Regras de Associação

Sarajane M. Peres e Clodoaldo A. M. Lima

19 de novembro de 2015

Material baseado em:
HAN, J. & KAMBER, M. Data Mining: Concepts and Techniques. 2nd. 2006

Regras de Associação

Algumas regras ....

Quem compra cerveja, também compra fraldas.

Quem compra pão, também compra leite.

Quem compra queijo, também compra presunto. Quem compra presunto, também compra queijo.

Quem compra coca-cola, também compra sonho de valsa.

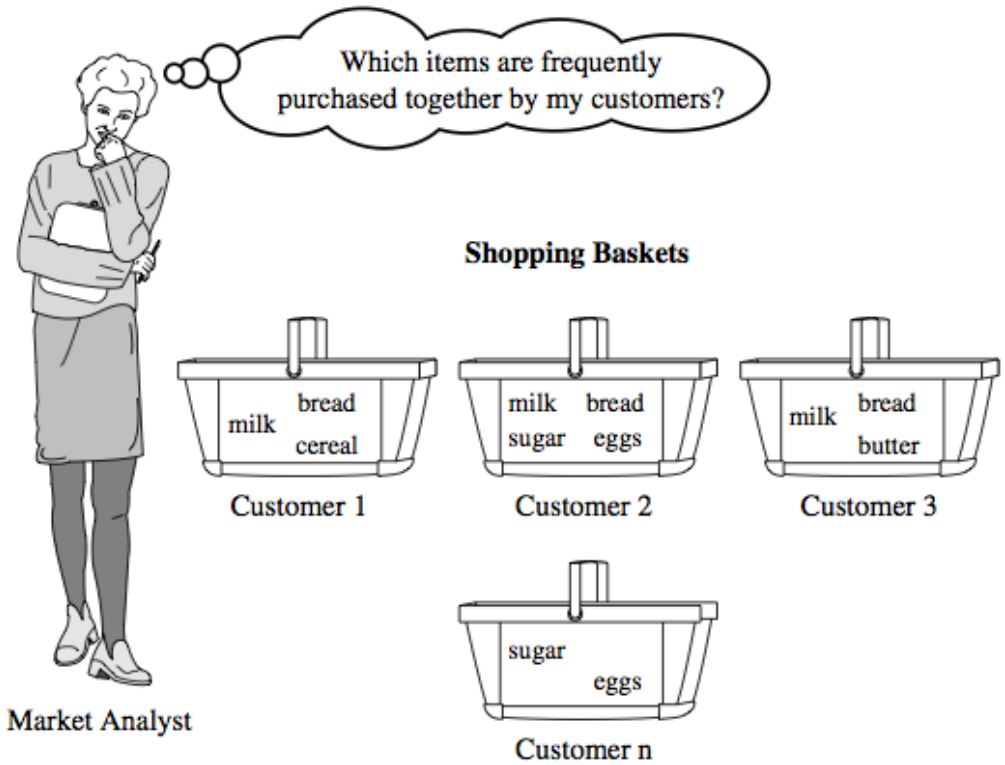
Famílias com muitos filhos, não possuem televisão.

Quem votou a favor de religião nas escolas, votou a favor do "país X".

Regras de Associação

Exemplo - Análise da cesta de compras

Este processo analisa os hábitos de compra de clientes por meio da descoberta de associações entre diferentes itens que aparecem nas "cestas de compras". A descoberta destas associações ajuda os varejistas no desenvolvimento de estratégias de marketing já que revelam quais itens são frequentemente comprados juntos pelos clientes.



Regras de Associação

Exemplo - Análise da cesta de compras

Pensando no nosso universo como sendo um conjunto de itens disponíveis em uma loja, a cada item podemos associar uma variável booleana que representa a presença ou ausência daquele item em um evento.

Assim, cada "compra" (ou transação) pode ser representada por um vetor booleano de valores associados a estas variáveis. Os vetores booleanos, então, podem ser analisados como padrões de compras que refletem itens que são *frequentemente associados* ou *comprados juntos*.

Esses padrões podem ser representados na forma de **regras de associação**.

Exemplo

A informação sobre "clientes que compram computadores também tendem a comprar software antivírus" pode ser representada na regra de associação:

$$computer \Rightarrow antivirus\ software\ [support = 2\%, confidence = 60\%]$$

Regras de Associação

Regra de associação

$$computer \Rightarrow antivirus\ software\ [support = 2\%, confidence = 60\%]$$

Interpretando a regra ....

Suporte (*support*) e **confiança** (*confidence*) são duas medidas de "interessabilidade" (*interestingness*), que refletem respectivamente a **utilidade** e **confiabilidade** da regra descoberta.

Um suporte de 2% para uma regra de associação significa que 2% de todas as transações sob análise mostram que computadores e antivirus são comprados juntos.

A confiança de 60% significa que 60% das compras onde os clientes compraram computadores também apresentam o item antivirus como item vendido.

Tipicamente, regras de associação são consideradas de interesse se elas satisfazem tanto um **suporte mínimo** quanto uma **confiança mínima**.

