Мере интересантности правила

Ненад Митић

Математички факултет
nenad@matf.bg.ac.rs

Мере интересантности правила

- Често се добија јако велики број правила придруживања
- Велики број правила из формираног скупа
  - често није користан за даље истраживање/анализу
  - може да доведе до тога да се превиди неко од потенцијално интересантних правила
  - потребно је дефинисати мере које ће бити коришћене у процесу елиминације неинтересантних правила
- Не постоји мера која је најбоља у свим случајевима
- Меру треба бирати у зависности од претходне анализе података над којима се врши истраживање
- Свака од мера има своје предности и недостатке, као и случајеве у којима погрешно рангира интересантност правила

Табела контингената

За илустрацију мера се обично користи скуп који садржи учесталост појављивања ставки представљен у облику *табеле контингената*. Табела контингената за пар бинарних променљивих A и B има следећи изглед

	B	$\overline{B}$	
A	f_{11}	f_{10}	f_{1+}
$\overline{A}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

f_{10} - број транс. које садрже само A
f_{11} - број транс. које садрже и A и B
f_{01} - број транс. које садрже само B
f_{00} - број транс. које не садрже ни A ни B
f_{1+} - бројач подршке за A
f_{0+} - бројач подршке за $\overline{A}$
f_{+1} - бројач подршке за B
f_{+0} - бројач подршке за $\overline{B}$

A - присутно у трансакцији
$\overline{A}$ - није присутно у трансакцији
f_{ij} - бројач учесталости

Табела контингената

Учесталост појављивања ставки је могуће представити и у случају више променљивих у облику мултидимензионалне табеле контингената. На пример, тродимензионална табела контингената за ставке a, b и c има следећи изглед

c	b	$\overline{b}$	
a	f_{111}	f_{101}	f_{1+1}
$\overline{a}$	f_{011}	f_{001}	f_{0+1}
	f_{+11}	f_{+01}	f_{++1}

$\overline{c}$	b	$\overline{b}$	
a	f_{110}	f_{100}	f_{1+0}
$\overline{a}$	f_{010}	f_{000}	f_{0+0}
	f_{+10}	f_{+00}	f_{++0}

где x означава присутност, $\overline{x}$ одсутност у трансакцији $(x \in \{a, b, c\})$, а f_{ijk} означава број трансакција које садрже одговарајућу комбинацију a,b и c:

f_{000} - број транс. које не садрже ни a, ни b, ни c
f_{001} - број транс. које садрже само c
f_{010} - број транс. које садрже само b
f_{011} - број транс. које садрже b и c али не и a
f_{100} - број транс. које садрже само a
f_{101} - број транс. које садрже a и c али не и b
f_{110} - број транс. које садрже a и b али не и c
f_{111} - број транс. које садрже a и b и c

f_{+00} - број транс. које не садрже b и c
f_{+01} - број транс. које не садрже b али садрже c
f_{+10} - број транс. које не садрже c али садрже b
f_{+11} - број транс. које садрже b и c
f_{0+0} - број транс. које не садрже a и c
f_{0+1} - број транс. које не садрже a али садрже c
f_{1+0} - број транс. које не садрже c али садрже a
f_{1+1} - број транс. које садрже a и c
f_{++0} - број транс. које не садрже c
f_{++1} - број транс. које садрже c

Ограничења мере подршка/поузданост

Као једноставан пример за то да правило које је најбоље рангирано према одређеној мери не мора да буде интересантно може да послужи већ посматран скуп трансакција:

ИдТ	Ставке
1	Хлеб, Млеко
2	Млеко, Пелене, Пиво
3	Хлеб, Млеко, Пелене, Пиво
4	Хлеб, Млеко, Пиво
5	Хлеб, Млеко, Пелене, Пиво

Како свака трансакција у потрошачкој корпи садржи *Млеко*, правило

$$X \Rightarrow \text{Млеко, где } X \subseteq \{\text{Хлеб, Пелене, Пиво}\}$$

је бескорисно без обзира што има поузданости 100%.

Ограничења мере подршка/поузданост

Као други пример где комбинација подршка/поузданост не даје увек коректну слику може да послужи следећа табела у којој су приказана анкета да ли анкетирана особа пије чај и/или кафу.

