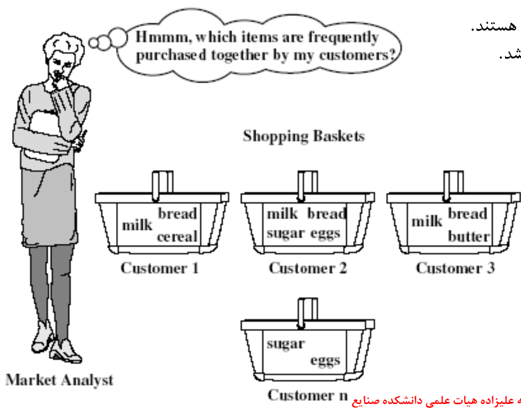
قواعد تلازمی

مفاهیم و آلگوریتمها

قواعد تلازمی

- تحليل وابستگيها يك حالت غير نظارتي داده كاوي مي باشد كه به جستجو براي يافتن ارتباط در مجموعه داده ها مي پردازد. به عبارتي ديگر تحليل وابستگيها مطالعه ويژگي‌ها يا خصوصياتي مي‌باشد كه با يكديگر همراه هستند. این روش در سال 1993 توسط **Agrawal** مطرح شد.



- این الگوریتم در وهله اول در تحلیل سبد خرید و اینکه چقدر اقلام خریداري شده بوسیله مشتریان با هم ارتباط دارند ، مورد استفاده قرار گرفت .

تعاریف و مفاهیم اصلی

قوانين وابستگي به شكل اگر و آنگاه به همراه معيارهاي پشتيبان و اطمينان مربوط به قوانين مي‌باشند.

$I = \{ i1 , i2 , \ldots , im \}$: مجموعه‌ای از کل ایتمهای خریداری شده است

T: هر زیرمجموعه‌ای از **I** می‌باشد که از آن بعنوان تراکنش یاد می‌کنیم.

D: مجموعه تراکنشهای موجود در T است

TID : شناسه منحصر به فرد و یکتایی است که به هریک از تراکنشهااختصاص می‌یابد.

نمای کلی یک قانون وابستگی به فرم زیر می‌باشد:

$$X \Rightarrow Y \text{ [support , Confidence]}$$

می‌باشد بطوریکه داریم:

$X \subset I, Y \subset I$ و $X \cap Y = \varnothing$

تعاریف و مفاهیم اصلی

- **پشتیبان یا Support**: *نشان دهنده درصد یا تعداد مجموعه تراکنشهایی در* ***D*** *است که شامل هر دوی* ***X*** *و* ***Y*** ***(X ∪ Y)*** *باشند.*
- **اطمینان یا Confidence**: میزان وابستگي يك قلم كالاي خاص را به ديگري بيان مي كند ومطابق فرمول زیر محاسبه می‌شود:

$$confidence = support(X \cup Y)/support(X)$$

مثال

مجموعه ای از کل اقلام خریداری شده در **I** آمده است.

I={خیار ، جعفری ، پیاز ، گوجه‌فرنگی ، نمک ، نان ، زیتون ، پنیر ،کره}

- مجموعه D که شامل تک تک تراکنشها و خریدها میباشد ، به فرم زیر است:

- *D={T1,T2,T3,T4,T5,T6,T7,T8}*

- ***T1***{*جعفری ، پیاز ، زیتون ، خیار ،گوجه‌فرنگی*}،
- ***T2***{*جعفری ، خیار ،گوجه‌فرنگی*}،
- *T3*{*نان ، نمک ،گوجه‌فرنگی ، پیاز ، جعفری ، خیار*}،
- ***T4***{*نان ،پیاز ،خیار ،گوجه فرنگی*}
- ***T5***{*پیاز ،نمک ، گوجه‌فرنگی*}،
- ***T6***{*پنیر ،نان*}،
- ***T7***{*خیار ،پنیر ،گوجه‌فرنگی*}،
- ***T8***{*کره ،نان*}}

مثال

فرض کنیم قانون وابستگی به فرم زیر را داریم:

X => Y [support , Confidence]

{خیار, گوجه‌فرنگی} ⟹ {پیاز, جعفری}

X = {گوجه‌فرنگی, خیار}

Y = {جعفری, پیاز}

X=>Y = {گوجه‌فرنگی ، خیار ، جعفری، پیاز} = {T1, T3}

Support(***X=>Y)*** = 2/8 = 0.25

مثال

زیرا مجموعه $X \cup Y$، 2 عضو و مجموعه D، 8 عضو دارد و بدین معنی است که الگوی خرید "گوجه‌فرنگی ، خیار ، جعفری ،پیاز" در 25% کل سبد خرید رخ می‌دهد.

