



Aula 6

Análise de Associações

Introdução

- Há uma classe de modelos que podem ser aplicados há um **conjunto de dados de interesse que não incluem um atributo alvo**
- Esses modelos são chamados de **modelos descritivos** e são usualmente gerados por meio de **aprendizado não supervisionado**
- São bem simples, intuitivos e são utilizados para **descobrir regularidades ou associações entre os dados**
- Um modelo descritivo do conjunto de dados pode ser obtido por meio da **Análise de Associações**

Introdução

- O objetivo da análise de associações é **encontrar recorrências entre itens** de um grande conjunto de transações
- **A descoberta de relações entre itens em grandes quantidades de transações de negócios podem ajudar em muitos processos de tomada de decisões**
 - Análise de comportamento do consumidor
 - Recomendação de itens
 - Definição de *layouts*
 - Desenvolvimento de estratégia de *marketing*
 - Definir descontos de alguns produtos para encorajar o consumidor a levar outros produtos
 - ...

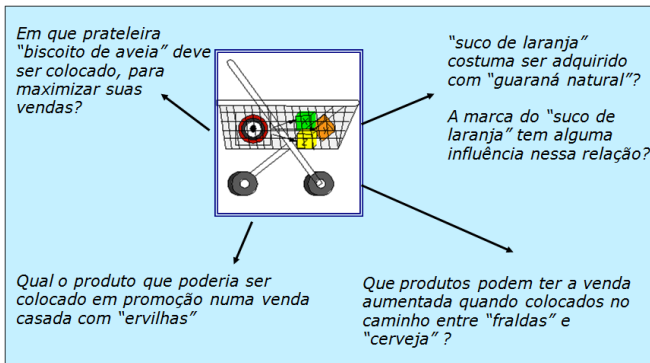
Análise de Cestas de Mercado

- A área de pesquisa relacionada com a análise de associação teve seu início na tentativa de identificar itens relacionados nas compras de clientes em mercados (*Market Basket Analysis*)
- Quando um consumidor compra um **conjunto de itens** em algum posto de venda, todos os itens da compra são armazenados
- Esse conjunto de itens é chamado de **transação**
- **OBSERVAÇÃO:** normalmente são armazenados a hora, os itens e seus respectivos preços dos itens comprados (para o método básico, apenas os itens são necessários)

Análise de Cestas de Mercado

- A análise de cestas de mercados permite identificar vendas recorrente relacionadas à:
 - Compra de um produto
 - Compra de grupos de produtos
 - Se a compra de um produto aumenta as chances de outro produto ser vendido
- Com isso, pode-se obter regra do tipo *“Em 68% das vezes que alguém compra cereal também compra leite”* ou *“se um cliente compra cereal, ele compra leite em 68% das vezes”*
- **Principais usos:**
 - Planejar iniciativas promocionais
 - Definir o local de produtos nas prateleiras

Análise de Cestas de Mercado



<http://www.devmedia.com.br/>

mineracao-de-dados-com-market-basket-analysis-revista-sql-magazine-111/27853

Análise de Cestas de Mercado

- O caso clássico da literatura é o da rede de supermercados WalMart
 - Foi descoberta a relação de compra entre uma marca de fraldas e uma marca de cerveja (para compras realizadas por homens na sextas-feiras ao fim do dia)
 - Esposas pediam para os maridos comprarem fraldas para o fim de semana (quando saíssem do trabalho)
 - Os maridos aproveitavam e levavam a cerveja para relaxar durante o fim de semana
 - O gerente decidiu colocar as fraldas perto das cervejas → vendas aumentaram ainda mais

Mineração *Web*

- A mineração de dados da *web* é particularmente útil para entender os **padrões de navegação** e a **frequência com que grupos de páginas web são visitadas** (em uma única sessão ou sessões consecutivas)
- Neste caso, a lista de páginas visitadas é armazenada como transação (pode-se armazenar também o número da sequência e o tempo e duração da visita)

Mineração Web

- Pode-se gerar regras do tipo *“Se um indivíduo visita o site timesonline.co.uk, então dentro de uma semana ele também visitará o site economist.com com probabilidade de 87%”*
- Regras de associação do tipo acima podem influenciar a estrutura de *links* entre páginas de maneira a:
 - Facilitar a navegação
 - Recomendar páginas ou sequências de páginas específicas
 - Colocar propagandas ou mensagens promocionais específicas

Mineração Web

[illegible]

Características

- Porém, as regras extraída para um processo de apoio à decisão devem ser **interpretáveis** e **não triviais**
- **Regras interpretáveis**
 - Regras representam um paradigma clássico para representação de conhecimento (**paradigma simbólico**)
 - São populares devido à sua estrutura **simples e intuitiva**
 - $A \implies B$
 - $A, B \implies C$
 - ...

