دوره درسی ترم ۲ سال ۱۳۹۱

داده کاوی

میثم مدنی
دانشگاه صنعتی شریف
فصل هفتم،
http://www.madani.pro
Email: me@madani.pro

چشم‌انداز

داده کاوی

میثم مدنی

۱ مباحثی در کاوش الگو

مباحثی در کاوش الگو

داده کاوی

میثم مدنی

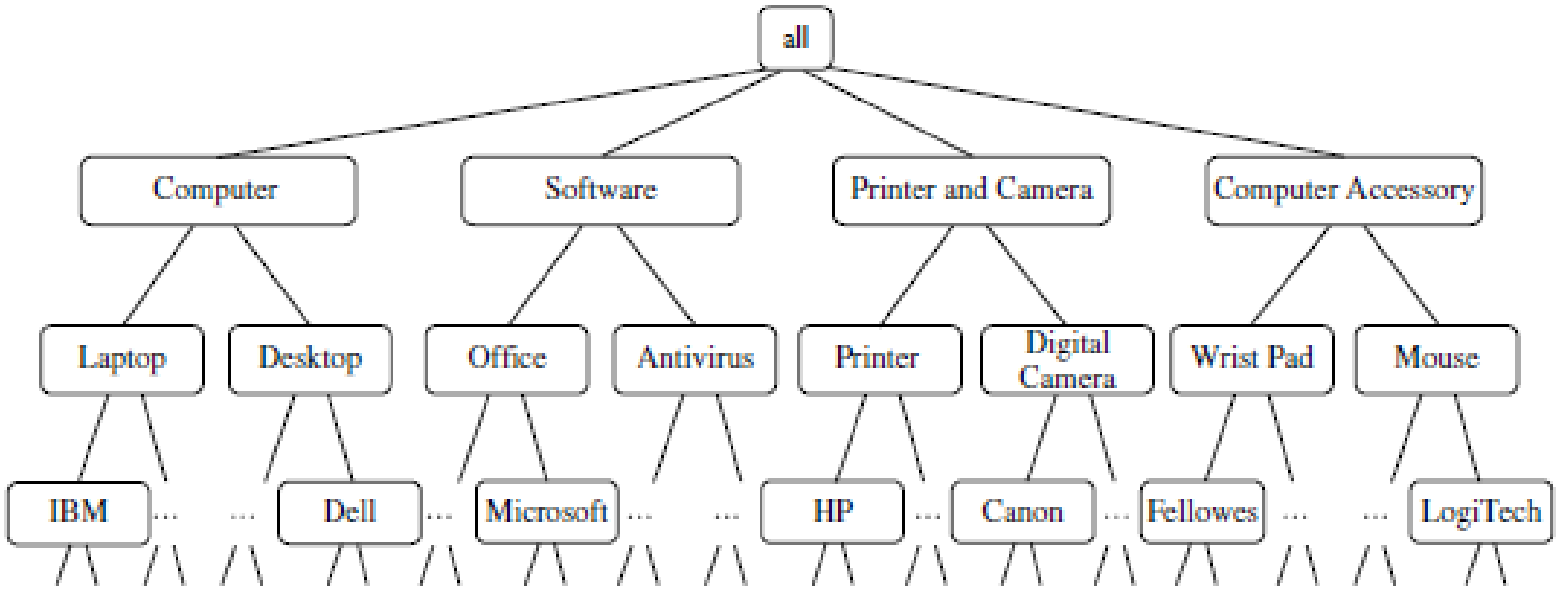
مرور کلی

کاوش الگو در فضای چند بعدی

کاوش قواعد شراکت چند سطحی

مسئله سلسله مراتب مطرح است و قاعده شراکت به معنای قبل قابل اکتشاف نخواهد بود!