X={ T3,T4,T7,T2,T1}

یعنی {خیار ,گوجه‌فرنگی} در T3,T4,T7,T2,T1 خریداری شده‌اند. بنابراین:

Support(x)=8/5=0.62

Confidence = *support(X => Y)/support(X)*

= (2/8)/(5/8)=2/5=0.40

یعنی هنگامی که افراد "خیار و گوجه‌فرنگی" را خریداری می‌کنند ، در 40% اوقات ، "جعفری و پیاز" را هم می‌خرند .

الگوریتم Apriori

L_1 = {large 1-itemsets}

For ($k = 2; L_{k-1} \neq \phi ; k++$) do begin

C_k = apriori-gen (L_{k-1});

forall transactions $t \in D$ do begin

C_t = subset(C_k,t)

forall candidates $c \in C_t$ do

$c.count++$;

end

end

$L_k = \{ c \in C_k | c.count \geq minsup \}$

end

$Answer = \bigcup_k L_k$;

بزرگترین آیتمها را در نظر می‌گیرد

CKهای جدید را مطابق الگوریتم بیان شده تولید می کند

Support هر یک از آیتمهای CK را محاسبه می‌کند

فقط آنهایی که از حداقل Support بزرگتر یا مساویند را در LK قرار می‌دهد

الگوریتم Apriori مثال:

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

L1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

C2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

L2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

itemset
{2 3 5}

itemset	sup
{2 3 5}	2

الگوریتم Apriori

مطابق الگوریتم برای محاسبه C3 بصورت زیر عمل می‌کنیم:

L_2 = { {1,3}, {2,3}, {2,5}, {3,4} }

برای محاسبه C3 مطابق این الگوریتم فقط آنهایی که مولفه اول برابر دارند یعنی {{2,3},{2,5}} هستند و از آنجائیکه برای محاسبه تمامی Ck ها باید قانون لکسیکوگرافیک رعایت شود تنها حالت ممکن {2,3,5} می‌باشد. بنابراین

C3={2,3,5}

استخراج قوانین

- پروسه استخراج قوانین وابستگی
- ابتدا همه frequent itemsets ها را که دارای حداقل support هستند بیابید.
- برای تمامی frequent itemset ها
  - همه زیرمجموعه‌های آنها را استخراج کنید
  - همه قوانین ممکن را استخراج کنید
  - قوانینی را بپذیرید که از حداقل confidence برخوردارهستند.

- آیتمها در هر مجموعه‌ای مطابق با قانون لکسیکوگرافیک چیده می‌شوند به عنوان مثال
- اگر LK ={a[1],a[2],…a[k]} باشد ، مطابق این قانون باید رابطه زیر برقرار باشد.
  a[1]<a[2]<…<a[k]

مثال دوم

- minimum support s=30%,
  minimun confidence c= 60%

Transaction ID	Items Purchased
1	{آب پرتقال, لیموناد}
2	{شیر, آب پرتقال, شیشه پاک کن}
3	{آب پرتقال, پاک کننده, لیموناد}
4	{شیشه پاک کن, لیموناد}
5	{لیموناد, چیپس}

Candidate 1-itemsets (C1)

1-itemset	Support
آب پرتقال	60%
لیموناد	80%
شیر	20%
شیشه پاک کن	40%
پاک کننده	20%
چیپس	20%

Large 1-itemsets (L1)

1-itemset	Support
آب پرتقال	60%
لیموناد	80%
شیشه پاک کن	40%

Candidate 2-itemsets (C2)