Características

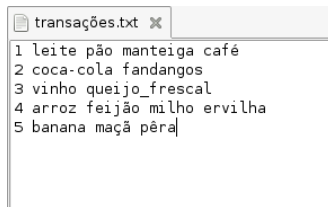
- **Regras triviais**

- Regras que não produzem conhecimento óbvio
 - *“Quem compra café também compra filtro de café”*
 - Ou ainda regras que refletem efeitos de uma campanha promocional

- Além disso, deve-se tomar cuidado com regras que invertem a causa e o efeito
 - *“Compradores de seguro de carro também compram carro com probabilidade de 100%”*

Transações

- **As transações podem estar armazenadas em diferentes formatos**



```
transações.txt x
1 leite pão manteiga café
2 coca-cola fandangos
3 vinho queijo_frescal
4 arroz feijão milho ervilha
5 banana maçã pêra|
```

Exemplo de um arquivo de transações no qual os itens de uma transação estão disposto de maneira sequencial

Transações

- As transações podem estar armazenadas em diferentes formatos

case ID	attribute
TRANS_ID	ITEM_ID
11	B
11	D
11	E
12	A
12	B
12	C
12	E
13	B
13	C
13	D
13	E

Exemplo de um arquivo de transações no qual cada item de uma transação corresponde a uma entrada na base de dados

Transações

- As transações podem estar armazenadas em diferentes formatos

transaction ID	milk	bread	butter	beer
1	1	1	0	0
2	0	0	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Exemplo de um arquivo de transações dispostos em uma tabela na qual cada coluna corresponde a um item

http://en.wikipedia.org/wiki/Association_rule_learning

Conceitos Básicos e Notações

- $\mathcal{I} = \{I_1, I_2, \dots, I_m\}$: **itemset** (ou conjunto de itens)
- T : **transação**, que corresponde a um *itemset* não vazio tal que $T \subseteq \mathcal{I}$
- **Regra de Associação**: uma implicação da forma $A \Rightarrow B$, na qual $A \subset \mathcal{I}$, $B \subset \mathcal{I}$, $A \neq \emptyset$, $B \neq \emptyset$, e $A \cap B = \emptyset$

Conceitos Básicos e Notações

- **Suporte Relativo:**

$$\text{sup}(A \Rightarrow B) = P(A \cup B) = \frac{\text{cont}(A \& B)}{|T|}$$

na qual $P(A \cup B)$ é a probabilidade de ocorrer o conjunto de itens item em A e B no conjunto de transações T , $\text{cont}(A \& B)$ é a frequência com que os itens de A e B apareceram nas mesmas transações e $|T|$ é o número total de transações

Conceitos Básicos e Notações

- **Confiança:**

$$\text{conf}(A \Rightarrow B) = P(B|A) = \frac{\text{sup}(A \Rightarrow B)}{\text{sup}(A)} = \frac{P(A \cup B)}{P(A)} = \frac{\text{cont}(A \& B)}{\text{cont}(A)}$$

na qual $P(A)$ é a probabilidade de ocorrer o conjunto de itens em A no conjunto total de transações, i.e.,

$$P(A) = \text{cont}(A) / |T|$$

Conceitos Básicos e Notações

- **Em geral, a mineração de regras de associação pode ser dividida em dois passos**
 - 1 **Encontrar todos os itemsets frequentes:** cada um dos *itemsets* irá ocorrer com frequência igual ou superior ao limiar determinado pelo usuário (suporte mínimo)
 - 2 **Gerar regras de associação fortes a partir dos itemsets frequentes:** gerar regras usando os *itemsets* frequentes extraídos no passo anterior de forma que a confiança das regras seja maior que um limiar de confiança definido pelo usuário (confiança mínima)

Conceitos Básicos e Notações

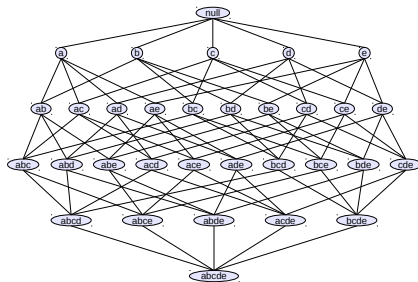
- O segundo passo é muito menos custoso computacionalmente, portanto a custo computacional para a extração de regras é determinado pelo primeiro passo
- Tipicamente as regras de associação são consideradas fortes ou interessante se estas satisfazem ambos o suporte e confiança mínimos
- O valores de suporte e confiança mínimos são definidos pelo usuário ou especialistas de domínio

Algoritmo APRIORI - Geração de *Itemsets* Frequentes

- O algoritmo Apriori [Agrawal and Srikant, 1994] é um dos algoritmos utilizados para a extração dos *itemsets* frequentes que são utilizados para gerar as regras de associação
- **É executado um processo iterativo no qual k -itemsets são utilizados para gerar $(k + 1)$ -itemsets**
 - Primeiramente, o conjunto de 1-itemset frequente é gerado (L1)
 - A seguir, L1 é utilizado para gerar o conjunto de 2-itemsets frequentes (L2)
 - L2 é utilizado para gerar L3 e assim por diante, até que nenhum k -itemset seja encontrado