	Кафа	$\overline{\text{Кафа}}$	
Чај	15	5	20
$\overline{\text{Чај}}$	75	5	80
	90	10	100

На први поглед важи правило Чај $\Rightarrow$ Кафа - јер је подршка правила 15% и поузданост 75%

Међутим, како је sup(Кафа $\Rightarrow$ $\overline{\text{Чај}}$)=0.75 и sup(Кафа)=0.9 то је

conf(Кафа $\Rightarrow$ $\overline{\text{Чај}}$)= 0.75/0.9=0.83.3%, што значи да је прво правило бескорисно односно погрешно јер у суштини информација да особа пије чај смањује вероватноћу да пије и кафу.

Ограничења мере подршка/поузданост

1. Табела контингената садржи информације о коришћењу меда од стране особа које пију чај

	Мед	$\overline{\text{Мед}}$	
Чај	100	100	200
$\overline{\text{Чај}}$	20	780	800
	120	880	1000

   На први поглед правило Чај $\longrightarrow$ Мед има поузданост 50% из чега може да се изведе закључак да особина да се пије чај не утиче на коришћење меда

   Међутим, проценат оних који користе мед је 12% $\Longrightarrow$ информација да неко пије чај повећава вероватноћу да користи мед са 12% на 50%!

   $\Longrightarrow$Правило Чај $\longrightarrow$ Мед јесте од интереса!

2. Проблем: поузданост не узима у обзир подршку десне стране правила

Статистичка перспектива

Неке чињенице које се користе у наредним примерима

1. Подршка $sup(A)$ мери вероватноћу појављивања A: $sup(A) = \frac{f_{1+}}{N}$
2. Подршка $sup(A, B)$ мери вероватноћу да се A и B заједно појављују: $P(A, B) = sup(A, B) = \frac{f_{11}}{N}$
3. Ако су A и B независни $\rightarrow P(A, B) = P(A) \times P(B)$, односно $sup_{nez}(A, B) = sup(A) \times sup(B) = \frac{f_{1+}}{N} \times \frac{f_{+1}}{N}$
4. Одступање $sup(A, B)$ од $sup(A) \times sub(B)$ је знак статистичке зависности A и B
5. Поузданост мери одступање $sup(A, B)$ од $sup(A)$ али не и од $sup(B)$

Лифт

Једна од мера која коригује недостатак поузданости и која узима у обзир и подршку десног дела правила $A \longrightarrow B$) је *Лифт*.

$$Lift(A, B) = I(A, B) = \frac{conf(A \longrightarrow B)}{sup(B)}$$

Лифт вредности веће од 1 означавају да је последична (десна) страна правила много чешћа у трансакцијама које садрже леву (узрочну) страну правила него у трансакцијама које је не садрже, док вредност мања од 1 означавају правила чија је позданост мања од очекиване. У пракси, ако се користи Лифт као мера интересантности, пожељно је користити вредности Лифт>1.1

Piatetsky-Shapiro мера

- Ростоје случајеви када Лифт не пружа адекватну информацију. На пример
  - правило које има мању подршку и већи Лифт може у неким случајевима да буде мање интересантно од правила које има већу подршку и мањи Лифт пошто је такво правило применљиво на већи део материјала
  - у случају потрошачке корпе на већи број потрошача који купују неке артикле добит од куповине неког артикла од стране већег броја потрошача би у том случају била задовољавајућа

- Мера која спаја величину и јачину ефекта правила придруживања је *Piatetsky-Shapiro*. У литератури се ова мера још назива и мера *полуге* (енг. *leverage*) или моћ правила

$$PS = s(A,B) - s(A) \times s(B) = \frac{f_{11}}{N} - \frac{f_{1+}f_{+1}}{N^2}$$

- $PS(A,B) = 0 \longrightarrow A$ и B су међусобно независни
- $PS(A,B) > 0 \longrightarrow A$ и B су позитивно корелисани
- $PS(A,B) < 0 \longrightarrow A$ и B су негативно корелисани