Regras de Associação

Definições

Seja $I = \{I_1, I_2, ..., I_m\}$ um **conjunto de itens**. Seja D, um conjunto de dados transacionais onde cada transação T é um conjunto de itens tal que $T \subseteq I$. Cada transação possui um identificador TID. Seja A um subconjunto de itens. É dito que T contém A **se e somente se** $A \subseteq T$.

Uma **regra de associação** é uma implicação da forma $A \Rightarrow B$, onde $A \subset I, B \subset I$ e $A \cap B = \emptyset$.

A regra $A \Rightarrow B$ vale no conjunto de transações D com **suporte** s, onde s é a porcentagem de transações em D que contém $A \cup B$.

A regra $A \Rightarrow B$ tem **confiança** c no conjunto de transações D, onde c é a porcentagem de transações em D contendo A que também contém B.

Regras de Associação

Uma regra que satisfaça tanto um suporte mínimo (*min_sup*) quando uma confiança mínima (*min_conf*) é chamada de **forte**.

Definições

Um conjunto de itens é chamado de **itemset**. Um itemset que contem k itens é um k-**itemset**. O conjunto $\{computer, antivirus\}$ é um 2-itemset.

A **frequência de ocorrência de um itemset** é o número de transações que contém o itemset. Isto também é conhecido como **frequência**, **suporte** ou **contagem** de um itemset. Se um itemset I satisfaz um suporte mínimo ele é dito um **itemset frequente**. Um conjunto de k-itemsets frequentes é chamado de L_k

Regras de Associação

$$confidence(A \Rightarrow B) = \frac{support(A \cup B)}{support(A)}.$$

Definições

A regra acima mostra que a confiança de uma regra $A \Rightarrow B$ pode ser facilmente derivada do suporte de A e do suporte de $A \cup B$. Isto é, uma vez que o suporte de A, B e $A \cup B$ são conhecidos, é possível derivar $A \Rightarrow B$ e $B \Rightarrow A$, e checar se tais regras são fortes.

O problema de minerar regras de associação pode ser reduzido ao problema de minerar itemsets frequentes.

Regras de Associação

Procedimento geral

- **Encontrar todos os itemsets frequentes:** Por definição, para ser considerado frequente, o itemset deve ocorrer pelo menos tão frequentemente quanto um suporte mínimo predeterminado, *min_sup*.

- **Gerar regras de associação fortes a partir dos itemsets frequentes:** Por definição, estas regras devem satisfazer um suporte mínimo e uma confiança mínima.

Regras de Associação

Desafio

O principal desafio de mineração de itemsets frequentes em grandes bases de dados é que, frequentemente, esse processo gera um número muito grande de itemsets frequentes. Isto acontece porque, **se um itemset é frequente, cada um de seus subconjuntos também o é**. Para superar esta dificuldade, introduz-se o conceito de *itemsets frequentes fechados* ou *maximal itemset frequente*.

Definições

Um itemset X é **fechado** em um conjunto de dados S, se não existir nenhum super-itemset próprio[a] Y ($X \subset Y$) tal que Y tenha o mesmo suporte que X.

Um itemset X é um **itemset frequente fechado** no conjunto de dados S se X é tanto fechado quanto frequente em S.

Um itemset X é um **maximal itemset frequente** (ou **max-itemset**) no conjunto de dados S se X é frequente, e não existe um super-itemset Y tal que $X \subset Y$ e Y é frequente em S.

---

[a] Y contém pelo menos um item a mais que X.

Regras de Associação

Definições

Seja C o conjunto de itemsets frequentes fechados para o conjunto de dados S satisfazendo um suporte mínimo, *min_sup*. Seja M o conjunto de maximal itemsets frequentes para S satisfazendo o *min_sup*.

Suponha que nós tenhamos o suporte de cada itemset em C e em M. Note que C e sua informação de suporte pode ser usada para derivar todo o conjunto de itemsets frequentes. Assim nós temos que C contém a informação completa referente aos itemsets frequentes.

Por outro lado, M registra somente o suporte dos itemsets máximos.

Regras de Associação

Ilustração - Itemsets frequentes fechados e máximos

Suponha que o banco de dados transacional tem somente duas transações: $\{\langle a_1, a_2, ..., a_{100}\rangle; \langle a_1, a_2, ..., s_{50}\rangle\}$
Suponha que o $min\_sup = 1$. Existem dois itemsets frequentes fechados (e seus suportes): $C = \{\{a_1, a_2, ..., a_{100}\} : 1; \{a_1, a_2, ..., a_{50}\} : 2\}$
Existe um itemset frequente máximo: $M = \{\{a_1, a_2, ..., a_{100}\} : 1\}$

O conjunto de itemsets frequentes fechados (C) contém a informação completa referente aos itemsets frequentes. Por exemplo, de C, é possível derivar:

1. $\{a_2, a_{45} : 2\}$ desde que $\{a_2, a_{45}\}$ é um sub-itemset de $\{a_1, a_2, ..., a_{50}\} : 2\}$;
2. $\{a_8, a_{55} : 1\}$ desde que $\{a_8, a_{55}\}$ é um sub-itemset de $\{a_1, a_2, ..., a_{100}\} : 1\}$;

De M é possível somente afirmar que ambos itemsets são frequentes, mas não é possível afirmar os seus suportes.

