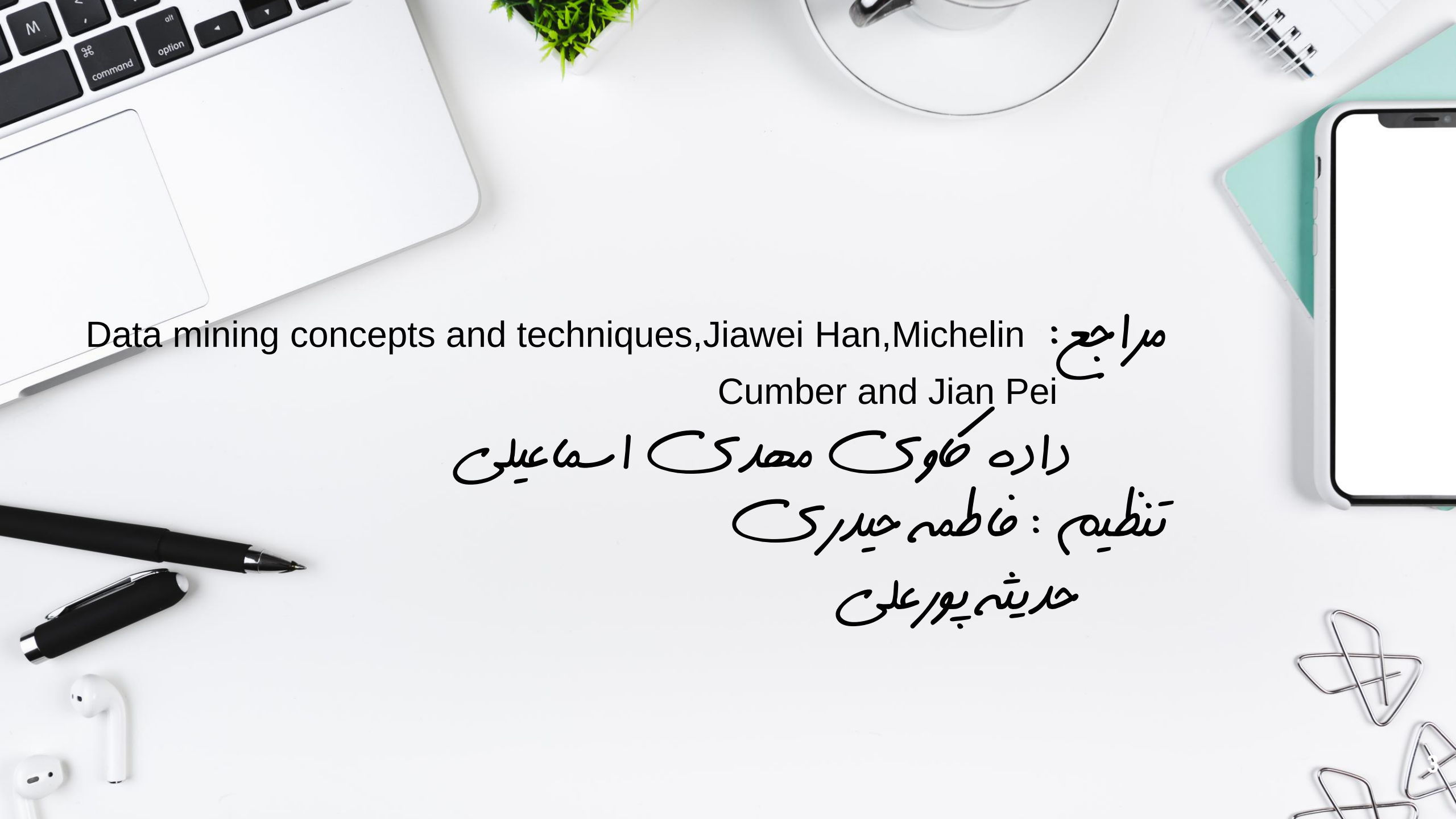

بسم الله الرحمن الرحیم

داده کاوی

دکتر حسام عمرانپور

دانشگاه صنعتی نوشیروانی بابل



مراجع:

Data mining concepts and techniques,Jiawei Han,Michelin Cumber and Jian Pei

داده کاوی مهدی اسماعیلی

تنظیم: فاطمه حیدری

حدیثه پورعلی

ادامه فصل 1

آماده سازی داده ها

4- Data Reduction

4-3: گسسته سازی:

4-3-1. Entropy Based

4-3-2. Chi Merge

4- Data Reduction

4-3-1. Entropy Based

الگوریتم:

1. ابتدا مقادیر (نمونه ها) را مرتب می کنیم.
2. برای هر نمونه i :
   - داده ها به دو زیرمجموعه افراز می شوند : D_1 , D_2
   - Info این دو زیرمجوعه محاسبه می شود.
3. هر داده ای که Info کوچکتری داشت به عنوان مرز یک سطح انتخاب می شود.
4. 2 و 3 تکرار می شوند تا به شرط خاتمه برسیم.(شرط خاتمه می تواند یک Info خاص و یا یک تعداد از سطوح گسسته موردنظر باشد.)

4- Data Reduction

- $Info(D) = \frac{|D_1|}{|D|} \times Entropy(D_1) + \frac{|D_2|}{|D|} \times Entropy(D_2)$
- $Entropy(D_k) = -\sum_{i=1}^{k} p_i \log_2 p_i$

  - $|D|$: تعداد نمونه ها
  - $|D_k|$: تعداد نمونه های افراز i
  - p_i : احتمال رخداد کلاس c_i در D_k
  - c_i : شماره کلاس
  - **سوال:** کدام Info بهتر است؟

- $Info(D) = \sum_{i1}^{l} \frac{|D_i|}{|D|} \times Entropy(D_i)$

4- Data Reduction

4-3-2. Chi merge

الگوریتم :

1. ابتدا مقادیر را مرتب (صعودی) می کنیم.
2. هر عنصر در یک بازه قرار قرار می گیرد (n بازه) روش پایین به بالاست.
3. برای هر دو بازه مجاور محاسبه می شود و دوباره با کوچکترین مقدار k^2 ادغام k^2 می شوند.

4- Data Reduction

- $K^2 = \sum_i^2 \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}}$
- $E_{ij} = \frac{R_i \times C_j}{N}$

- K : تعداد کلاس
- A_{ij} : تعداد نمونه ها در i_امین بازه و در j_امین کلاس
- E_{ij} : فراوانی مورد انتظار برای A_{ij}

4- Data Reduction

Att	Class
1	A
3	B
2	A
8	A
9	A
11	B
23	B
37	A
39	B
45	A
46	A
59	A

- مثال :

[0,2) [2,5) [5,7.5) [7.5,8.5) [8.5,10).....

	Class A	Class B	
[7.5,8.5)	$A_{11} = 1$	$A_{12} = 0$	$R_1 = 1$
[8.5,10)	$A_{21} = 1$	$A_{22} = 0$	$R_2 = 1$
	$C_1 = 2$	$C_2 = 0$	$N = 2$

$$E_{11} = \frac{2}{2} = 1 \qquad E_{12} = \frac{0}{2}$$

$$E_{21} = \frac{2}{2} = 1 \qquad E_{22} = \frac{0}{2}$$

فصل 2

الگوهای مکرر و قوانین انجمنی

Frequent Patterns and Association Rules

- الگوهای مکرر و قوانین انجمنی
  - تحلیل سبد خرید

TID	Items
1	{Bread , Milk}
2	{Bread , Egg , Butter , Cheese}
3	{Milk , Butter , Cheese , Chicken }
4	{Bread , Cheese , Chicken}
5	{Bread , Milk , Butter , Cheese}

Frequent Patterns and Association Rules

- Support **(پشتیبان)** : تعداد (درصد) تراکنش هایی که شامل اقلام مجموعه باشد.

Support {Cheese , Butter} = 3 $\Rightarrow$ $\frac{3}{5} = \frac{6}{10} = 60\%$

- اگر در یک مجموعه D قلم داشته باشیم، $2^D - 1$ مجموعه حالت وجود دارد.
- از این حالات دنبال حالاتی هستیم که مقدار support آن از minsupport (توسط کاربر مشخص می شود) بیشتر باشد.