TID	*Items Purchased*
T100	Apple 17″ MacBook Pro Notebook, HP Photosmart Pro b9180
T200	Microsoft Office Professional 2010, Microsoft Wireless Optical Mouse 5000
T300	Logitech VX Nano Cordless Laser Mouse, Fellowes GEL Wrist Rest
T400	Dell Studio XPS 16 Notebook, Canon PowerShot SD1400
T500	Lenovo ThinkPad X200 Tablet PC, Symantec Norton Antivirus 2010
…	…



کاوش قواعد شراکت چند سطحی

برای ما خیلی مهم است که بتوانیم قاعده شراکتی مثل نوت بوک دل به موس وایرلس بیابیم. به چیزی به نام تجرید Abstraction نیاز داریم.

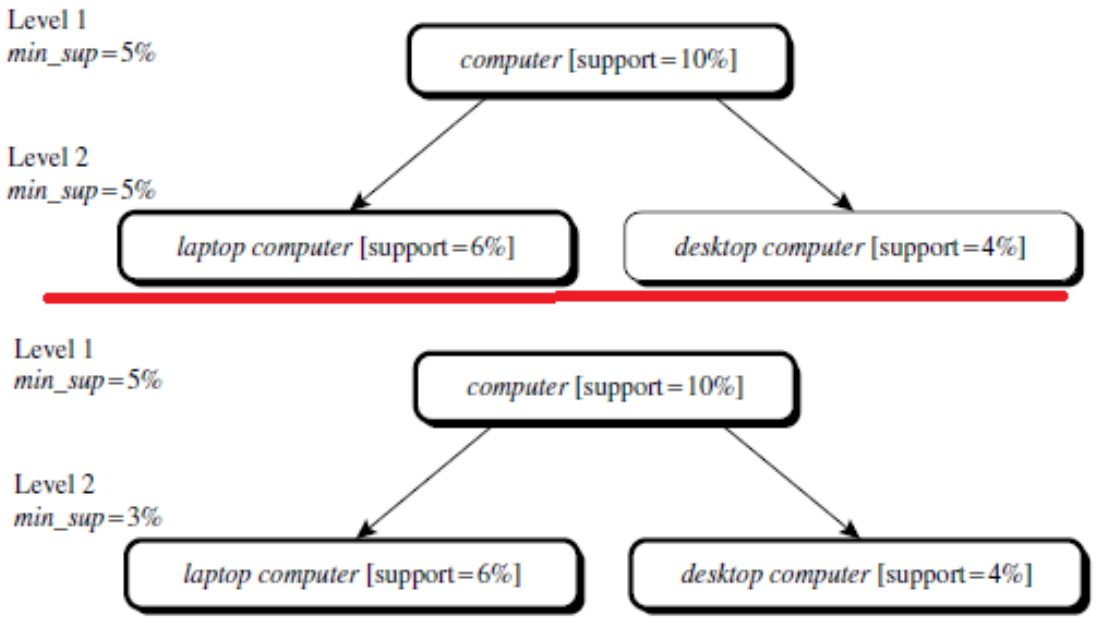
به قواعدی که به تجریدی این چنین نیاز دارند **قواعد شراکت چندسطحی** می گویند.

برخی روشها در زیر آمده است:

۱ - استفاده از پشتیبان یکنواخت برای تمامی سطوح

۲ - استفاده از مینیمم پشتیبان کاهشی

۳ - استفاده از مینیمم پشتیبان بر مبنای گروه یا مورد



کاوش قواعد شراکت چند سطحی

۱. قاعده یک **جد** قاعده دوم است اگر قاعده دوم با جایگزینی مواردش با اجدادشان به قاعده یک تبدیل شود.
۲. به این صورت یک سلسله مراتب روی قواعد شراکت ایجاد خواهد شد.
۳. یک قاعده **زائد** است اگر با توجه به اجدادش پشتیبان و اطمینانش به مقدار مورد انتظار نزدیک باشد.
۴. مثلا ما در laptop ۴ برند داشتیم و انتظار می رود پشتیبان آن یک چهارم شود.