Algoritmo APRIORI - Geração de *Itemsets* Frequentes

- Em geral, um conjunto de dados que contém k itens pode gerar potencialmente $2^k - 1$ *itemsets* frequentes
- Um *lattice* pode ser utilizado para listar todos os possíveis *itemsets*

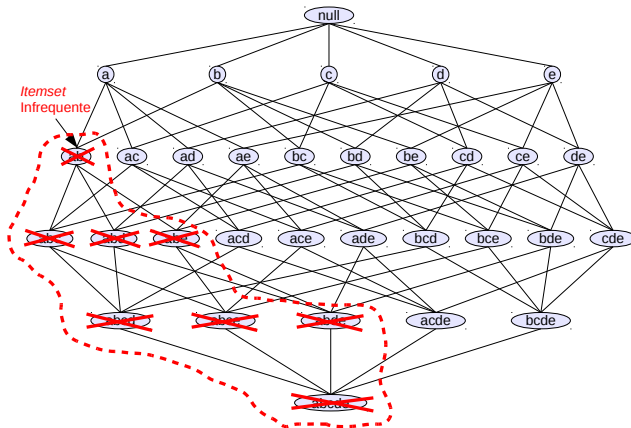


Lattice dos conjuntos de *itemsets* gerados por $\mathcal{I} = \{a, b, c, d, e\}$

Algoritmo APRIORI - Geração de *Itemsets* Frequentes

- Portanto, para um conjunto de dados com 100 itens distintos
→ 1.26×10^{30} possibilidades de itemsets seriam analisados
- Para melhorar a eficiência da geração do *itemsets* frequentes em cada nível, uma propriedade importante é utilizada para reduzir o espaço de busca é utilizada no algoritmo APRIORI
- **Propriedade do Apriori**
 - “*Todo subconjunto não vazio de um itemset frequente também é frequente*”
 - Se um *itemset* I não satisfaz o suporte mínimo, a adição de um *itemset* A em I não fará com que $(A \cup I)$ seja frequente

Algoritmo APRIORI - Geração de *Itemsets* Frequentes



Redução do espaço de busca pela utilização da propriedade do Apriori

Algoritmo APRIORI - Geração de *Itemsets* Frequentes

- Para verificar as propriedades anteriores, vamos considerar o seguinte conjunto de transações

PILHA	MANTEIGA	PÃO
1	1	1
0	1	1
0	1	1
0	1	0
0	0	1

Algoritmo APRIORI - Geração de *Itemsets* Frequentes

- Se um *itemset* é frequente, todos os possíveis subconjuntos de itens desse *itemset* também serão frequentes já que o suporte destes será no mínimo igual ao suporte do *itemset* completo

PILHA	MANTEIGA	PÃO
1	1	1
0	1	1
0	1	1
0	1	0
0	0	1



PILHA	MANTEIGA	PÃO
1	1	1
0	1	1
0	1	1
0	1	0
0	0	1

Algoritmo APRIORI - Geração de *Itemsets* Frequentes

- Se um *itemset* é infrequente, todos os possíveis *itemsets* derivados a partir deste *itemset* também serão infrequentes já que o suporte destes será no máximo o suporte do *itemset* de menor tamanho

PILHA	MANTEIGA	PÃO
1	1	1
0	1	1
0	1	1
0	1	0
0	0	1



PILHA	MANTEIGA	PÃO
1	1	1
0	1	1
0	1	1
0	1	0
0	0	1

Algoritmo APRIORI - Geração de *Itemsets* Frequentes

- A princípio, existem várias formas de gerar os candidatos a *itemsets* → número de combinações pode ser muito grande
- Os requisitos para a geração dos *itemsets* frequentes são
 - 1 Não se deve-se gerar muitos candidatos desnecessários (um candidato é desnecessário se ao menos um de seus subconjuntos é infrequente)
 - 2 Deve-se assegurar que o conjunto de candidatos é completo (nenhum itemset frequente é excluído da geração de candidatos)
 - 3 Não se deve gerar o mesmo itemset candidato mais de uma vez

Algoritmo APRIORI - Geração de *Itemsets* Frequentes

- **Método $F_{k-1} \times F_1$**

- Estender cada $(k - 1)$ -*itemset* com os 1-*itemsets* frequentes
- $\{\text{Cerveja, Fralda}\} \cup \{\text{Pão}\}$ formam o *itemset* candidato $\{\text{Cerveja, Fralda, Pão}\}$
- O método é completo pois cada k -*itemset* é composto de um $(k - 1)$ -*itemset* e um 1-*itemset* frequente
- Não prevê que o mesmo *itemset* candidato não seja gerado mais de uma vez

Exemplo: $\{\text{Cerveja, Fralda, Leite}\}$ pode ser gerado pelas junções de

- $\{\text{Cerveja, Fralda}\} \cup \{\text{Leite}\}$
- $\{\text{Cerveja, Leite}\} \cup \{\text{Fralda}\}$
- $\{\text{Fralda, Leite}\} \cup \{\text{Cerveja}\}$