2-itemset	Support
{آب پرتقال, لیموناد}	40%
{آب پرتقال, شیشه پاک کن}	20%
{لیموناد, شیشه پاک کن}	20%

Large 2-itemsets (L2)

2-itemset	Support
{آب پرتقال, لیموناد}	40%

- آب پرتقال → لیموناد (confidence = 66.67%)
- لیموناد → آب پرتقال (confidence = 50%)

- آب پرتقال → لیموناد (confidence = 66.67%)

مثال سوم

- minimum support count is 2, (s=2/9=22%)

TID	List of Item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2, I3,I5
T900	I1,I2,I3

Candidate 1-itemsets (C1)

1-itemset	Support count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Large 1-itemsets (L1)

1-itemset	Support count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Candidate 2-itemsets (C2)

2-itemset	Support count
{I1,I2}	4
{I1, I3}	4
{I1,I4}	1
{I1,I5}	2
{I2,I3}	4
{I2,I4}	2
{I2,I5}	2
{I3,I4}	0
{I3,I5}	1
{I4,I5}	0

Large 2-itemsets (L2)

2-itemset	Support count
{I1,I2}	4
{I1, I3}	4
{I1,I5}	2
{I2,I3}	4
{I2,I4}	2
{I2,I5}	2

Candidate 3-itemsets (C3)

3-itemset	Support count
{I1,I2,I3}	2
{I1, I2,I5}	2

Large 3-itemsets (L3)

3-itemset	Support count
{I1,I2,I3}	2
{I1, I2,I5}	2

الگوریتم AprioriTid

- همانگونه که قبلاً نیزذکرشد الگوریتم Apriori همه پایگاه‌داده را درهر گذر می‌پیماید تا Support ها را محاسبه کند و پیمودن همه پایگاه داده ممکن است در همه فازها مورد نیاز نباشد . بر مبنای این مشکل ، الگوریتم دیگری بنام AprioriTid ابداع شد. این الگوریتم نیزروشی مشابه با الگوریتمApriori را برای محاسبه CKها درهر فاز بکار می‌برد.

- تفاوت عمده ای که این الگوریتم با الگوریتم Apriori دارد در این است که این الگوریتم کل پایگاه داده را برای محاسبه Support بعد از مرحله اول نمی‌پیماید و از مجموعه C^K برای محاسبه Support استفاده می کند.

الگوریتم AprioriTid

$L_1 = \{large\ 1\text{-}itemsets\}$

$\hat{C}_1 = database\ D;$

For $(k = 2;\ L_{k-1} \neq \phi;\ k{+}{+})$ do begin

  $C_k = \text{apriori-gen}\,(L_{k-1});$

  $\hat{C}_k = \phi;$

  forall *entries* $t \in \hat{C}_{k-1}$ do begin

    $C_t = \{c \in C_k \mid (c - c[k]) \in t.set\text{-}of\text{-}items$

      $\wedge\, (c - c[k-1]) \in t.set\text{-}of\text{-}items\};$

    forall candidates $c \in C_t$ do

      $c.count{+}{+};$

      if $(C_t \neq \varphi)$ then $\hat{C}_k {+}{=} < t.TID, C_t >;$

    end

  end

  $L_k = \{\, c \in C_k \mid c.count \geq minsup \}$

end

$Answer = \bigcup_k L_k;$

آیتمهای با بیشترین فراوانی را محاسبه می کند

مقداردهی اولیه $C\hat{}_1$ با پایگاه داده موجود

C_Kهای جدید را مطابق الگوریتم بیان شده تولید می کند

یک انباره جدید ایجاد می‌کند

itemsetsهایی را در نظر بگیرید که قبلا وجود داشته اند

Supportرا محاسبه می‌کند

موجودیتهای خالی را حذف می‌کند

فقط آنهایی که ازحداقل Support بزرگتر یا مساویند را در L_K قرار می‌دهد

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

L_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

$C\hat{}_1$

TID	Set-of-itemsets
100	{ {1},{3},{4} }
200	{ {2},{3},{5} }
300	{ {1},{2},{3},{5} }
400	{ {2},{5} }

