(B)	(A E)



Element u nizu

Дискретне секвенце и ниске

```
GGTTCGCGCCTTCAGCCCCGCGCC
CGCAGGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCGCCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Дефиниција 3. Дискретна секвенца дужине n и димензије d садржи d особина са дискретним вредностима за сваку од n различитих временских тачака t_1, t_2, \dots, t_n . Свака од n компоненти \overline{Y}_i садржи атрибуте са дискретним карактеристикама (понашањем) који су прикупљени у i -тој временској тачки.

Просторни подаци

- Одређују просторне локације
- Атрибути (најчешће два) који одређују простор су контекстуални
- Други атрибути (којих може да буде више) могу да моделирају понашање

Просторни подаци

Дефиниција 4. Слог просторних података са d димензија садржи d атрибута који приказују понашање и један или више контекстуалних података који одређују локацију. Скуп d димензионих просторних података садржи d димензионих слогова $\overline{X}_1, \dots, \overline{X}_n$, и скуп од n локација L_1, \dots, L_n , таквих да је слог \overline{X}_i придружен локацији L_i .

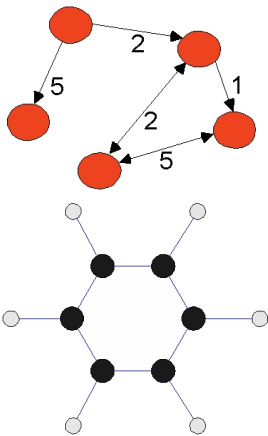
Просторно-временски подаци

Два типа просторно-временских података

- И просторни и временски атрибути могу да буду контекстуални (нпр. мерење варијације температуре мора кроз време)
- Временски атрибут је контекстуалан, а просторни моделира понашање (нпр. анализа трајекторија)

Графовски и мрежни подаци

- Вредности могу да буду придружене чворовима у мрежи
- Однос између вредности података је приказан преко грана
- У неким случајевима атрибути могу да буду додељени чворовима



Графовски и мрежни подаци

Дефиниција 5. Мрежа $G = (N, A)$ садржи скуп од N чворова и скуп грана A , при чему гране из A представљају релације између чворова. Скуп атрибута \overline{X}_i може да буде придружен чвору i , или скуп атрибута \overline{Y}_{ij} може да буде придружен грани (i, j) .

Најзначајнији градивни блокови у ИП

- Подаци
- Правила придруживања
- Класификација
- Кластеровање
- Анализа и визуелизација резултата

Подаци у процесу ИП

- Мултидимензионална база \mathcal{D} са n слогова и d атрибута
- Претходна припрема података (препроцесирање)
 - Редукција
 - Откривање аномалија
 - ...
- Представљање у облику матрице података D димензије $n \times d$
- Релације између колона
- Релације између врста
- Релације између група атрибута у врстама

Аномалије и елементи ван граница

Дефиниција 6. (Откривање аномалија) За дату матрицу података D одредити редове у матрици који су јако различити од остатка редова

- Податак је елемент ван граница (енг. *outlier*) ако је у значајној мери различит од осталих података
- Синоними: аномалија, абнормалност, дискоординација, ...
- Пример: откривање упада у рачунарски систем, идентификација СПАМ порука, злоупотреба кредитних картица, медицинска дијагностика, спровођење закона,

Правила придруживања

- Најједноставнији облик: ретка бинарна база (0/1, са највећим бројем 0)
- Колоне - ставке, врста трансакције:
 $(i,j) = 1 \rightarrow$ трансакција i садржи ставку j

Бр. транс.	хлеб	млеко	пелене	пиво	јаја	кола
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

Правила придруживања

Дефиниција 7. У датој бинарној матрици D величине $n \times d$ посматрају се сви подскупови колона A такви да све вредности у тим колонама у одговарајућој врсти имају вредност 1. Тада важе следеће ознаке:

- A је скуп ставки
- $\#(A)$ означава број појављивања скупа ставки A у комплетном скупу
- N представља број редова у комплетном скупу
- $A \Rightarrow B$ означава да је скуп ставки B придружен скупу ставки A

Правила придруживања

Дефиниција 8. Нека су A и B два скупа ставки. Тада се подршка (енг. *support*) правила придруживања $A \Rightarrow B$ у ознаци *sup* дефинише као количник броја трансакција које садрже A и B у односу на укупан број трансакција

$$\text{sup}(A \Rightarrow B) = \frac{\#(A \cup B)}{N}$$

Дефиниција 9. Нека су A и B два скупа ставки. Тада се поузданост (поверење, енг. *confidence*) правила придруживања $A \Rightarrow B$ у ознаци *conf* дефинише као количник броја трансакција које садрже A и B у односу на број трансакција које садрже A

$$\text{conf}(A \Rightarrow B) = \frac{\#(A \cup B)}{\#(A)}$$

Правила придруживања

- Задатак је одредити правила придруживања (енг. *association rules*) која повезују атрибуте у истој инстанци
- *Интересантна* су правила која имају одређен ниво подршке и поузданости
- За налажење интересантних правила често се не користе апсолутне фреквенције већ χ^2 мера
- Елементи матрице не морају да буду бинарне вредности

Правила придруживања

Бр. транс.	хлеб	млеко	пелене	пиво	јаја	кола
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

$$\#(\{\text{млеко, хлеб, пелене}\})=2$$

за $\{\text{млеко, пелене}\} \Rightarrow \{\text{пиво}\}$

$$\text{sup}(\{\text{млеко, хлеб, пелене}\})=2/5$$

$$\text{sup} = \frac{\#(\{\text{млеко, пелене, пиво}\})}{N} = 2/5$$

$$\text{conf} = \frac{\#(\{\text{млеко, пелене, пиво}\})}{\#(\{\text{млеко, пелене}\})} = 2/3$$

Кластеровање

- Груписање врста по 'сличности'
- Кластеровање се у литератури јавља и као *класификација без надзора*

x_1^1	x_1^2	x_1^3	x_1^4	...	x_1^d
x_2^1	x_2^2	x_2^3	x_2^4	...	x_2^d
x_3^1	x_3^2	x_3^3	x_3^4	...	x_3^d
x_4^1	x_4^2	x_4^3	x_4^4	...	x_4^d
x_5^1	x_5^2	x_5^3	x_5^4	...	x_5^d
x_6^1	x_6^2	x_6^3	x_6^4	...	x_6^d
...
x_n^1	x_n^2	x_n^3	x_n^4	...	x_n^d

Класификација (наставак)

Дефиниција 10. (Класификација података) За дату матрицу података D из базе \mathcal{D} величине $n \times d$ и вредност ознака класа у интервалу $\{1, \dots, k\}$ које су придружене сваком од n слогова (редова) формира се модел за тренирање \mathcal{M} , који може да се користи за предвиђање ознаке класа d -димензионалног слога $\bar{Y} \notin \mathcal{D}$.

- Модел се формира над *подацима за тренинг*
- *Тест подаци* $\notin \mathcal{D}$ (скуп који се користи за формирање модела)

Неки примери ИП апликација

- Распоређивање производа у радњама
- Препоруке купцима
- Аномалије у веб логовима
- ...