$$buys(X, \text{“laptop computer”}) \Rightarrow buys(X, \text{“HP printer”})$$
$$[support = 8\%, confidence = 70\%]$$

$$buys(X, \text{“Dell laptop computer”}) \Rightarrow buys(X, \text{“HP printer”})$$
$$[support = 2\%, confidence = 72\%]$$

کاوش قواعد شراکت چند بعدی

۱ قواعدی که دو یا بیشتر بعد دارند را **قاعده شراکت چند بعدی** می گویند.
۲ قواعد شراکتی چند بعدی که پیش بینی تکراری دارند را **هیبرید** می گویند.

$$age(X, \text{“20...29”}) \wedge occupation(X, \text{“student”}) \Rightarrow buys(X, \text{“laptop”}).$$

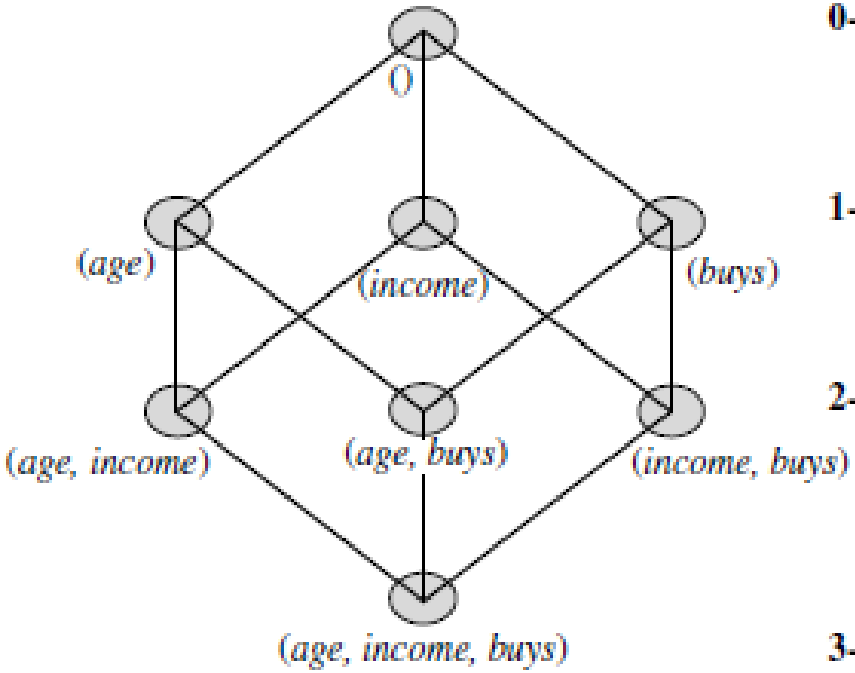
$$age(X, \text{“20...29”}) \wedge buys(X, \text{“laptop”}) \Rightarrow buys(X, \text{“HP printer”}).$$

روش‌های کشف قواعد شراکت را می توان به دو دسته تقسیم کرد.

۱ کمی
۲ اسمی

کاوش قواعد شراکت کمی

۱ بر مبنای مکعب داده



۲ بر مبنای خوشه (بالا به پایین، پایین به بالا)

۳ روش آماری(تشخیص استثناها)

$$sex = female \Rightarrow meanwage = \$7.90/hr \ (overall\_mean\_wage = \$9.02/hr).$$

$$population\_subset \Rightarrow \textbf{mean\_}of\textbf{\_}values\_for\_the\_subset,$$

کاوش قواعد شراکت کمی

۱ یک الگوی کمیاب الگویی است که تکراری نباشد!
۲ فرض کنید دنبال الگویی می گردیم که شامل حداقل یک مورد با ارزشی بیش از ۵۰۰ دلار داشته باشد.
۳ کاوش کارا از این الگو کمی مشکل است.
۴ باید چاره ای اندیشید

تعریف: اگر مجموعه‌های تکراری X, Y به ندرت با هم رخ دهند $\sup(X \cap Y) < \sup(X) \times \sup(Y)$ آنگاه این دو مجموعه را همبسته منفی می گویند. و $X \cap Y$ را الگوی همبسته منفی می گویند.
اگر $\sup(X \cap Y) << \sup(X) \times \sup(Y)$ دو مجموعه به طور قویا منفی همبسته اند.