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

$C\hat{}_2$

TID	Set-of-itemsets
100	{ {1 3} }
200	{ {2 3},{2 5} {3 5} }
300	{ {1 2},{1 3},{1 5}, {2 3}, {2 5}, {3 5} }
400	{ {2 5} }

L_2

Itemset	Support
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

itemset
{2 3 5}

$C\hat{}_3$

TID	Set-of-itemsets
200	{ {2 3 5} }
300	{ {2 3 5} }

L_3

Itemset	Support
{2 3 5}	2

22

سمیه علیزاده هیات علمی دانشکده صنایع

مقایسه Apriori و AprioriTid

- همانگونه که قبلاً نیزذکرشد الگوریتم Apriori همه پایگاه داده را درهر گذر می‌پیماید تا Support ها را محاسبه کند و پیمودن همه پایگاه داده ممکن است در همه فازها مورد نیاز نباشد.
- تفاوت عمده ای که این الگوریتم با الگوریتم Apriori دارد در این است که این الگوریتم کل پایگاه داده را برای محاسبه Support بعد از مرحله اول نمی‌پیماید و از مجموعه C^K برای محاسبه Support استفاده می کند.

مقایسه Apriori و AprioriTid

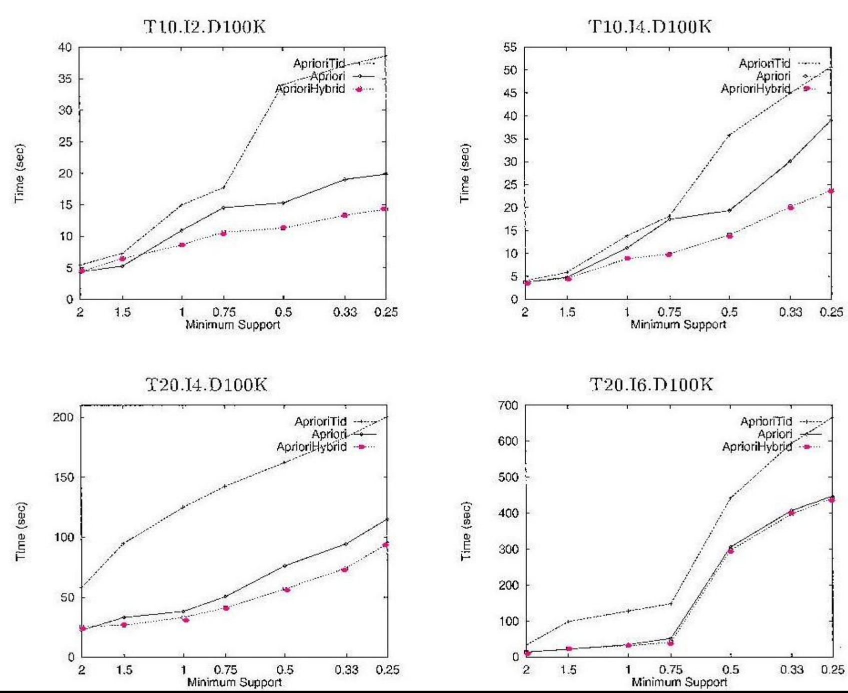
- درالگوریتم AprioriTid مقادیر C^k بجای پایگاه داده در نظر گرفته می‌شوند . اگر C^k بتواند در حافظه Fit شود ، این الگوریتم سریعتر از Apriori عمل خواهد کرد.
- زمانیکه C^k خیلی بزرگ باشد نمی‌تواند در حافظه جای بگیرد و در نتیجه زمان محاسبه بسیار بالا می رود ،بنابراین الگوریتم Apriori سریعتر از الگوریتم AprioriTid عمل خواهد کرد.

الگوریتم Apriori Hybrid

خصوصیات این الگوریتم به ترتیب زیر هستند:

- این الگوریتم درفازهای اولیه اجرا مطابق الگوریتم Apriori عمل می‌کند.
- اندازه تخمینی C^k را بصورت زیر در نظر می‌گیرد:
- وقتی که انتظار می‌رود C^k مناسب حافظه شود به الگوریتم AprioriTid سوئیچ کرده و مطابق این الگوریتم پیش می‌رود.
- اگرچه تغییر از Apriori به AprioriTid زمان براست ، اما در بسیاری ازموارد نتایج مثبتی دارد.