Algoritmo APRIORI - Geração de *Itemsets* Frequentes

- **Método $F_{k-1} \times F_1$**

- Uma forma de evitar a geração de candidatos duplicados é mantendo uma ordem lexicográfica dos *itemsets* frequentes
 - Cada $(k - 1)$ -*itemset* X é então estendido com itens frequentes que são lexicograficamente maiores que X
 - {Cerveja, Fralda} pode ser aumentado com {Leite}
 - {Fralda, Pão} não pode ser aumentado com {Cerveja}
- Ainda pode haver geração de candidatos desnecessário
 - Supondo que temos os candidatos {Cerveja, Fralda} e {Leite}
 - Será gerado o *itemset* {Cerveja, Fralda, Leite}
 - Porém, pode ser que {Cerveja, Leite} não seja frequente, e portanto {Cerveja, Fralda, Leite} também não será

Algoritmo APRIORI - Geração de *Itemsets* Frequentes

- **Método** $F_{k-1} \times F_{k-1}$

- A geração de um *itemset* candidato é feita pela junção de dois $(k - 1)$ -*itemsets* somente se os primeiros $(k - 2)$ -*itemsets* são idênticos
- Por exemplo, os *itemsets* {Cerveja, Fralda} e {Cerveja, Leite} são unidos para formar o *itemset* candidato {Cerveja, Fralda, Leite}
- Por exemplo, não é necessário unir os *itemsets* {Cerveja, Fralda} com {Fralda, Leite} para formar o *itemset* {Cerveja, Fralda, Leite} pois esse já foi obtido pela junção dos *itemsets* {Cerveja, Fralda} e {Cerveja, Leite}
- Pode ainda gerar candidatos desnecessários em menor volume
→ ex: {Fralda, Leite} pode ser infrequente

- Uma regra de associação pode ser extraída por particionar o *itemset* Y em dois subconjuntos não vazios, X e $Y - X$, tal que $X \Rightarrow Y - X$ satisfaz a confiança mínima
- Seja $Y = \{a, b, c\}$, pode-se gerar as seguintes regras
 - $\{a, b\} \Rightarrow \{c\}$
 - $\{a, c\} \Rightarrow \{b\}$
 - $\{b, c\} \Rightarrow \{a\}$
 - $\{a\} \Rightarrow \{b, c\}$
 - $\{b\} \Rightarrow \{a, c\}$
 - $\{c\} \Rightarrow \{a, b\}$
- Cada k -*itemset* frequente Y pode produzir $2^k - 2$ regras de associação (ignorando regras com $\emptyset \Rightarrow Y$ e $Y \Rightarrow \emptyset$)
- Computar a confiança de uma regra de associação não requer verificações adicionais no conjunto de transações

Exemplo

Extrair regras com $\text{sup-mín} = 30\%$ e $\text{conf-mín} = 70\%$

T	Leite	Café	Cerveja	Pão	Manteiga	Arroz	Feijão
1	N	S	N	S	S	N	N
2	S	N	S	S	S	N	N
3	N	S	N	S	S	N	N
4	S	S	N	S	S	N	N
5	N	N	S	N	N	N	N
6	N	N	N	N	S	N	N
7	N	N	N	S	N	N	N
8	N	N	N	N	N	N	S
9	N	N	N	N	N	S	S
10	N	N	N	N	N	S	N

Exemplo

- Calculando o suporte dos 1-*itemsets*
 - $\text{sup}(\{\text{Leite}\}) = 20\%$
 - $\text{sup}(\{\text{Café}\}) = 30\%$
 - $\text{sup}(\{\text{Cerveja}\}) = 20\%$
 - $\text{sup}(\{\text{Pão}\}) = 50\%$
 - $\text{sup}(\{\text{Manteiga}\}) = 50\%$
 - $\text{sup}(\{\text{Arroz}\}) = 20\%$
 - $\text{sup}(\{\text{Feijão}\}) = 20\%$
- Somente $\{\text{Café}\}$, $\{\text{Pão}\}$ e $\{\text{Manteiga}\}$ são frequentes
- Combinando os 1-*itemsets* frequentes e calculando o suporte
 - $\text{sup}(\{\text{Café}, \text{Manteiga}\}) = 30\%$
 - $\text{sup}(\{\text{Café}, \text{Pão}\}) = 30\%$
 - $\text{sup}(\{\text{Manteiga}, \text{Pão}\}) = 40\%$
- Somente $\{\text{Café}, \text{Manteiga}\}$ e $\{\text{Manteiga}, \text{Pão}\}$ são frequentes