فرض کنید یک فروشگاه ابزار خیاطی داریم

از هر بسته A, B ۱۰۰ بسته فروخته باشیم. و تنها یک فروش از هر دو داشته باشیم.
$\sup(X \cap Y) = 0.005 << \sup(X) \times \sup(Y) = 0.25$
فرض کنید فروش 10^6 فاکتور داریم.
$\sup(X \cap Y) = 10^{-6} >> \sup(X) \times \sup(Y) = 10^{-8}$!!!

کاوش قواعد شراکت کمی

مشکل تراکنش های خالی را این گونه م یتوان حل کرد NullTransaction

تعریف:

مجموعه‌های تکراری X, Y را به طور منفی همبسته می گویند اگر $(P(X|Y) + P(Y|X))/2 < \epsilon$ که ϵ آستانه الگوی منفی است.

فرض کنید یک فروشگاه ابزار خیاطی داریم

از هر بسته A, B ۱۰۰ بسته فروخته باشیم. و تنها یک فروش از هر دو داشته باشیم.

$$\epsilon = 0.02, \text{minsup} = 0.0001$$

$$(P(X|Y) + P(Y|X))/2 = (0.01 + 0.01)/2 = 0.01 < 0.02$$

فرض کنید فروش 10^6 فاکتور داریم.

$$(P(X|Y) + P(Y|X))/2 = (0.01 + 0.01)/2 = 0.01 < 0.02$$

کاوش الگوی تکراری مقید

- فرآیند داده کاوی ممکن است هزاران قاعده از داده ها کشف کند که بسیاری از آنها بی ربط یا بی استفاده برای کاربر است.
- کاربران اغلب دید خوبی از الگوهای مناسب و شکل الگوی مورد نظر را دارند.
- همچنین شرایطی ممکن است برای قواعد مورد نظر وجود داشته باشد.
- یک روش مناسب تعیین شرایط و خصوصیات به عنوان قید است که فضای جستجو را کاهش می دهد.
- به این فرآیند کاوش مقید می گویند Constraint – Based Mininng قیود می توانند شامل موارد زیر باشند

۱. **قیدهای تعیین نوع دانش** شراکت، همبستگی، خوشه‌بندی و طبقه‌بندی
۲. **قیود داده** تعیین نوع داده ها
۳. **قید بعد، سطح**
۴. **قید تمایل** تعیین آستانه هایی مانند پشتیبان، اطمینان و همبستگی
۵. **قید قاعده** تعیین نوع، شکل یا شرایط قاعده ای است که می خواهد کاوش شود این ها به عنوان ابر قاعده توصیف خواهند شد.

کاوش با ابرقاعده قواعد شراکت

- ابرقاعده این امکان را می دهد تا شکل های از قواعد را که برای ما جلب هستند پیدا کنیم
- فرض کنید که دسترسی به مشخصات خریداران داریم
- می خواهیم الگوهایی را کشف کنیم که خصیصه خریدار را به کالاهای خریداری شده وی مرتبط کند.
- به طور تخصصی تر می خواهیم بدانیم کدام جفت خصیصه ها منجر به خرید آفیس می شود؟

$$P_1(X, Y) \wedge P_2(X, W) \Rightarrow buys(X, \text{“office software”}).$$

$$age(X, \text{“30..39”}) \wedge income(X, \text{“41K..60K”}) \Rightarrow buys(X, \text{“office software”}).$$

$$P_1 \wedge P_2 \wedge \cdots \wedge P_l \Rightarrow Q_1 \wedge Q_2 \wedge \cdots \wedge Q_r, \ )$$

کاوش با ابرقاعده قواعد شراکت

قید قاعده ارتباط بین متغیرها، مقادیر اولیه متغیرها و قیود در توابع تجمعی در قواعد اکتشافی است.