Regras de Associação

Minerando: itemsets frequentes booleanos, de único nível e dimensão única.

Explorando o algoritmo **Apriori**: o algoritmo básico para encontrar itemsets frequentes, e a partir disso, gerar regras de associação fortes.

Trata-se de um algoritmo proposto por R.Agrawal e R. Srikant, em 1994. O nome do algoritmo é baseado no fato que o algoritmo usa conhecimento *a priori* sobre propriedades de itemset frequentes. Nele é empregado uma abordagem iterativa onde k-itemsets são usados para explorar $(k+1)$-itemsets. De forma resumida:

- o conjunto de 1-itemsets frequentes é encontrado por meio da varredura do banco de dados para contagem de cada item, e da descoberta daqueles itens que satisfazem um suporte mínimo. O resultado é chamado de L_1.
- L_1 é usado para encontrar L_2, o conjunto de 2-itemsets frequentes, o qual é usado para encontrar L_3 e assim por diante, até que nenhum k-itemset frequente possa ser encontrado. Encontrar L_k requer uma leitura completa do banco de dados.

Regras de Associação

Apriori Property

Todos os subconjuntos não vazios de um itemset frequente deve também ser frequente.

Essa propriedade permite reduzir o esforço de busca por itemsets frequentes. Ela é baseada nas seguintes observações:

- se um itemset I não satifaz o suporte mínimo, min_sup, então I não é frequente; ou seja, $P(I) < min\_sup$.
- se um item A é adicionado ao itemset I, então o itemset resultante (i.e. $I \cup A$) não pode ocorrer com mais frequência do que I;
- portanto, $I \cup A$ não é frequente também; ou seja, $P(I \cup A) < min\_sup$.

Esta propriedade pertence à classe de propriedades chamadas **antimonotônicas** no sentido que *se um conjunto não pode passar num teste, todos os seus superconjuntos falharão no mesmo teste*. Ela é chamada *antimonotônica* porque ela é uma propriedade *monotônica* no contexto de falhas em teste.