Exemplo

- Calculando a confiança para regras geradas de {Café,Pão}
 - $\text{conf}(\{\text{Café}\} \Rightarrow \{\text{Pão}\}) = 30\%/30\% = 100\%$
 - $\text{conf}(\{\text{Pão}\} \Rightarrow \{\text{Café}\}) = 30\%/50\% = 60\%$
- Calculando a confiança para regras geradas de {Café,Manteiga}
 - $\text{conf}(\{\text{Café}\} \Rightarrow \{\text{Manteiga}\}) = 30\%/30\% = 100\%$
 - $\text{conf}(\{\text{Manteiga}\} \Rightarrow \{\text{Café}\}) = 30\%/50\% = 60\%$
- Calculando a confiança para regras geradas de {Manteiga,Pão}
 - $\text{conf}(\{\text{Manteiga}\} \Rightarrow \{\text{Pão}\}) = 40\%/50\% = 80\%$
 - $\text{conf}(\{\text{Pão}\} \Rightarrow \{\text{Manteiga}\}) = 40\%/50\% = 80\%$

- $\text{sup}(\{\text{Café}, \text{Manteiga}, \text{Pão}\}) = 30\%$
- Calculando a confiança para regras geradas de $\{\text{Café}, \text{Manteiga}, \text{Pão}\}$
 - $\text{conf}(\{\text{Café}, \text{Pão}\} \Rightarrow \{\text{Manteiga}\}) = 30\%/30\% = 100\%$
 - $\text{conf}(\{\text{Café}, \text{Manteiga}\} \Rightarrow \{\text{Pão}\}) = 30\%/30\% = 100\%$
 - $\text{conf}(\{\text{Manteiga}, \text{Pão}\} \Rightarrow \{\text{Café}\}) = 30\%/40\% = 75\%$
 - $\text{conf}(\{\text{Café}\} \Rightarrow \{\text{Pão}, \text{Manteiga}\}) = 30\%/30\% = 100\%$
 - ...

Medidas de Interesse

- O número de regras geradas utilizando somente suporte e confiança geralmente é muito grande
 - Regras não interessantes
 - Dificuldade da análise por parte do usuário
- Há também o problema do cobertor curto
 - Valores de suporte e confiança altos geram muitas regras, porém, geralmente óbvias
 - Valor de suporte e confiança baixo geram muitas regras, porém, pode haver regras potencialmente interessantes

Medidas de Interesse

- Itens com suporte muito alto acabam por compor a maioria das regras
- Mesmo com valores de suporte e confiança altos, as regras de associação podem expressar relações não válidas
- Medidas de interesse são comumente utilizadas para avaliar a qualidade de uma regra de associação

Medidas de Interesse

Tabela: Conjunto de regras de exemplo

ID	Regra	Sup _A	Sup _B	Sup	Conf
1	{cenoura} \Rightarrow {batata_inglesa}	0,7701	0,8175	0,7038	0,9138
2	{acerola} \Rightarrow {ovos}	0,0427	0,8886	0,0379	0,8889
3	{filé_de_viola} \Rightarrow {açúcar_refinado}	0,0877	0,8649	0,0758	0,8649
4	{milho_verde} \Rightarrow {ervilhas}	0,3294	0,3791	0,2701	0,8201
5	{fruta_do_conde} \Rightarrow {melancia}	0,0450	0,1422	0,0308	0,6842
6	{banana_nanica} \Rightarrow {banana_prata}	0,1209	0,7607	0,0735	0,6078

Medidas de Interesse

- Independência entre itens
 - A confiança da regra 3 $\{\text{filé_de_viola}\} \Rightarrow \{\text{açúcar_refinado}\}$ é 0,8649
 - Se observamos, o suporte de $\{\text{açúcar_refinado}\}$ é 0,8649, portanto a compra de $\{\text{filé_de_viola}\}$ não aumenta nem diminui a chance de comprar $\{\text{açúcar_refinado}\}$
 - A compra de $\{\text{açúcar_refinado}\}$ independe da compra de $\{\text{filé_de_viola}\}$
 - O mesmo acontece para a regra 2 $\{\text{acerola}\} \Rightarrow \{\text{ovos}\}$

Medidas de Interesse

- Independência entre itens
 - Os *itemsets* são independentes se

$$\text{conf}(A \Rightarrow B) = \text{sup}(B)$$

- Lembrando que

$$\text{conf}(A \Rightarrow B) = \text{sup}(A \cup B) / \text{sup}(A)$$

- Portanto, os *itemsets* são **independentes** se

$$\text{sup}(A \cup B) = \text{sup}(A) \times \text{sup}(B)$$

- $\text{sup}(A) \times \text{sup}(B)$ representa o **suporte esperado** (sup_esp) de $A \cup B$, i.e., $\text{sup_esp}(A \cup B) = \text{sup}(A) \times \text{sup}(B)$