25



دیگر الگوریتمها

- این الگوریتم ازاولین الگوریتمهایی بود که برای استخراج همه آیتمهای بزرگ ازپایگاه داده درسال 1993توسط R.Agrawal, T. Imielinski, and A. Swami ابداع گردید و درحقیت نام این الگوریتم برگرفته از نامهای اول ابداع کنندگان آن می‌باشد.
- این الگوریتم چندین گذررا برروی پایگاه داده انجام می‌دهد و درهرگذرهمه تراکنشها را می پیماید. گامهای این الگوریتم به صورت زیر می‌باشند:
- برای هر یک ازتراکنشها بزرگترین آیتمها انتخاب می‌شوند .
- آیتمهای کاندیدا(CK) با گسترش هر یک از این آیتمهای بزرگ به سایر آیتمها در هر تراکنش ساخته می شوند.

الگوریتم AIS

Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

L_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

C_2

	Itemset	Support
TID_1	{1 3}*	2
	{1 4}	1
	{3 4}	1
TID_2	{2 3}*	2
	{2 5}*	3
	{3 5}*	2
TID_3	{1 2}	1
	{1 5}	1

C_3

Itemset	Support
{1 3 4}	1
{2 3 5}*	2
{1 3 5}	1

الگوریتم AIS

- از معایب این روش اینستکه در هر گذر تعدادی مجموعه آیتم انتخاب می شوند که از حداقل مقدار پشتیبان(در اینجا 2) برخوردار نیستند و بایستی کنار گذاشته شوند.
- بعنوان مثال در C2 مجموعه آیتمهای اضافی عبارتند از {1,4},{3,4},{1,2},{1,5}

الگوریتم SETM

این الگوریتم توسط Houtsma در سال 1995 ابداع شد و در سال 1996 نسخه دوم آن بمنظور محاسبه Itemsetهای بزرگ در SQL توسط Srikant مطرح شد

مشابه الگوریتم AIS ، این الگوریتم نیز چندین گذربرروی پایگاه داده انجام می‌دهد. گامهای این الگوریتم به قرار زیر می‌باشند:

- پشتیبان هر یک از آیتم‌های مجزا محاسبه و بزرگترین آنها انتخاب می‌شوند

آیتمهای کاندیدا(CK) با گسترش هر یک از این آیتمهای بزرگ به سایر آیتمها در هر تراکنش ساخته می شوند. علاوه بر آن در این مرحله TID های مربوط به هر یک از CKرا در یک ساختار ترتیبی نگهداری می کند و سپس support هر یک از CKها با جمع‌کردن این ساختار ترتیبی محاسبه می کند.

الگوریتم SETM

Database

TID	Item
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

L_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	{3}

C_2

Itemset	TID
{1 3}	100
{1 4}	100
{3 4}	100
{2 3}	200
{2 5}	200
{3 5}	200
{1 2}	300
{1 3}	300
{1 5}	300
{2 3}	300
{2 5}	300
{3 5}	300
{2 5}	400

C_3

TID

Itemset	TID
{1 3 4}	100
{2 3 5}	200
{1 3 5}	300
{2 3 5}	300

معایب الگوریتمهای SETM وAIS

- این الگوریتمها خیلی کند هستند
- Itemset های زیادی با Support/Confidence پائینتر از مینیمم Support/Confidence در نظرگرفته شده توسط کاربر تولید می کنند.