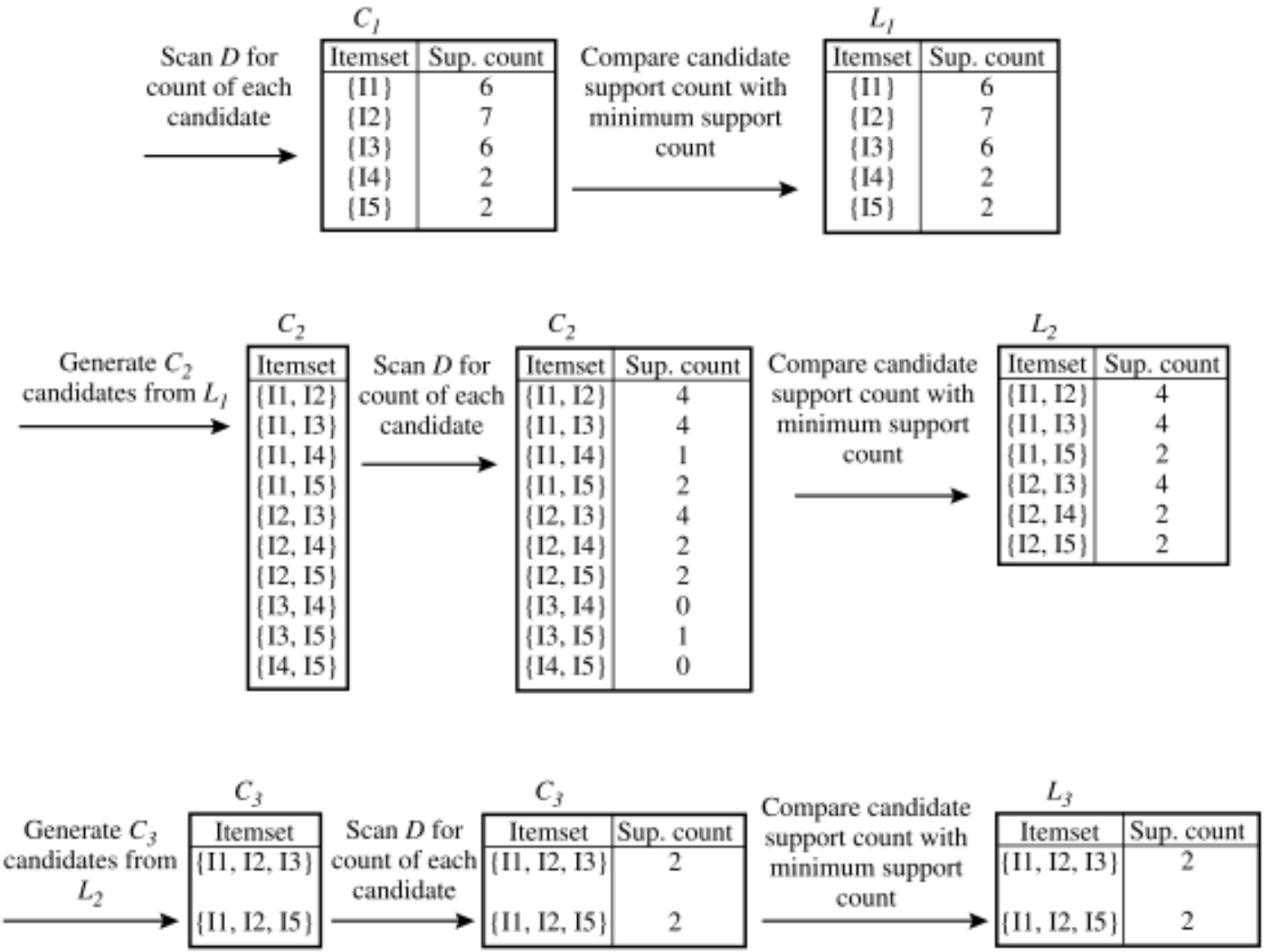
Regras de Associação

Um exemplo

No conjunto de dados há nove transações, $|D| = 9$. A figura do próximo slides mostra como o Apriori encontra os itemset frequentes em D. Suponha $min\_sup = 2$.

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

Regras de Associação



Regras de Associação

(a) Join: $C_3 = L_2 \bowtie L_2 = \{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}\} \bowtie$
$\{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}\}$
$= \{\{I1, I2, I3\}, \{I1, I2, I5\}, \{I1, I3, I5\}, \{I2, I3, I4\}, \{I2, I3, I5\}, \{I2, I4, I5\}\}.$

(b) Prune using the Apriori property: All nonempty subsets of a frequent itemset must also be frequent. Do any of the candidates have a subset that is not frequent?

- The 2-item subsets of $\{I1, I2, I3\}$ are $\{I1, I2\}$, $\{I1, I3\}$, and $\{I2, I3\}$. All 2-item subsets of $\{I1, I2, I3\}$ are members of L_2. Therefore, keep $\{I1, I2, I3\}$ in C_3.
- The 2-item subsets of $\{I1, I2, I5\}$ are $\{I1, I2\}$, $\{I1, I5\}$, and $\{I2, I5\}$. All 2-item subsets of $\{I1, I2, I5\}$ are members of L_2. Therefore, keep $\{I1, I2, I5\}$ in C_3.
- The 2-item subsets of $\{I1, I3, I5\}$ are $\{I1, I3\}$, $\{I1, I5\}$, and $\{I3, I5\}$. $\{I3, I5\}$ is not a member of L_2, and so it is not frequent. Therefore, remove $\{I1, I3, I5\}$ from C_3.
- The 2-item subsets of $\{I2, I3, I4\}$ are $\{I2, I3\}$, $\{I2, I4\}$, and $\{I3, I4\}$. $\{I3, I4\}$ is not a member of L_2, and so it is not frequent. Therefore, remove $\{I2, I3, I4\}$ from C_3.
- The 2-item subsets of $\{I2, I3, I5\}$ are $\{I2, I3\}$, $\{I2, I5\}$, and $\{I3, I5\}$. $\{I3, I5\}$ is not a member of L_2, and so it is not frequent. Therefore, remove $\{I2, I3, I5\}$ from C_3.
- The 2-item subsets of $\{I2, I4, I5\}$ are $\{I2, I4\}$, $\{I2, I5\}$, and $\{I4, I5\}$. $\{I4, I5\}$ is not a member of L_2, and so it is not frequent. Therefore, remove $\{I2, I4, I5\}$ from C_3.

(c) Therefore, $C_3 = \{\{I1, I2, I3\}, \{I1, I2, I5\}\}$ after pruning.

Regras de Associação

Algorithm: Apriori. Find frequent itemsets using an iterative level-wise approach based on candidate generation.

Input:

- D, a database of transactions;
- *min_sup*, the minimum support count threshold.

Output: L, frequent itemsets in D.

Regras de Associação

Method:

(1) L_1 = find_frequent_1-itemsets(D);
(2) for ($k = 2; L_{k-1} \neq \phi; k++$) {
(3) C_k = apriori_gen(L_{k-1});
(4) for each transaction $t \in D$ { // scan D for counts
(5) C_t = subset(C_k, t); // get the subsets of t that are candidates
(6) for each candidate $c \in C_t$
(7) c.count++;
(8) }
(9) $L_k = \{c \in C_k | c.count \geq min\_sup\}$
(10) }
(11) return $L = \cup_k L_k$;

Regras de Associação

```
procedure apriori_gen(L_{k-1}:frequent (k-1)-itemsets)
(1)     for each itemset l_1 ∈ L_{k-1}
(2)         for each itemset l_2 ∈ L_{k-1}
(3)             if (l_1[1] = l_2[1]) ∧ (l_1[2] = l_2[2]) ∧ ... ∧ (l_1[k-2] = l_2[k-2]) ∧ (l_1[k-1] < l_2[k-1]) then {
(4)                 c = l_1 ⋈ l_2; // join step: generate candidates
(5)                 if has_infrequent_subset(c, L_{k-1}) then
(6)                     delete c; // prune step: remove unfruitful candidate
(7)                 else add c to C_k;
(8)             }
(9)     return C_k;
```

Regras de Associação

```
procedure has_infrequent_subset(c: candidate k-itemset;
                L_{k-1}: frequent (k-1)-itemsets); // use prior knowledge
(1)     for each (k-1)-subset s of c
(2)         if s ∉ L_{k-1} then
(3)             return TRUE;
(4)     return FALSE;
```

Regras de Associação

Gerando as regras

- para cada itemset frequente l, gere todos os subconjuntos não vazios de l;
- para todo conjunto s não vazio de l, crie a regras $s \Rightarrow (l - s)$ onde $\frac{suporte(l)}{suporte(s)} \geq min\_conf$.