Однос камата

Однос камата (енг. *interest ratio*, *interest factor*) се за скуп ставки $i_1, i_2, \ldots, i_k$ дефинише као

$$I(i_1, i_2, \ldots, i_k) = \frac{sup(i_1, i_2, \ldots, i_k)}{\prod_{j=1}^{k} sup(i_j)}$$

У случају бинарних променљивих A и B однос камата је једнак

$$I(A, B) = \frac{sup(A, B)}{sup(A) \cdot sup(b)} = \frac{N \cdot f_{11}}{f_{1+} \cdot f_{+1}}$$

Важи

$$I(A, B) = \begin{cases} = 1, & \text{ако су А и В независни} \\ > 1, & \text{ако су А и Б позитивно корелисани} \\ < 1, & \text{ако су А и Б негативно корелисани} \end{cases}$$

У случају када је нека ставка екстремно ретка, однос камате даје нетачне резултате. На пример, ако се ставка X јавља само у једној трансакцији у великој бази података, свака ставка Y која се јавља заједно са њом у тој трансакцији може да се упари и формира скуп ставки $\{X, Y\}$ са јако великим односом камате.

У случају бинарних променљивих, однос камате се поклапа са лифт мером.

Deployability

Deployability - могућност ширења (развијања, распоређивања) - проценат скупа података који задовољава услов на левој страни правила али не и последични део на десној страни. У контексту потрошачке корпе ова мера би означавала проценат броја купаца који су купили артикле наведене на левој страни правила, али (још увек) нису купили оне наведене на десној страни. Могућност ширења се дефинише као

$$deployability = \frac{(Antecedent\ \ Support\ \ in\ \#\ \ of\ \ Records) - (Rule\ \ Support\ \ in\ \#\ \ of\ \ Records)}{Number\ \ of\ \ Records} * 100$$

где *Antecedent Support* означава број слогова у којима се јавља лева страна правила, док *Rule Support* означава број слогова у којима се јављају и лева и десна страна правила

Коефицијент корелације

Као једна од мера може се користити и Пирсонов коефицијент корелације

$$\rho = \frac{E[X \cdot Y] - [X] \cdot [Y]}{\sigma(X) \cdot \sigma(Y)}$$

где је $E[X]$ очекивање од X, а $\sigma(X)$ стандардна девијација од X

$$\rho_{ij} = \frac{sup(i,j) - sup(i) \cdot sup(j)}{\sqrt{sup(i) \cdot sup(j) \cdot (1 - sup(i)) \cdot (1 - sup(j))}}$$

где су $sup(i)$, $sup(j)$ и $sup(i,j)$ релативне подршке ставки i, j и скупа ставки {i,j}

χ^2 мера

Једна од симетричних мера која третира присуство и одсуство ставки на идентичан начин је χ^2 мера

Нека су O_i и E_i осмотрена и очекивана вредност апсолутне подршке ставки у стању i (присутан, одсутан). Тада се χ^2 мера скупа ставки X дефинише као

$$\chi^2(X) = \sum_{i=1}^{2^{|X|}} \frac{(O_i - E_i)^2}{E_i}$$

Вредности које су близу 0 означавају статистичку независност између ставки. Веће вредности означавају већу зависност између ставки, али не носе информацију да ли је та зависност позитивна или негативна.

χ^2 тест задовољава особину затворења навише, што као последицу има могућност конструкције ефикасног алгоритма за одређивање интересантних к-ставки.

ИС мера

ИС мера је алтернативна мера која може да се примени у случају асиметричних бинарних променљивих. Укључује однос између $sup(A,B)$ и $sup(A)$ и $sup(B)$:

$$IS(A,B) = \sqrt{I(A,B) \times sup(A,B)} = \frac{sup(A,B)}{\sqrt{sup(A) \times sup(B)}}$$

- ИС расте када расту И (однос камате) и подршка
- Ако два обрасца имају исти однос камате ИС даје предност ономе са већом подршком
- ИС је еквивалентан косинусној мери за бинарне променљиве