- **مجموعه اقلام مکرر** : مجموعه هایی که مقدار پشتیبان آن از minsupport بیشتر باشد.

Frequent Patterns and Association Rules

- **قوانین انجمنی**

مجموعه A ⟶ مجموعه B

{Milk} ⟶ {Bread}

- Support **یک قانون** : درصدی از مجموعه های موجود که شامل (مجموعه A ∪ مجموعه B) باشد.
- Confidence **یک قانون** : درصدی از مشتریانی که A خریده اند، B را هم خریده باشند.

$$\text{Confidence} = \frac{\text{تعداد قوانینی که شامل } A \text{ و } B \text{ باشد}}{\text{تعداد قوانینی که شامل } A \text{ باشد}}$$

Frequent Patterns and Association Rules

$X \rightarrow Y \quad , \; X \cap Y = \emptyset$

Support $(X \rightarrow Y) = P(X \cap Y)$

Confidence $(X \rightarrow Y) = P(X|Y) = \frac{P(X \cap Y)}{P(X)}$

- **سوال :** آیا رابطه زیر بر قرار است ؟

$$X \rightarrow Y \quad \equiv \quad Y \rightarrow X$$

Frequent Patterns and Association Rules

- مثال :

TID	Items
1	{1,3,4}
2	{2,3,5}
3	{1,2,3,5}
4	{2,5}

$R_1 : \{1\} \rightarrow \{3\}$

Support = 50% confidence = 100%

$R_2 : \{1\} \rightarrow \{4\}$

Support = 25% confidence = 50%

$R_3 : \{2,3\} \rightarrow \{5\}$

Support =50% confidence = 100%

$R_4 : \{2\} \rightarrow \{3,5\}$

Support =50% confidence = 66.6%

Frequent Patterns and Association Rules

- Strong Rules (**قوانین قوی**) : قوانینی که confidence و support آن ها بزرگتر از minconfidence و minsupport باشد.
  - **سوال :** minconfidence و minsupport چقدر باشد تا R_1 و R_3 در مثال قبل به عنوان قوانین قوی انتخاب شوند؟
  - **سوال :** تراکنشی که confidence آن زیاد و support آن کم باشد، چه قانونی است؟
  - **سوال :** تراکنشی که confidence آن کم و support آن زیاد باشد، چه قانونی است؟
- فرض کنیم d قلم داده وجود دارد، تعداد قوانین انجمنی برابر است با :

$$\sum_{k}^{d-1}\left(\binom{d}{k}\times\sum_{j=1}^{d-k}\binom{d-k}{j}\right)=3^{d}-2^{d+1}+1$$

Frequent Patterns and Association Rules

الگوریتم Apriori جهت یافتن الگوهای مکرر

- مانند الگوریتم BFS عمل می کند.
  - مرحله 1 : مجموعه اقلام یکتایی
  - مرحله 2 : مجموعه اقلام 2 تایی
  - مرحله k : مجموعه اقلام k تایی
- پس از محاسبه مقدار پشتیبان مرحله k و مقایسه با minsup الگوهای مکرر شناسایی می شوند. آن هایی که مکرر نیستند را ادامه نمی دهیم (از آن ها یال خارج نمی شود)، چون اگر یک الگوی مکرر داشته باشیم زیرمجموعه های آن مکرر هستند.(اگر یک زیرمجموعه مکرر نباشد، مجموعه ای که شامل این زیرمجموعه باشد، مکرر نخواهد بود.)
- در این حالت، فضای جستجو کاهش میابد.(با هرس کردن درخت)

Frequent Patterns and Association Rules

- **خاصیت یکنواختی** (monotone) : یک رابطه f دارای خاصیت یکنواختی است اگر :

$$\forall\, x, y \in J : (X \subseteq Y) \Rightarrow f(x) \leq f(y)$$

- **خاصیت پادیکنواختی** (Antimonotone) :

$$\forall\, x, y \in J : (X \subseteq Y) \Rightarrow f(y) \leq f(x)$$

- الگوریتم Apriori دارای خاصیت پادیکنواختی است.