مثال : فرض کنید که دسترسی به مشخصات خریداران داریم

- *item(item_ID, item_name, description, category, price)*
- *sales(transaction_ID, day, month, year, store_ID, city)*
- *trans_item(item_ID, transaction_ID)*

- یک حالت این است که در مورد الگوهای ارزان جستجو کنیم (با مجموع قیمت کمتر از ده دلار) $\text{sum(I.price)} < 10$
- الگوهای گران قیمت $\text{sum(I.price)} > 50$

هرس فضای الگوها با قیود هرس کننده

۵ نوع قید داریم

۱ **پادیکنوا** قید $\text{sum}(\text{I.price}) < 10$ را در نظر بگیرید k امین تکرار را در الگوریتم استقرایی در نظر بگیرید اگر مجموعه ای این قید را ارضا نکند دیگر نیازی نیست ادامه دهیم. این شاخه را هرس می کنیم.
در واقع اگر یک مجموعه این قید را برآورده نکند مجموعه های شامل آن هم آن قید را ارضا نمی کنند.

۲ **یکنوا** مشابه قسمت قبل اما برعکس!
$\text{sum}(\text{I.price}) > 50$ در واقع اگر یک مجموعه این قید را برآورده نکند مجموعه های مشمول در آن هم آن قید را ارضا نمی کنند.

۳ **فشرده** می توانیم مستقیما با قید تمام فضای مربوط به آن را مشخص کنیم.
$\min(\text{I.price}) > 50$

۴ **تبدیل‌پذیر** هیچ یک از موارد قبل نیست اما اگر آِیتم ها به نحو مناسبی مرتب شوند یکی از موارد قبل خواهد شد. مثلا $\text{avg}(\text{I.price}) > 50$ هیچ یک از موارد زیر نیست اما اگر از کوچک به بزرگ وارد کنیم پادیکنوا خواهد شد.

۵ **تبدیل ناپذیر** هیچ کدام!

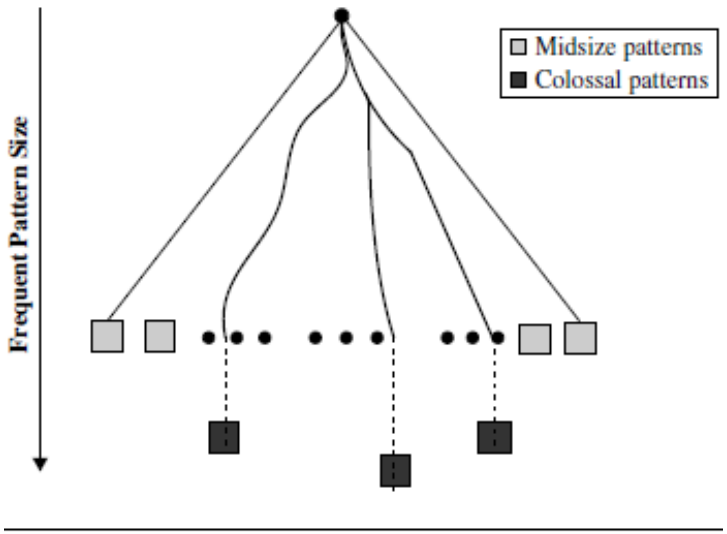
خبر خوب اینکه اکثر قیود از چهار نوع اول هستند.