Коефицијент корелације

За бинарне променљиве, Пирсонов коефицијент корелације може да се мери користећи ρ коефицијент

$$\rho = \frac{f_{11} \cdot f_{00} - f_{01} \cdot f_{10}}{\sqrt{f_{1+} \cdot f_{+1} \cdot f_{0+} \cdot f_{+0}}}$$

	p	$\overline{p}$	
q	880	50	930
$\overline{q}$	50	20	70
	930	70	1000

	r	$\overline{r}$	
s	20	50	70
$\overline{s}$	50	880	930
	70	930	1000

Ограничења

- Иако се p и q заједно јављају чешће него r и s важи $\rho(p,q) = \rho(r,s) = 0.232$ јер даје једнаку важност присуству и одсуству ставки у трансакцијама
- Однос камата: $I(p,q) = 1.02, I(r,s) = 4.08$, док $IS(p,q) = 0.9346, IS(r,s) = 0.286$ Конфликт?

Косинусна мера колоне

- Косинусна мера може да се примени и на колоне ради рачунања сличности међу ставкама у тим колонама
- Најчешће се рачуна користећи вертикалну *ИдТ* репрезентацију листи одговарајућих бинарних вектора
- Ако су A и B пар вектора тада је $A \bullet B = sup(A, B)$, а $\mid A \mid = \sqrt{sup(A)}$ је величина вектора A, па се добија да је

$$IS(A, B) = \frac{sup(A, B)}{\sqrt{sup(A) \times sup(B)}} = \frac{A \bullet B}{\mid A \mid \times \mid B \mid} = cos(A, B)$$

- Симетрична мера

Поузданост свих

Поузданост свих је мера која се односи на скупове ставки (не на правила!). Подржава затворење наниже и рачуна се по формули

$$all - confidence(X) = \frac{sup(X)}{max_{x \in X}(sup(x))}$$

где је $max_{x \in X}(sup(x))$ величина подршке ставке са највећом подршком у скупу трансакција X

Значење: сва правила која могу да се изведу из X имају подршку бар једнаку $all - confidence(X)$

Особине мера

- Поред претходно наведених у литератури се јавља велики број мера.
- Неке мере су добре за неке примене, али не и за неке друге
- Који критеријум треба користити при процени квалитета мере?

Особине мера

Према ауторима Piatetsky-Shapiro добра мера мора да задовољава 3 особине:

- $M(A,B) = 0$ ако су A и B статистички независне
- $M(A,B)$ се монотоно повећава са $P(A,B)$ када $P(A)$ и $P(B)$ остају непромењене
- $M(A,B)$ се монотоно смањује са $P(A)$ [или $P(B)$] када $P(A,B)$ и $P(B)$ [или $P(A)$] остају непромењене

Особине мера

Додатно питање је како се мера M понаша у случају

- пермутације променљивих $M(A,B) = M(B,A)$? Ако важи да је $M(A,B) = M(B,A)$ таква мера је симетрична
- скалирања вредности у реду или колони
- инверзије (нпр. код вектора бинарних вредности прелазак 0 у 1 и обратно)
- додавања "празних" слогова (слогова који не садрже A и B, тј. када се повећава само f_{00}

На наредна два слајда су приказане неке од мера и њихово рангирање према табели контингената са 10 случајева. Из примера се види да не постоји мера која је рангирана као најбоља у свим случајевима.

Особине мера

#	Measure	Formula
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i), -\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\Big(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\Big)$
9	Gini index (G)	$\max\Big(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]$ $-P(A)^2-P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A), P(A\|B))$
12	Laplace (L)	$\max\left(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\right)$
13	Conviction (V)	$\max\left(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})}\right)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B)-P(A)P(B)$
17	Certainty factor (F)	$\max\left(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)}\right)$
18	Added Value (AV)	$\max(P(B\|A)-P(B), P(A\|B)-P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B), P(A\|B)-P(A))$

Особине мера

Мере приказане на претходном слајду се рангирају према табли контингената у различите редоследе (1-најбоље; 10- најгоре)

Example	f_{11}	f_{10}	f_{01}	f_{00}
E1	8123	83	424	1370
E2	8330	2	622	1046
E3	9481	94	127	298
E4	3954	3080	5	2961
E5	2886	1363	1320	4431
E6	1500	2000	500	6000
E7	4000	2000	1000	3000
E8	4000	2000	2000	2000
E9	1720	7121	5	1154
E10	61	2483	4	7452