Frequent Patterns and Association Rules

- مثال :

TID	Items
1	{1,2,3,4,5,6}
2	{2,3,4,5,6,7}
3	{1,4,5,8}
4	{1,4,6,9,10}
5	{2,4,5,10,11}

Minsupport = 60%

Frequent Patterns and Association Rules

Itemset1	Support	Frequent_Itemset1
1	60%	√
2	60%	√
3	40%	×
4	100%	√
5	80%	√
6	60%	√
7	20%	×
8	20%	×
9	20%	×
10	40%	×
11	20%	×

Itemset2	Support	Frequent_Itemset2
1,2	20%	×
1,4	60%	√
1,5	40%	×
1,6	40%	×
2,4	60%	√
2,5	60%	√
2,6	40%	×
4,5	80%	√
4,6	60%	√
5,6	40%	×

Frequent Patterns and Association Rules

Itemset3	Support	Frequent_Itemset3
1,4,2	_	×
1,4,5	_	×
1,4,6	_	×
2,4,1	_	×
2,4,5	60%	√
2,4,6	_	×
2,5,1	_	×
2,5,4	_	×
4,6,1	_	×
4,5,6	_	×

Itemset4	Support	Frequent_Itemset4
2,4,5,1	_	×
2,4,5,6	_	×
….	….	….

Frequent Patterns and Association Rules

- مجموعه ی تعداد حالات این مثال :

$$\binom{11}{1} + \binom{11}{2} + \binom{11}{3} > 11 + 10 + 1$$

- برای پیدا کردن قوانین مکرر :

از آن هایی که جدول 2تایی هستند، شروع می کنیم و آن هایی که frequent item آن ها تیک خورده از minsup بیشتر است، فقط کافی است حالت های قانون را از مجموعه ها دربیاوریم و برای 3تایی ها مثلا اولی به دومی و سومی و از سومی به اولی و دومی و از دومی به اولی و سومی و بالعکس.یعنی از دومی و سومی به اولی و از سومی به دومی و اولی و از اولی به دومی و سومی و .....

Frequent Patterns and Association Rules

- نمایش تراکنش به صورت نمونه_ویژگی:
  - مثال :

ID	Items
1	{1,3,5,7,8}
2	{3,5}
3	{2,1}
4	{1,3,4}
5	{2,1}
6	{5,7,6}

ویژگی\نمونه	1	2	3	4	5	6	7	8
1	1	0	1	0	1	0	1	1
2	0	0	1	0	1	0	0	0
3	1	1	0	0	0	0	0	0
4	1	0	1	1	0	0	0	0
5	1	1	0	0	0	0	0	0
6	0	0	0	0	1	1	1	0

Frequent Patterns and Association Rules

- تبدیل نمونه و ویژگی ← تراکنش i :

ویژگی\نمونه	D_1	D_2	D_3	D_4	D_5
s_1	0	1	0	1	1
s_2	1	0	0	1	0
s_3	1	0	0	1	0
s_4	0	1	0	0	1
s_5	1	0	0	1	1

ID	Items
1	{2,4,5}
2	{1,4}
3	{1,4}
4	{2,5}
5	{1,4,5}

Frequent Patterns and Association Rules

- **سوال :** اگر اعداد اعشاری باشند یا اعداد به صورت 0 و 1 نباشند، باید چه کار کنیم ؟

ویژگی\نمونه	D_1	D_2	D_3	D_4	D_5
s_1	36	52.5	25	-3	0.4
s_2	13	0	2	1.5	0
s_3	7	8	9	2	0
s_4	11	20	14	10	15

Frequent Patterns and Association Rules

- **فشرده سازی الگوهای مکرر :**

هنگامی که minsup کوچک باشد یا الگوهای مکرر طول بزرگ داشته باشند، حافظه ی زیادی برای ذخیره سازی تمام الگوهای مکرر نیاز است.لذا دو مفهوم معرفی می شود :