Medidas de Interesse

- Dependência negativa entre itens
 - A confiança da regra $\{banana_nanica\} \Rightarrow \{banana_prata\}$ é 0,6078
 - A probabilidade de qualquer cliente comprar $\{banana_prata\}$ é 0,7607
 - A compra de $\{banana_nanica\}$ diminui a chance de comprar $\{banana_prata\}$
 - Os *itemsets* de uma regra possuem dependência negativa se

$$sup(A \cup B) < sup_esp(A \cup B)$$

Medidas de Interesse

- Dependência positiva entre itens

- A confiança da regra 4 $\{\text{milho_verde}\} \Rightarrow \{\text{ervilhas}\}$ é 0,8201
- A probabilidade de qualquer cliente comprar $\{\text{ervilhas}\}$ é 0,3791
- A compra de $\{\text{milho_verde}\}$ aumenta a chance de comprar $\{\text{ervilhas}\}$
- Os clientes que compraram $\{\text{milho_verde}\}$ tem maior probabilidade de comprar $\{\text{ervilhas}\}$
- Os *itemsets* de uma regra possuem dependência positiva se

$$\text{sup}(A \cup B) > \text{supesp}(A \cup B)$$

Medidas de Interesse

- A medida confiança, por não considerar a dependência entre os *itemsets*, pode gerar um número muito grande de regras que apresentam relacionamentos falsos
- Utilização de medidas de interesse objetivas
 - Índices estatísticos utilizados para selecionar regras interessantes
 - Suporte e confiança são exemplos de medidas de interesse objetivas

Lift

- A medida Lift, também conhecida por *Interest*, é uma das mais utilizada para avaliar dependências entre *itemsets*

$$lift(A \Rightarrow B) = \frac{conf(A \Rightarrow B)}{sup(B)} = \frac{sup(A \cup B)}{sup(A) \times sup(B)}$$

- $lift(A \Rightarrow B) = 1$: A e B são independentes
- $lift(A \Rightarrow B) > 1$: A e B são positivamente dependentes
- $lift(A \Rightarrow B) < 1$: A e B são negativamente dependentes
- Intervalo: $[0 - \infty[$
- Simétrica
- Quanto maior o valor de Lift, mais interessante a regra

Rule Interest

- A medida *Rule Interest*, também conhecida por Piatetsky-Shapiro, novidade ou *leverage*, também é utilizada para avaliar dependência entre os *itemsets*

$$ri(A \Rightarrow B) = sup(A \cup B) - sup_esp(A \cup B)$$

- $ri(A \Rightarrow B) = 0$: A e B são independentes
- $ri(A \Rightarrow B) > 0$: A e B são positivamente dependentes
- $ri(A \Rightarrow B) < 0$: A e B são negativamente dependentes
- Intervalo: $[-0, 25 - 0, 25]$
- Simétrica
- Quanto maior o valor da medida, mais interessante é a regra

Convicção

- Objetivo é avaliar a “implicação” da regra

$$\text{conv}(A \Rightarrow B) = \frac{\text{sup}(A) \times \text{sup}(\neg B)}{\text{sup}(A \cup \neg B)}$$

- Avalia o quanto A e $\neg B$ se afastam da independência
- Caso exista independência completa entre o antecedente e o consequente da regra, o valor da convicção será igual a 1
- Regras onde o antecedente nunca aparece sem o consequente (confiança de 100%) terão valor de convicção igual a ∞
- Intervalo: $[0 - \infty[$
- Assimétrica

Medidas de Interesse

Tabela: Conjunto de regras de exemplo

ID	Regra	Sup _A	Sup _B	Sup _{A∪B}	Conf _{A∪B}	Lift _{A∪B}	RI _{A∪B}	Convicção _{A∪B}
1	{cenoura} ⇒ {batata_inglesa}	0,7701	0,8175	0,7038	0,9138	1,1179	0,0742	2,1198
2	{acerola} ⇒ {ovos}	0,0427	0,8886	0,0379	0,8889	0,9989	0,0000	0,9910
3	{filé_de_viola} ⇒ {açúcar_refinado}	0,0877	0,8649	0,0758	0,8649	0,9993	0,0000	0,9957
4	{milho_verde} ⇒ {ervilhas}	0,3294	0,3791	0,2701	0,8201	2,1630	0,1452	3,4490
5	{fruta_do_conde} ⇒ {melancia}	0,0450	0,1422	0,0308	0,6842	4,8133	0,0244	2,7184
6	{banana_nanica} ⇒ {banana_prata}	0,1209	0,7607	0,0735	0,6078	0,7992	-0,0184	0,6104

Outras medidas de interesse objetivas

- Outras medidas de interesse objetivas podem ser encontradas em:
 - [Geng and Hamilton, 2006]
 - [Guillet and Hamilton, 2007]
 - [Tan et al., 2002]

Regras de Associação por Atributos Binário Simétricos

- Há situações em que os valores 0 e 1 (ou ausência e presença) são igualmente importantes
- Por exemplo, ao preencher um formulário, o usuário selecionará como 1 as opções que ele escolheu
- Porém, regras interessantes podem ser extraídas com o que o usuário não escolheu
- *“Se o usuário não gosta de high school musical então ele gosta de futebol”*