Para o itemset frequente $\{I1, I2, I5\}$ ...

$$
\begin{aligned}
I1 \wedge I2 &\Rightarrow I5, & confidence &= 2/4 = 50\% \\
I1 \wedge I5 &\Rightarrow I2, & confidence &= 2/2 = 100\% \\
I2 \wedge I5 &\Rightarrow I1, & confidence &= 2/2 = 100\% \\
I1 &\Rightarrow I2 \wedge I5, & confidence &= 2/6 = 33\% \\
I2 &\Rightarrow I1 \wedge I5, & confidence &= 2/7 = 29\% \\
I5 &\Rightarrow I1 \wedge I2, & confidence &= 2/2 = 100\%
\end{aligned}
$$

Regras de Associação

Classificação para padrões frequentes

- Baseado na completude dos padrões a serem minerados
- Baseado nos níveis de abstração envolvidos no conjunto de regras
- Baseado no número de dimensões dos dados envolvidos na regras
- Baseado nos tipos de valores manuseados nas regras
- Baseado nos tipos de regras as serem mineradas
- Baseado nos tipos de padrões a serem minerados

Regras de Associação

Baseado na completude dos padrões a serem minerados

É possível minerar o conjunto completo de itemsets frequentes, os itemsets frequentes fechados e os itemsets frequentes máximos, dado um suporte mínimo. Além disso é possível minerar:

- **itemset frequentes restritos:** aqueles que satisfazem um conjunto de restrições definidas pelo usuário;
- **itemsets frequentes aproximados:** aqueles que possuem suporte aproximado ao mínimo;
- **itemsets frequentes near-match:** aqueles que quase alcançam o suporte mínimo;
- **top k-itemsets frequentes:** os k itemsets mais frequentes de acordo com um valor de k pré-determinado.

Regras de Associação

Baseado nos níveis de abstração envolvidos no conjunto de regras

Suponha que um conjunto de regras de associação inclua as seguintes regras, onde X é uma variável que representa um cliente:

$$buys(X, \text{“computer”}) \Rightarrow buys(X, \text{“HP-printer”})$$
$$buys(X, \text{“laptop-computer”}) \Rightarrow buys(X, \text{“HP-printer”})$$

Nessas regras os itens comprados possuem diferentes níveis de abstração (*“computer”* tem um nível de abstração mais alto do que *“laptop-computer”*). Essas regras são ditas **regras de associação multinível**. Se todos os itens referenciados no conjunto de regras forem do mesmo nível de abstração, elas serão ditas **regras de associação de nível único**.

Regras de Associação

Baseado no número de dimensões dos dados envolvidos na regras

Se itens ou atributos em uma regra de associação possuem uma única dimensão, as regras são ditas **regras de associação de dimensão única**. Caso contrário, são ditas **regras de associação multidimensionais**. Elas podem ser respectivamentes exemplificadas como:

$$buys(X, \text{“}computer\text{”}) \Rightarrow buys(X, \text{“}antivirus\text{-}software\text{”})$$

$$age(X, \text{“}30...39\text{”}) \wedge income(X, \text{“}42K...48K\text{”}) \Rightarrow buys(X, \text{“}high\text{-}resolution\ TV\text{”})$$

No caso da última regra, as dimensões são: *age*, *income* e *buys*.

Regras de Associação

Baseado nos tipos de valores manuseados nas regras

Se a regra envolve associações entre a presença ou a ausência de itens, ela é uma **regra de associação booleana**.

Se uma regra descreve associações entre itens ou atributos quantitativos, então ela é uma **regra de associação quantitativa**. Nestas regras, valores quantitativos para itens ou atributos são particionados em intervalos. No caso da última regra do slide anterior, os atributos quantitativos *age* e *income* foram discretizados.

Regras de Associação

Baseado nos tipos de regras as serem mineradas

Regras de associação são as regras mais comuns em mineração de dados. Contudo, a descoberta de associações pode ser aprofundada por meio da descoberta de correlações estatísticas, levando a **regras de correlação**.

Ainda é possível minerar **strong gradient relationships** entre itemsets, onde o gradiente é o raio da medida de um itemset quando comparado com a medida de seus pais (um itemset generalizado), seu filho (um itemset especializado) ou seu irmão (um itemset comparável). Por exemplo

A média de vendas da câmera digital Sony aumenta em 16% *quando vendida junto com o computador laptop Sony*.

Câmera e computadores são itens irmãos e Sony é um item pai.

Regras de Associação

Baseado nos tipos de padrões a serem minerados

Além de mineração de itens frequentes de um banco de dados transacional, ainda é possível minerar:

- **Padrões sequenciais**: busca por susequências frequentes em um conjunto de dados sequencial, onde uma sequência registra uma ordem de eventos. Por exemplo, estudar a ordem na qual itens são frequentemente comprados:

  Clientes tendem a comprar primeiro um PC, e depois uma câmera digital, e só então um cartão de memória.