Constraint	Antimonotonic	Monotonic	Succinct
$v \in S$	no	yes	yes
$S \supseteq V$	no	yes	yes
$S \subseteq V$	yes	no	yes
$min(S) \leq v$	no	yes	yes
$min(S) \geq v$	yes	no	yes
$max(S) \leq v$	yes	no	yes
$max(S) \geq v$	no	yes	yes
$count(S) \leq v$	yes	no	weakly
$count(S) \geq v$	no	yes	weakly
$sum(S) \leq v\ (\forall a \in S, a \geq 0)$	yes	no	no
$sum(S) \geq v\ (\forall a \in S, a \geq 0)$	no	yes	no
$range(S) \leq v$	yes	no	no
$range(S) \geq v$	no	yes	no
$avg(S)\ \theta\ v, \theta \in \{\leq, \geq\}$	convertible	convertible	no
$support(S) \geq \xi$	yes	no	no
$support(S) \leq \xi$	no	yes	no
$all\_confidence(S) \geq \xi$	yes	no	no
$all\_confidence(S) \leq \xi$	no	yes	no

کاوش الگوهای بسیار بزرگ و چند بعدی

کاوش الگوهای بسیار بزرگ با آمیختگی الگو

فرض کنید داده های زیر را داریم



row/col	1	2	3	4	…	38	39
1	2	3	4	5	…	39	40
2	1	3	4	5	…	39	40
3	1	2	4	5	…	39	40
4	1	2	3	5	…	39	40
5	1	2	3	4	…	39	40
…	…	…	…	…	…	…	…
39	1	2	3	4	…	38	40
40	1	2	3	4	…	38	39
41	41	42	43	44	…	78	79
42	41	42	43	44	…	78	79
…	…	…	…	…	…	…	…
60	41	42	43	44	…	78	79

در اغلب روش های جستجو از رشد تعداد آیتم ها استفاده می کنیم.
این روش ها نمی توانند از پس تولید حجم عظیم الگوهای با اندازه میانه بر بیایند.
به روش جدیدی نیازی داریم.

کاوش الگوهای بسیار بزرگ با آمیختگی الگو

خواص اصلی آمیختگی الگو

- درخت را در یک حالت با عرض محدود طی می کند.
- تعداد متناهی الگو با اندازه متوسط به عنوان نقاط شروع آمیختن انتخاب به سمت پایین درخت انتخاب می شوند.
- به مسئله فضای جستجو نمایی بر نمی خوریم.

۱ **الگوی هسته** برای یک الگوی α مجموعه موارد $\beta \subseteq \alpha$ را $\tau-$هسته می گوییم اگر

$$\frac{|D_\alpha|}{|D_\beta|} \geq \tau, \quad 0 \leq \tau \leq 1$$

که در آن $|D_\alpha|$ تعداد الگوهای شامل α در داده هاست.

۲ τ را **نسبت هسته** می گوییم.

۳ الگوی α را $(\mathrm{d}, \tau)-$ قوی می گوییم اگر d حداکثر تعداد مواردی باشد که می تواند از α برداشته شود تا یک الگوی $\tau-$ هسته باقی بماند

$$d = \max_{\beta}\{|\alpha| - |\beta| | \beta \subseteq \alpha, \text{ and } \beta \text{ is a } \tau\text{-core pattern of } \alpha\}$$

مثال

فرض کنید هر کدام از الگوهای زیر صد بار تکرار داشته باشد.

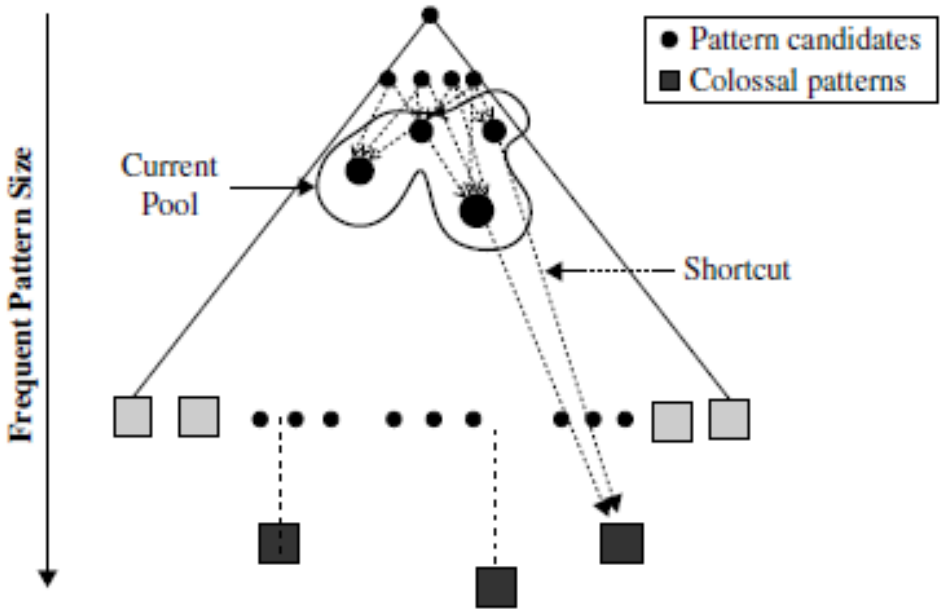
$$\{\alpha_1 = (abe),\ \alpha_2 = (bcf),\ \alpha_3 = (acf),\ \alpha_4 = (abcfe)\}$$