Regras de Associação par Atributos Binário Simétricos

- Pode-se utilizar algoritmos específicos para minerar regras desse tipo
- Ou pode-se processar a base original para que se possam aplicar algoritmos tradicional para extrair regras de associação
- Ex:
 - Por exemplo, para um atributo simétrico comunicação → valor 1 caso o usuário deseja receber informações no e-mail sobre determinado produto ou serviço e 0 caso contrário
 - Esse atributo pode ser convertido em dois atributos binários assimétricos → `consente_comunicação` e `não_consente_comunicação`

Regras de Associação por Atributos Binário Simétricos

Opção 1	Opção 2		Opção 1 SIM	Opção 1 NÃO	Opção 2 SIM	Opção 2 NÃO
1	0		1	0	0	1
0	1		0	1	1	0
1	1		1	0	1	0
1	1		1	0	1	0
1	0		1	0	0	1
1	0		1	0	0	1
1	1		1	0	1	1



Regras de Associação para Atributos Categóricos

- Em algumas bases de dados os atributos podem ser categóricos ao invés de binários
- **Ex:** estados, níveis de educação, cor de cabelo, sexualidade, ...
- Neste caso, também é possível converter os atributos categóricos para atributos binários assimétricos → cada atributo receberá o valor zero se o valor do mesmo ocorrer no registro e 1 caso contrário

Regras de Associação para Atributos Categóricos

Cabelo	Nível de Ensino
Loiro	Fundamental
Acaju	Médio
Preto	Superior
Preto	Superior
Loiro	Médio
Acaju	Superior
Preto	Médio



Cabelo_Loiro	Cabelo_Acaju	Cabelo_Preto	Nível_Fundamental	Nível_Médio	Nível_Superior
1	0	0	1	0	0
0	1	0	0	1	0
0	0	1	0	0	1
0	0	1	0	0	1
1	0	0	0	1	0
0	1	0	0	0	1
0	0	1	0	1	0

Regras de Associação para Atributos Contínuos

- Há algoritmos específicos para extrair regras de associação de atributos contínuos (idade, altura, salário, ...)
- Entretanto, pode-se também pré-processar a base de dados de maneira a aplicar os algoritmos para extração de regras de associação tradicionais
 - Os atributos contínuos são transformados em atributos discretos por meio de discretização

Regras de Associação para Atributos Contínuos

Idade	Altura (m)
21	1,35
15	1,51
14	1,65
58	1,71
70	1,93
50	1,75
23	1,85



Idade	Altura (m)
Adulto	Anão
Jovem	Baixo
Jovem	Médio
Idoso	Médio
Idoso	Alto
Adulto	Médio
Adulto	Alto



Idade_Jovem	Idade_Adulto	Idade_Idoso	Altura_Anão	Altura_Baixo	Altura_Médio	Altura_Alto
0	1	0	1	0	0	0
1	0	0	0	1	0	0
1	0	0	0	0	1	0
0	0	1	0	0	1	0
0	0	1	0	0	0	1
0	1	0	0	0	1	0
0	1	0	0	0	0	1

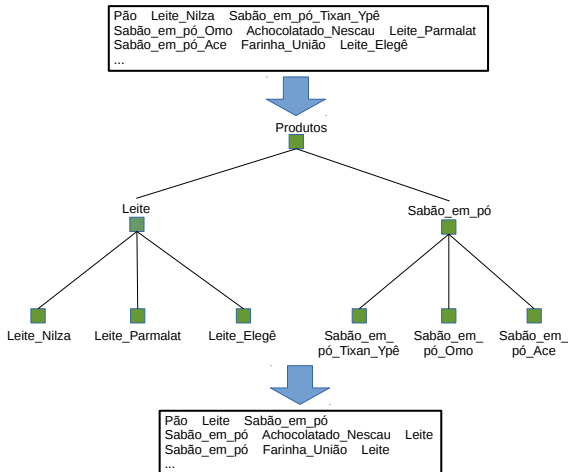
Regras de Associação Multidimensionais

- Conjunto de dados podem ser armazenados em diferentes *datawarehouses* e *data marts* e organizados de acordo com várias dimensões lógicas
- Pode-se também aplicar regras de associação para considerar múltiplas dimensões
- **Ex:** *“se um consumidor compra uma câmera digital, possui 30-40 anos de idade, têm gastos anuais médios de US\$300 – 500, então ele irá comprar uma impressora com probabilidade de 78%”*.
- **OBSERVAÇÃO:** dimensões com valores numéricos ou categóricos devem ser discretizadas e apenas uma valor referente a cada dimensão devem aparecer nas regras

Regras de Associação de Múltiplos Níveis

- Em algumas aplicações, não é possível extrair regras de associação fortes devido a rareficação dos dados
- Pode ser que alguns itens não sejam incluídos nas transações como *itemsets* frequentes devido à baixa ocorrência nas transações
- Entretanto, os objetos pertencentes às transações usualmente pertencem à uma hierarquia de conceitos
- Com isso, é possível remediar a rareficação de determinados itens transferindo à análise para um nível mais alto na hierarquia de conceitos → número de um determinado item irá aumentar nas transações