- **Padrões estruturados**: busca por subestruturas (grafos, latices, árvores, sequências, conjuntos, itens únicos ou combinações de tais estruturas) frequentes em um conjunto de dados estruturado. Trata-se de um caso mais geral de mineração de padrões frequentes.

Regras de Associação

Minerando regras de associação multinível

Para algumas aplicações pode ser difícil encontrar regras de associação no nível mais baixo de abstração. Isso ocorre por conta da esparsidade dos dados nos níveis mais baixos.

Regras de associação descobertas em níveis mais altos de abstração representam conhecimento de senso comum. Entretanto, o que pode ser de senso comum para um usuário pode não ser para outro.

Sistemas de mineração tem o objetivo de fornecer condições para descoberta de regras de associação de múltiplos níveis de abstração, com flexibilidade suficiente para transitar em diferentes espaços de abstração.

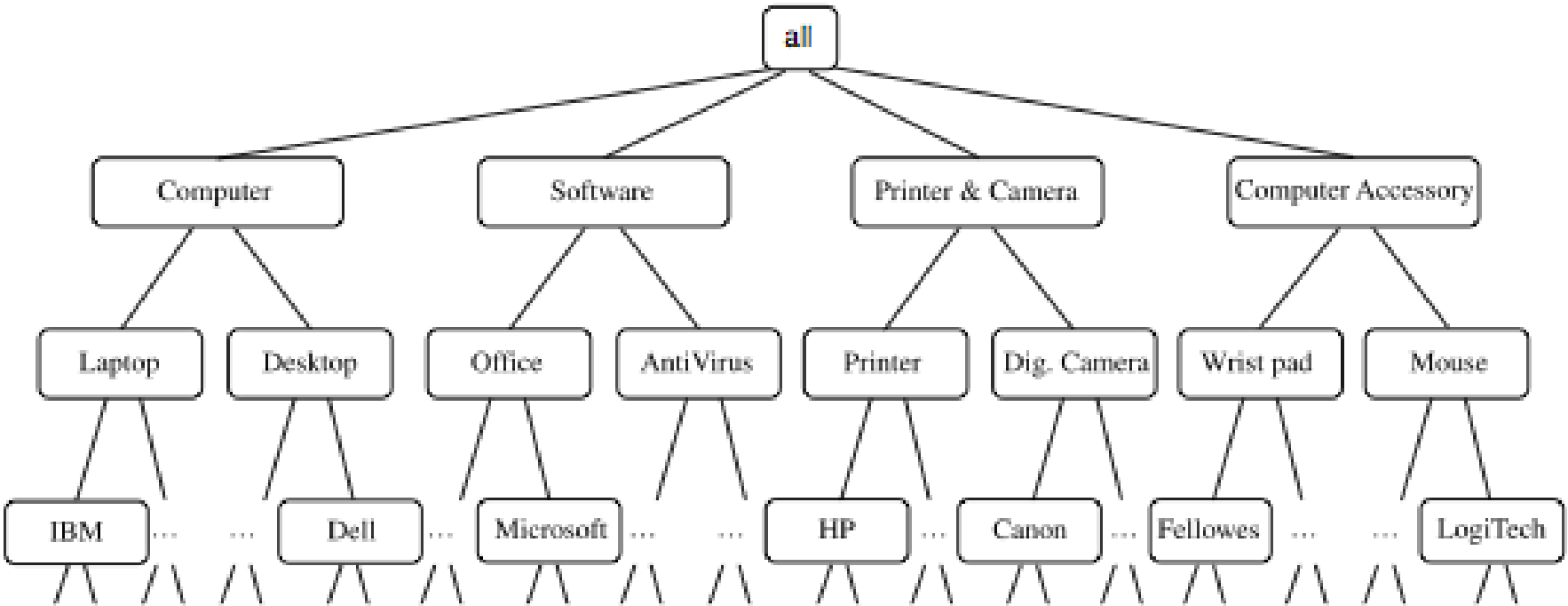
Regras de Associação

Exemplo

Considere a tabela abaixo, onde os itens comprados em cada uma das transações são mostrados. Considere também uma hierarquia de conceitos (próximo slide). O conceito de hierarquia define uma sequência de mapeamentos de um conjunto de conceitos de baixo nível para conceitos de alto nível.

TID	*Items Purchased*
T100	IBM-ThinkPad-T40/2373, HP-Photosmart-7660
T200	Microsoft-Office-Professional-2003, Microsoft-Plus!-Digital-Media
T300	Logitech-MX700-Cordless-Mouse, Fellowes-Wrist-Rest
T400	Dell-Dimension-XPS, Canon-PowerShot-S400
T500	IBM-ThinkPad-R40/P4M, Symantec-Norton-Antivirus-2003
...	...