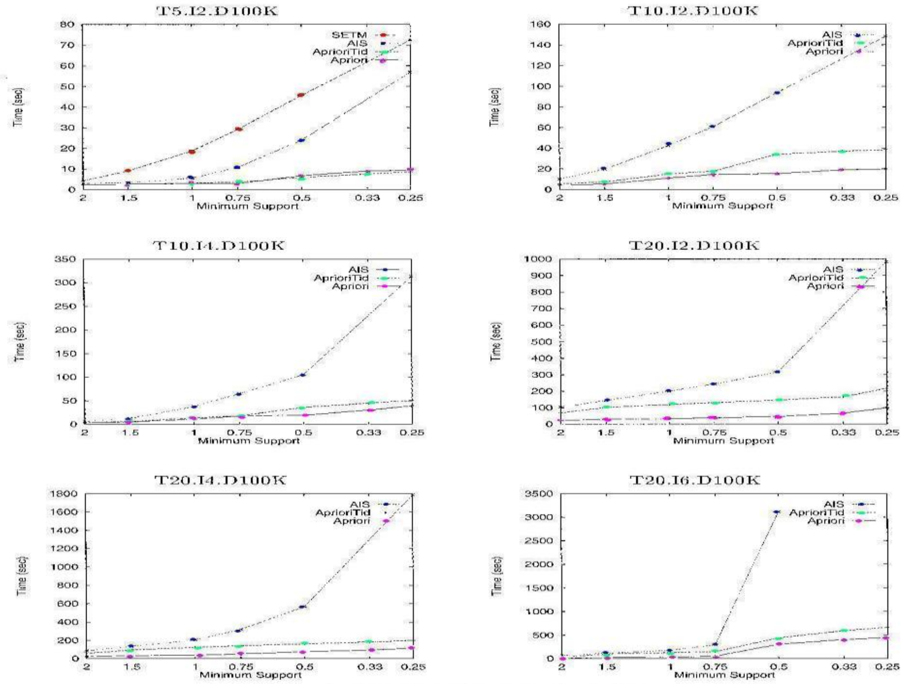


Figure 6: Execution times: Synthetic Data
20

Statistical-based Measures

- Measures that take into account statistical dependence

$$Lift = \frac{P(Y \mid X)}{P(Y)}$$

$$Interest = \frac{P(X,Y)}{P(X)P(Y)}$$

$$PS = P(X,Y) - P(X)P(Y)$$

$$\phi - coefficient = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1-P(X)]P(Y)[1-P(Y)]}}$$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Example: Lift/Interest

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea $\rightarrow$ Coffee

Confidence= P(Coffee|Tea) = 0.75

but P(Coffee) = 0.9

$\Rightarrow$ Lift = 0.75/0.9= 0.8333 (< 1, therefore is negatively associated)

Drawback of Lift & Interest

	Y	$\overline{Y}$	
X	10	0	10
$\overline{X}$	0	90	90
	10	90	100

	Y	$\overline{Y}$	
X	90	0	90
$\overline{X}$	0	10	10
	90	10	100

$$Lift = \frac{0.1}{(0.1)(0.1)} = 10$$

$$Lift = \frac{0.9}{(0.9)(0.9)} = 1.11$$

Statistical independence:

If P(X,Y)=P(X)P(Y) => Lift = 1

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Which Measures Should Be Used?

- **lift** *and* χ^2 are not good measures for correlations in large transactional DBs
- **all-conf** or **coherence** could be good measures (Omiecinski@TKDE'03)
- Both **all-conf** and **coherence** have the downward closure property
- Efficient algorithms can be derived for mining (Lee et al. @ICDM'03sub)