Regras de Associação de Múltiplos Níveis



Regras de Associação Sequenciais

- Frequentemente as transações são armazenadas de acordo com uma sequência temporal específica
- **Ex:** as transações sobre caminhos navegados na web são associados com a ordem/sequência temporal das visitas
- Em tais situações, é interessante extrair regras que levem em conta dependências temporais
- Pode-se estender o algoritmo Apriori para extrair tais tipos de regras

Regras de Associação Sequenciais

SID	Sequence
1	$\langle \{a, b\}, \{c\}, \{f, g\}, \{g\}, \{e\} \rangle$
2	$\langle \{a, d\}, \{c\}, \{b\}, \{a, b, e, f\} \rangle$
3	$\langle \{a\}, \{b\}, \{f, g\}, \{e\} \rangle$
4	$\langle \{b\}, \{f, g\} \rangle$



Pattern	Sup.
$\langle \{a\} \rangle$	3
$\langle \{a\}, \{g\} \rangle$	2
$\langle \{a\}, \{g\}, \{e\} \rangle$	2
$\langle \{a\}, \{f\} \rangle$	3
$\langle \{a\}, \{f\}, \{e\} \rangle$	2
$\langle \{a\}, \{c\} \rangle$	2
$\langle \{a\}, \{c\}, \{f\} \rangle$	2
$\langle \{a\}, \{c\}, \{e\} \rangle$	2
$\langle \{a\}, \{b\} \rangle$	2
$\langle \{a\}, \{b\}, \{f\} \rangle$	2
$\langle \{a\}, \{b\}, \{e\} \rangle$	2
$\langle \{a\}, \{e\} \rangle$	3
$\langle \{a, b\} \rangle$	2
$\langle \{b\} \rangle$	4
$\langle \{b\}, \{g\} \rangle$	3
$\langle \{b\}, \{g\}, \{e\} \rangle$	2
$\langle \{b\}, \{f\} \rangle$	4
$\langle \{b\}, \{f, g\} \rangle$	2
$\langle \{b\}, \{f\}, \{e\} \rangle$	2
$\langle \{b\}, \{e\} \rangle$	3
$\langle \{c\} \rangle$	2
$\langle \{c\}, \{f\} \rangle$	2
$\langle \{c\}, \{e\} \rangle$	2
$\langle \{e\} \rangle$	3
$\langle \{f\} \rangle$	4
$\langle \{f, g\} \rangle$	2
$\langle \{f\}, \{e\} \rangle$	2
$\langle \{g\} \rangle$	3
$\langle \{g\}, \{e\} \rangle$	2

Material Complementar

- Association rule learning

https://en.wikipedia.org/wiki/Association_rule_learning

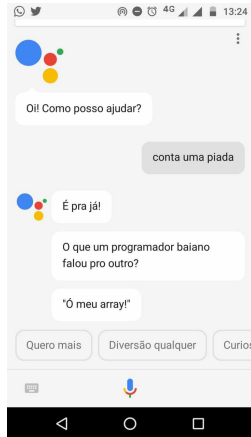
- Técnicas de Associação

<http://www.inf.ufrgs.br/~alvares/CMP259DCBD/RegrasDeAssociacao.pdf>

- Eles sabem o que você quer ver

<http://revistagalileu.globo.com/Revista/Common/0,,EMI334328-17773,00-ELES+SABEM+O+QUE+VOCE+QUER+VER.html>

Imagem do Dia



Inteligência Artificial
<http://lives.ufms.br/moodle/>

Rafael Geraldeli Rossi
rafael.g.rossi@ufms.br

Slides baseados em [Han et al., 2011], [Tan et al., 2005],
[Gonçalves, 2005] e nos *slides* do prof. Eduardo R. Hruschka

Referências Bibliográficas I



Agrawal, R. and Srikant, R. (1994).

Fast algorithms for mining association rules in large databases.

In *VLDB'94: Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.



Geng, L. and Hamilton, H. J. (2006).

Interestingness measures for data mining: A survey.

ACM Computing Surveys, 38(3):9.

Referências Bibliográficas II



Gonçalves, E. C. (2005).

Regras de associação e suas medidas de interesse objetivas e subjetivas.

INFOCOMP Journal of Computer Science, 4(1):26–35.



Guillet, F. and Hamilton, H. J., editors (2007).

Quality Measures in Data Mining, volume 43 of *Studies in Computational Intelligence*.

Springer.

Referências Bibliográficas III



Han, J., Kamber, M., and Pei, J. (2011).

Data Mining: Concepts and Techniques.

The Morgan Kaufmann Series in Data Management Systems.
Elsevier.



Tan, P.-N., Kumar, V., and Srivastava, J. (2002).

Selecting the right interestingness measure for association patterns.

In *SIGKDD'2002: Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 32–41. ACM.

Referências Bibliográficas IV



Tan, P.-N., Steinbach, M., and Kumar, V. (2005).
Introduction to Data Mining.
Addison-Wesley.