Regras de Associação



Regras de Associação

Exemplo

Os itens na tabela (slide anterior) são os conceitos em nível mais baixo na hierarquia. Observe como deve ser difícil encontrar padrões de compras interessantes neste nível (tão baixo) de abstração. Por exemplo: se "*IBM-ThinkPad-R40/P4M*" ou "*Symantec-Norton-Antivirus-2003*" ocorrem, cada um, em poucas transações, pode ser difícil encontrar regras de associação fortes envolvendo estes itens específicos. Poucas pessoas compram estes itens juntos. Contudo, poder-se-ia esperar que regras de associações fortes poderiam ser encontradas no nível de abstração "*IBM laptop computer*" e "*"antivirus software"*.

Regras de Associação

Minerando regras de associação multinível

Regras de associação multinível podem ser eficientemente mineradas usando hierarquias de conceitos e uma estrutura de suporte-confiança.

Geralmente, uma estratégia *top-down* é aplicada, na qual as contagens são acumuladas para o cálculo dos itemsets frequentes para cada nível de conceito, iniciando no nível 1 e terminando no nível mais específico, até que nenhum item frequente possa ser encontrado.

Qualquer algoritmo para contagem de itemsets frequentes pode ser usado nesta estratégia, incluindo o Apriori. Algumas variações, no entanto, podem ser aplicadas.

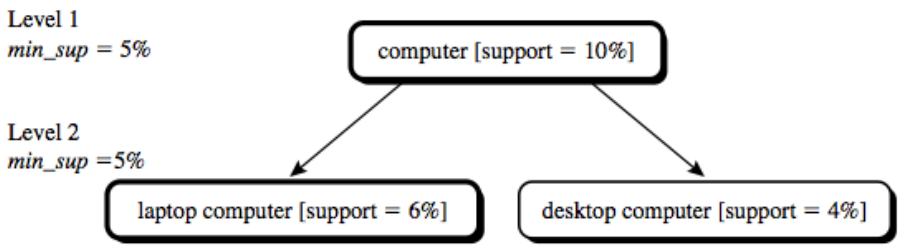
- usando suporte mínimo uniforme para todos os níveis;
- usando suporte mínimo reduzido nos níveis mais baixos;
- usando suporte mínimo baseado em grupo ou item.

Regras de Associação

Usando suporte mínimo uniforme para todos os níveis

O mesmo limiar para suporte mínimo é usado na mineração realizada em cada nível de abstração. Veja a figura abaixo. O limiar de suporte mínimo de 5% é aplicado. *Computer* e *laptop computer* são frequentes, enquanto *desktop computer* não.

Quando um limiar uniforme é usado, a busca é mais simples. Os usuários só precisam especificar um suporte mínimo, e o Apriori é diretamente aplicável baseando-se no conhecimento de que um nível mais alto é um superconjunto de seus descendentes: a busca evita examinar itemsets que contêm qualquer item cujo nível mais alto (ancestral) não tem um suporte mínimo.



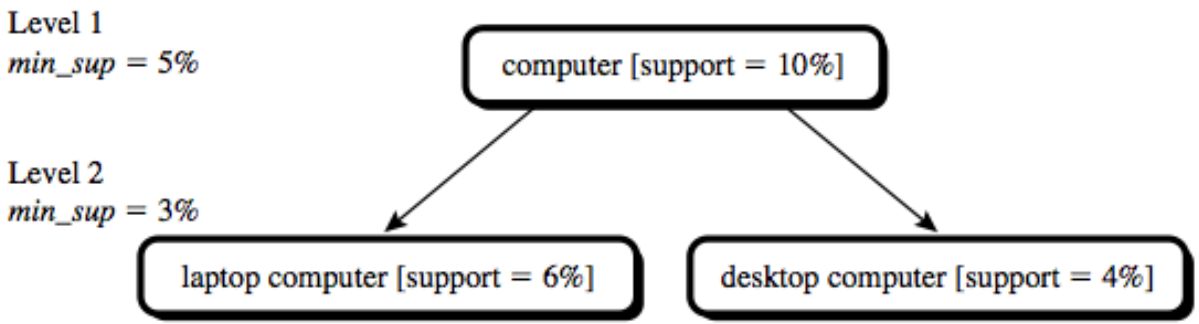
Desvantagem

Se um suporte mínimo muito alto é definido, a busca poderia perder algumas associações interessantes que ocorrem em níveis mais baixos de abstração. Se ele é muito baixo, pode gerar regras que não são interessantes nos níveis mais altos de abstração.

Regras de Associação

Usando suporte mínimo reduzido nos níveis mais baixos

Cada nível de abstração tem o seu próprio suporte mínimo. Quanto mais profundo, menor é o suporte mínimo. Veja o exemplo da figura.



Usando suporte mínimo baseado em grupo ou item

Quando se tem conhecimento da importância de **grupos**, é possível determinar suportes mínimos específicos para usuários, itens ou grupos. Por exemplo: um usuário poderia setar um suporte mínimo baseado no preço do produto, ou no item de interesse, de forma que ele pode se ater a regras de associação que contenham estas categorias.

Regras de Associação

Minerando regras de associação multinível

Note que nas duas últimas formas apresentadas, a propriedade Apriori pode não permanecer válida. Além disso, regras redundantes podem ser geradas.

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

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

Ambas as regras são úteis? A regra menos geral oferece informação nova?

Se considerarmos que 1/4 dos computadores laptops vendidos nessa loja são da IBM, o que podemos dizer sobre a utilidade da segunda regra?