اگر بگیریم $\tau = 0.5$ آنگاه (ab) یک الگوی هسته برای α_1 است که فقط در α_1 و α_2 قرار دارد.
بنابراین $\frac{|D_\alpha|}{|D_\beta|} = \frac{100}{200} \geq \tau$ پس α_1 $(2, 0.5)$ قوی است و α_4 یک $(4, 0.5)$ قوی است.

Transactions (# of Transactions)	*Core Patterns* ($\tau = 0.5$)
(*abe*) (100)	(*abe*), (*ab*), (*be*), (*ae*), (*e*)
(*bcf*) (100)	(*bcf*), (*bc*), (*bf*)
(*acf*) (100)	(*acf*), (*ac*), (*af*)
(*abcef*) (100)	(*ab*), (*ac*), (*af*), (*ae*), (*bc*), (*bf*), (*be*), (*ce*), (*fe*), (*e*), (*abc*), (*abf*), (*abe*), (*ace*), (*acf*), (*afe*), (*bcf*), (*bce*), (*bfe*), (*cfe*), (*abcf*), (*abce*), (*bcfe*), (*acfe*), (*abfe*), (*abcef*)

- **منبع اولیه** از موارد تکراری کوچک شروع می کنیم.
- **تکرار آمیختن الگوها** یک پارامتر را به عنوان ورودی می گیرد K ماکزیمم تعداد الگوهای قابل اکتشاف.

در هر مرحله K تا دانه انتخاب می کنیم به عنوان منبع (به صورت تصادفی) حال در هر مر حله تمام الگوهایی را که در گوی به سایز τ باشند را در نظر می گیریم.
تمام این الگوها را برای ابرالگوها ترکیب می کنیم.
این ابرالگوها یک منبع جدید تشکیل می دهند.



کاوش الگوهای تخمینی و فشرده

۱ فاصله دو الگوی تکراری

$$Pat\_Dist(P_1, P_2) = 1 - \frac{|T(P_1) \cap T(P_2)|}{|T(P_1) \cup T(P_2)|}.$$

۲ مثلا

$$T(P_1) = \{t_1, t_2, t_3, t_4, t_5\}$$
$$T(P_2) = \{t_1, t_2, t_3, t_4, t_6\},$$
$$Pat\_Dist(P_1, P_2) = 1 - \frac{4}{6} = \frac{1}{3}.$$

۳ یک الگو P با الگوی دیگر P'، $\lambda-$ پوشیده می شود اگر $O(P) \subseteq O(P')$ و $Dist(P, P') \leq \lambda$ که منظور از $O(P)$ مجموعه موارد P است

۴ تخمین

$$\delta \geq Pat\_Dist(P, P_r) = 1 - \frac{|T(P_r)|}{|T(P)|} \geq 1 - \frac{k}{min\_sup}.$$

$$\delta \geq Pat\_Dist(P, P_r) = 1 - \frac{|T(P_r)|}{|T(P)|} \geq 1 - \frac{k}{min\_sup}.$$

k is support P_r.