symbol	measure	range	formula
ϕ	ϕ-coefficient	-1…1	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
Q	Yule's Q	-1 … 1	$\frac{P(A,B)P(\overline{A},\overline{B})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{A},\overline{B})+P(A,\overline{B})P(\overline{A},B)}$
Y	Yule's Y	-1 … 1	$\frac{\sqrt{P(A,B)P(\overline{A},\overline{B})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{A},\overline{B})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}}$
k	Cohen's	-1 … 1	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
PS	Piatetsky-Shapiro's	-0.25 … 0.25	$P(A,B)-P(A)P(B)$
F	Certainty factor	-1 … 1	$\max(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)})$
AV	added value	-0.5 … 1	$\max(P(B\|A)-P(B), P(A\|B)-P(A))$
K	Klosgen's Q	-0.33 … 0.38	$\sqrt{P(A,B)}\max(P(B\|A)-P(B), P(A\|B)-P(A))$
g	Goodman-kruskal's	0 … 1	$\frac{\Sigma_j \max_k P(A_j,B_k)+\Sigma_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-max_k P(B_k)}$
M	Mutual Information	0 … 1	$\frac{\Sigma_i\Sigma_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_J)}}{\min(-\Sigma_i P(A_i)\log P(A_i)\log P(A_i),-\Sigma_i P(B_i)\log P(B_i)\log P(B_i))}$
J	J-Measure	0 … 1	$\max(P(A,B)\log(\frac{P(B\|A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}))$ $P(A,B)\log(\frac{P(A\|B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})$
G	Gini index	0 … 1	$\max(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A}[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B}[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]-P(A)^2-P(\overline{A})^2)$
s	support	0 … 1	$P(A,B)$
c	confidence	0 … 1	$max(P(B\|A), P(A\|B))$
L	Laplace	0 … 1	$\max(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2})$
IS	Cosine	0 … 1	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
γ	coherence(Jaccard)	0 … 1	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
α	all_confidence	0 … 1	$\frac{P(A,B)}{\max(P(A),P(B))}$
o	odds ratio	0 … ∞	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(\overline{A},B)P(A,\overline{B})}$
V	Conviction	0.5 … ∞	$\max(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})})$
λ	lift	0 … ∞	$\frac{P(A,B)}{P(A)P(B)}$
S	Collective strength	0 … ∞	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
χ^2	χ^2	0 … ∞	$\Sigma_i \frac{(P(A_i)-E_i)^2}{E_i}$

May 20, 2012

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

symbol	measure	range	formula
ϕ	ϕ-coefficient	-1…1	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
Q	Yule's Q	-1 … 1	$\frac{P(A,B)P(\overline{A},\overline{B})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{A},\overline{B})+P(A,\overline{B})P(\overline{A},B)}$
Y	Yule's Y	-1 … 1	$\frac{\sqrt{P(A,B)P(\overline{A},\overline{B})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{A},\overline{B})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}}$
k	Cohen's	-1 … 1	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
PS	Piatetsky-Shapiro's	-0.25 … 0.25	$P(A,B)-P(A)P(B)$
F	Certainty factor	-1 … 1	$\max(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)})$
AV	added value	-0.5 … 1	$\max(P(B\|A)-P(B), P(A\|B)-P(A))$
K	Klosgen's Q	-0.33 … 0.38	$\sqrt{P(A,B)}\max(P(B\|A)-P(B), P(A\|B)-P(A))$
g	Goodman-kruskal's	0 … 1	$\frac{\Sigma_j \max_k P(A_j,B_k)+\Sigma_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-max_k P(B_k)}$
M	Mutual Information	0 … 1	$\frac{\Sigma_i\Sigma_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_J)}}{\min(-\Sigma_i P(A_i)\log P(A_i)\log P(A_i),-\Sigma_i P(B_i)\log P(B_i)\log P(B_i))}$
J	J-Measure	0 … 1	$\max(P(A,B)\log(\frac{P(B\|A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}))$ $P(A,B)\log(\frac{P(A\|B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})$
G	Gini index	0 … 1	$\max(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A}[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B}[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]-P(A)^2-P(\overline{A})^2)$
s	support	0 … 1	$P(A,B)$
c	confidence	0 … 1	$max(P(B\|A), P(A\|B))$
L	Laplace	0 … 1	$\max(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2})$
IS	Cosine	0 … 1	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
γ	coherence(Jaccard)	0 … 1	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
α	all_confidence	0 … 1	$\frac{P(A,B)}{\max(P(A),P(B))}$
o	odds ratio	0 … ∞	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(\overline{A},B)P(A,\overline{B})}$
V	Conviction	0.5 … ∞	$\max(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})})$
λ	lift	0 … ∞	$\frac{P(A,B)}{P(A)P(B)}$
S	Collective strength	0 … ∞	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
χ^2	χ^2	0 … ∞	$\Sigma_i \frac{(P(A_i)-E_i)^2}{E_i}$

#	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)	$\sqrt{P(A,B)}\max(P(B\|A) - P(B), P(A\|B) - P(A))$