Regras de Associação

Mineração de Associação X Análise de Correlação

Frequentemente, muitas regras interessantes podem ser encontradas usando um limiar baixo para o suporte mínimo. Isso é possível porque o limiar de confiança da regra ajuda a avaliá-la melhor. Mas, mesmo assim, acontece de serem geradas regras que não são interessantes aos usuários.

Uma regra pode ser dita interessante mediante análises objetivas ou subjetivas. As análises subjetivas são feitas pelos próprios usuários, e podem diferir para usuários diferentes. Medidas objetivas (além do suporte e confiança) podem ser baseadas em estatísticas (significância estatística e análise de correlação).

Regras de Associação

Um exemplo: uma regra de associação "forte" ilusória.

Suponha uma análise de transações de vendas relacionadas à compra de jogos de computadores e vídeos. Considere que "jogo" se refere a transações contendo jogos de computadores, e "video" se refere a transações contendo vídeos. De 10.000 transações analisadas, os dados revelam que 6.000 incluem jogos de computadores, enquanto 7.500 incluem vídeos, e 4.000 incluem ambos. Suponha ainda que um programa de mineração de dados para descobrir regras tenha sido executado, usando um suporte mínimo de 30% e uma confiança mínima de 60%. Neste cenário, a seguinte regra de associação é descoberta:

$$buys(X, \text{“computer games”}) \Rightarrow buys(X, \text{“videos”}) \; [support = 40\%, \; confidence = 66\%]$$

A regra é forte e portanto seria retornada pelo programa, já que seu suporte $\frac{4.000}{10.000}$ e confiança $\frac{4.000}{6.000}$ satisfazem às medidas de interessabilidade definidas.

Regras de Associação

...

Contudo, a interessabilidade da regra é ilusória porque a probabilidade de comprar vídeos é de 75%, o que é maior do que 66%. De fato, jogos de computadores e vídeos são negativamente associados, já que a compra de um dos itens (no caso, jogos) diminui a probabilidade de compra do outro (no caso, vídeos).

Ou seja, o contexto mostra que se um jogo foi comprado, existirá menos chance (só 66%) de um vídeo ser comprado – o que fazer então a partir da descoberta dessa regra? Ela é de fato interessante?

A confiança da regra não mede, de fato, a força da regra. Ela mede a probabilidade condicional de um item dado um outro item (ou conjuntos de itens). Ela não é capaz de medir a correlação ou implicação existente entre os itens (ou conjuntos de itens).

O que é interessante: minerar relacionamentos interessantes entre os dados.

Regras de Associação

A medida de correlação pode ser usada para melhorar o framework *suporte/confiança* de regras de associação. Isso gera *regras de correlação* da forma:

$$A \Rightarrow B \ [support, \ confidence.correlation]$$

Isto é, a regra de correlação é medida não somente por seu suporte e confiança mas também pela correlação entre os itemsets A e B.

Regras de Associação

Lift

Lift é uma medida de correlação simples the funciona da seguinte forma. O ocorrência de um itemset A é **independente** da ocorrência de um itemset B se $P(A \cup B) = P(A)P(B)$; caso contrário, os itemsets A e B são dependentes ou correlatos (ou correlacionados) como eventos. Assim,

$$lift(A, B) = \frac{conf(A \Rightarrow B)}{sup(B)} = \frac{sup(AB)}{sup(A) * sup(B)}$$

Se o valor resultante é menor do que 1, então a ocorrência de A é *negativamente correlacionada com* a ocorrência de B. Se o valor resultante é maior do que 1, então A e B são *positivamente correlacionados*, significando que a ocorrência de um implica na ocorrência de outro. Se o resultado é igual a 1, então A e B são *independentes* e não há correlação entre eles.

Regras de Associação

Exemplo ....

Considerando os dados do último exemplo. Seja $\overline{jogos}$ as transações que não contém jogos de computadores, e $\overline{video}$ aquelas que não contém vídeos. As transações podem ser resumidas em uma tabela de contigência (veja abaixo).

A 2×2 contingency table summarizing the transactions with respect to game and video purchases.

	game	$\overline{game}$	Σ_{row}
video	4,000	3,500	7,500
$\overline{video}$	2,000	500	2,500
Σ_{col}	6,000	4,000	10,000

Da tabela, nós podemos ver que a probabilidade de comprar um jogo de computador é $P(game) = 0.60$, a probabilidade de comprar um video é $P(video) = 0.75$, e a probabilidade de comprar ambos é $P(game, video) = 0.40$. A medida *lift* da regra já citada é $P(game, video)/P(game)P(video) = 0.40/(0.60 * 0.75) = 0.89$.

Como o valor do *lift* é menor do que 1 existe uma correlação negativa entre a ocorrência de jogos e vídeos. O numerador é a probabilidade do cliente comprar ambos, e o numerador é a probabildiade que teria valido se duas compras tivessem sido feitas independentemente.

Regras de Associação

- Sarajane M. Peres - sarajane@usp.br
- Clodoaldo A. M. Lima - c.lima@usp.br

Escola de Artes, Ciências e Humanidades - EACH
Universidade de São Paulo - USP