#	ϕ	λ	α	Q	Y	κ	M	J	G	s	c	L	V	I	IS	PS	F	AV	S	ζ	K
E1	1	1	3	3	3	1	2	2	1	3	5	5	4	6	2	2	4	6	1	2	5
E2	2	2	1	1	1	2	1	3	2	2	1	1	1	8	3	5	1	8	2	3	6
E3	3	3	4	4	4	3	3	8	7	1	4	4	6	10	1	8	6	10	3	1	10
E4	4	7	2	2	2	5	4	1	3	6	2	2	2	4	4	1	2	3	4	5	1
E5	5	4	8	8	8	4	7	5	4	7	9	9	9	3	6	3	9	4	5	6	3
E6	6	6	7	7	7	7	6	4	6	9	8	8	7	2	8	6	7	2	7	8	2
E7	7	5	9	9	9	6	8	6	5	4	7	7	8	5	5	4	8	5	6	4	4
E8	8	9	10	10	10	8	10	10	8	4	10	10	10	9	7	7	10	9	8	7	9
E9	9	9	5	5	5	9	9	7	9	8	3	3	3	7	9	9	3	7	9	9	8
E10	10	8	6	6	6	10	5	9	10	10	6	6	5	1	10	10	5	1	10	10	7

Особине симетричних мера

У наредној табели су приказане особине неких симетричних мера

Ознака	Мера	Инверзија	Додавање празних слогова	Скалирање
ϕ	ϕ-коефицијент	да	не	не
α	количник шанси (odds ration)	да	не	да
κ	Cohen	да	не	не
Lift	Лифт (I, интересантност)	не	не	не
IS	Косинус	не	да	не
PS	Piatetsky-Shapiro	да	не	не
S	Јачина групе	да	не	не
ζ	Žakard	не	да	не
h	поузданост свих	не	да	не
s	Подршка	не	не	не

Симпсонов парадокс

Скривене променљиве које не учествују у анализи могу да утичу на резултат

- скуп трансакција које приказују куповину HDTV и трака за трчање

Купили HDTV	Купили траку за трчање		Збир
	Да	Не	
Да	99	81	180
Не	54	66	120
	153	147	300

- Правило {HDTV=Да}$\longrightarrow$ {трака за трчање=Да} има поузданост 55% (99/180)
- Правило {HDTV=Не}$\longrightarrow$ {трака за трчање=Да} има поузданост 45% (54/120)
- Да ли је коректан закључак да ће купац који купи HDTV врло вероватно купити и траку за трчање?!

Симпсонов парадокс

Дубља анализа: у скупу купаца постоје две групе: студенти и запослени. Ако се изврши анализа куповине по тим групама

Купац	Купили HDTV	Купили траку за трчање		Збир
		Да	Не	
Студент	Да	1	9	10
	Не	4	30	34
Запослени	Да	98	72	170
	Не	50	36	86

- Куповина HDTV и трака за трчање за студенте
  - Поузданост правила conf{HDTV=Да}$\longrightarrow$ {трака за трчање=Да}=10% (1/10)
  - Поузданост правила conf{HDTV=Не}$\longrightarrow$ {трака за трчање=Да}=11.8% (4/34)
- Куповина HDTV и трака за трчање за запослене
  - Поузданост правила conf{HDTV=Да}$\longrightarrow$ {трака за трчање=Да}=57.7% (98/170)
  - Поузданост правила conf{HDTV=Не}$\longrightarrow$ {трака за трчање=Да}=58.1% (50/86)
- За обе појединачне групе важи правило да ће купац који не купи HDTV врло вероватно купити и траку за трчање?!

Симпсонов парадокс

- Без обзира на алтернативну меру (корелација, лифт, ...) HDTV и траке за трчање су
  - позитивно повезане када су подаци комбиновани
  - негативно повезане када су подаци стратификовани
- Овакав случај обртања правила се назива Симпсонов парадокс!
- Домаћи задатак: Дати објашњење