1. **ماکسیمال (maximal)** : یک الگوی مکرر هنگامی ماکسیمال است که هیچ یک از ابرمجموعه های آن مکرر نباشد. درواقع برای تعیین ماکسیمال بودن {a,b} کافی است تمام ابرمجموعه های بلافصل آن در پایگاه داده بررسی شوند که مکرر نباشد :

اقلام : {a,b,c,d,e,f}

ابرمجموعه های بلافصل {a,b} : {a,b,c} , {a,b,d} , {a,b,e} , {a,b,f}

Frequent Patterns and Association Rules

- **مثال :**

ID	Items
1	1,2,3,…,20
2	1,2,….,10

Minsup = 50%

پاسخ :

Maximal = {1,2,3,…,20}

- دیگر لازم نیست الگوهای مکرر 4 عضوی بررسی شوند.چرا؟
- با داشتن ماکسیمال دیگر نیازی به نگهداری تمام الگوهای مکرر نیست.
- الگوهای مکرر ماکسیمال هیچ اطلاعاتی درباره support زیرمجموعه های خود ندارند، به جز این که فقط می دانند زیرمجموعه هایش مکرر است.

Frequent Patterns and Association Rules

2. **بسته** (closed) : مجموعه بسته مجموعه ای است که هیچ یک از ابرمجموعه های بلافصل آن مقدار پشتیبان دقیقا برابر با پشتیبان X نداشته باشد. اگر حداقل یک ابرمجموعه بلافصل X پشتیبان برابر X داشته باشد، X بسته نیست.اگر X مکرر باشد، X مکرر بسته است.

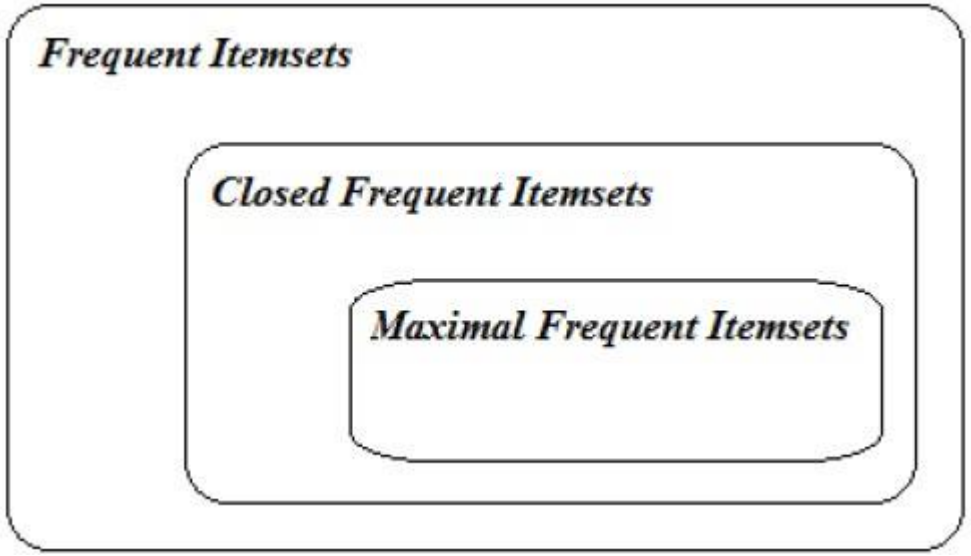
- به کمک الگوهای مکرر بسته می توانیم مقدار پشتیبان الگوهای مکرر که بسته نیستند را محاسبه نماییم. به این منظور باید مقدار پشتیبان کلیه ابرمجموعه های بلافصل آن ها را داشته باشیم. همه ی ابرمجموعه های بلافصل، مقدار پشتیبان کوچکتری از X خواهند داشت.

❑ **مثال :** closed مثال قبل را بدست آورید.

Closed = {1,2,…,10} , {1,2,3,…,20}

Frequent Patterns and Association Rules

- رابطه بین الگوهای مکرر، الگوهای مکرر بسته و الگوهای مکرر ماکسیمال :



- **نکته :** برای هر الگوی k تایی می توان $\sum_{i=1}^{k-1}\binom{k}{i} = 2^k - 2$ قانون انجمنی تولید کرد.