Aprendizado Descritivo

Aula 08 – Regras de associação e métricas de interesse

Professor Renato Vimieiro

DCC/ICEx/UFMG

Introdução

- Como discutimos anteriormente, o processo de extração de regras de associação em base de dados envolve duas etapas:
  - A mineração de padrões frequentes (que vimos ao longo das últimas aulas); e
  - A geração de regras interessantes a partir dos padrões
- A definição de ‘interesse’ é naturalmente subjetiva, pois depende do domínio da aplicação.
- Pode-se usar métricas na tentativa de tornar a medida de interesse mais objetiva
- O suporte dos padrões é uma forma de evitar regras espúrias, já que exclui padrões que ocorrem por chance
- Assim o processo de geração de regras se torna relativamente trivial a partir dos padrões frequentes

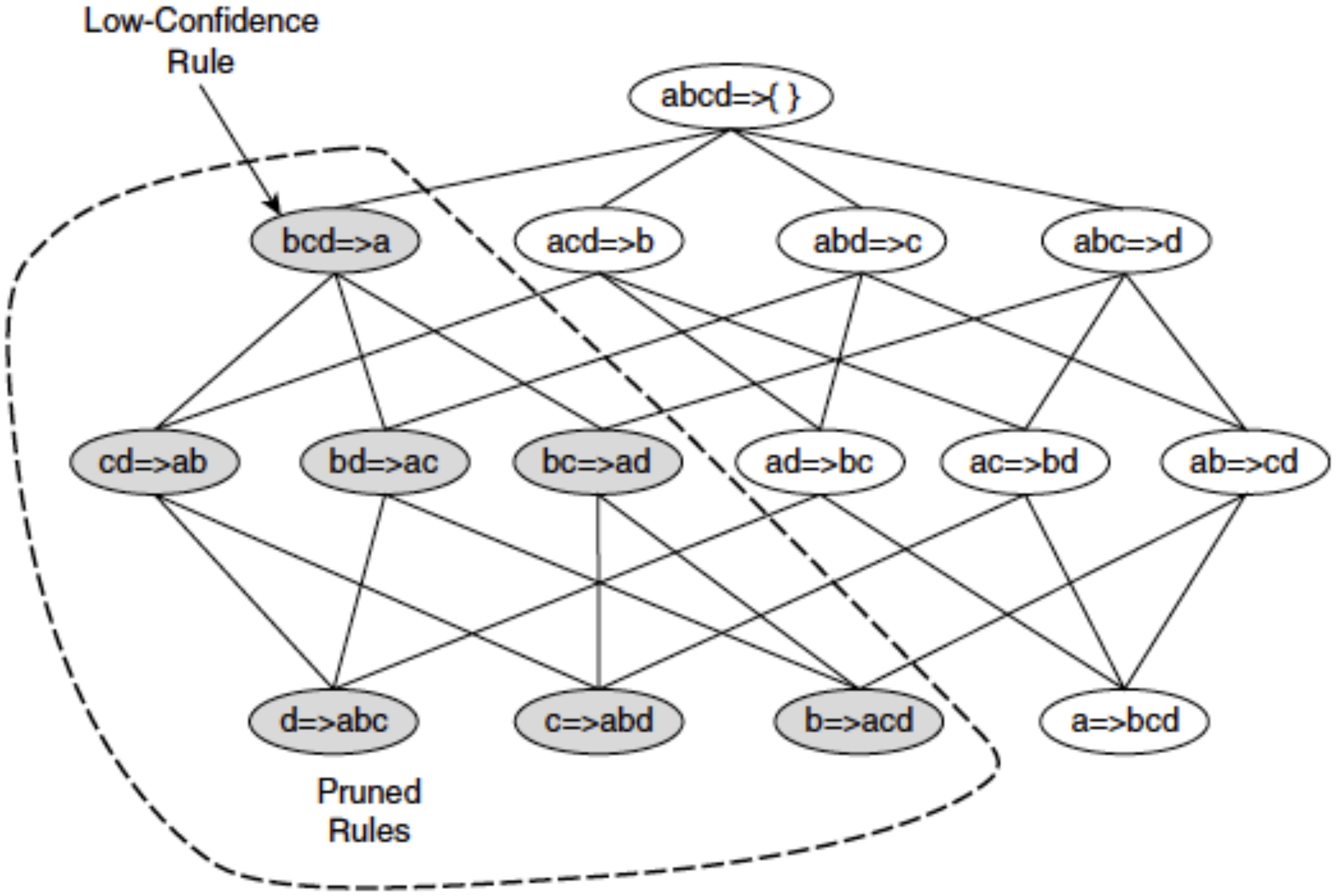
Regras de associação

- Uma regra de associação é uma implicação lógica do tipo $X \rightarrow \ Y$ em que X e Y são itemsets e $X \cap Y = \emptyset$
- O interesse (representatividade) de uma regra é definida por:
  - Suporte: $\mathrm{sup}(\mathrm{X} \rightarrow \mathrm{Y}) = \frac{\mathrm{sup}(X \cup Y)}{|D|}$
  - Confiança: $\mathrm{conf}(\mathrm{X} \rightarrow \mathrm{Y}) = \frac{\mathrm{sup}(X \cup Y)}{\mathrm{sup}(X)}$
- Uma regra é interessante se o suporte é maior que um limiar de suporte mínimo (minsup), e a confiança maior que uma confiança mínima (minconf)
- Enquanto o suporte mostra quão aplicável (frequente) a regra é na base de dados, a confiança mostra a força da associação entre os itemsets (quão provável é se observar os itens de Y nas transações da cobertura de X)

Regras de associação

- Como as regras interessantes são aquelas que satisfazem o suporte mínimo, elas podem ser geradas a partir dos itemsets frequentes
- Podem ser geradas $2^k - 2$ regras a partir de um k-itemset
  - $r(X) = \{(X - Y) \rightarrow Y \mid Y \ \subseteq X \ e \ 0 < |Y| < k\}$
  - Note que, como X é frequente, o suporte dos antecedentes e consequentes da regra também são itemsets frequentes e, portanto, já tiveram suporte computado. Assim, não há necessidade de novas passadas na base de dados para computar a confiança da regra
- Embora não seja anti-monotônica, se $(\mathrm{X} - Y) \rightarrow Y$ não satisfaz a restrição de confiança, então $(X - Y') \rightarrow Y'$ também não satisfaz para todo $Y' \supset Y$

Regras de associação



Retirado de Figura 6.15 de Tan et al.

Regras de associação

- Dessa forma, a geração das regras pode ser sistematizada em um procedimento similar ao usado no Apriori

Algorithm 6.3 Procedure **ap-genrules**(f_k, H_m).

$k = |f_k|$ {size of frequent itemset.}
$m = |H_m|$ {size of rule consequent.}
if $k > m + 1$ **then**
    $H_{m+1} = \text{apriori-gen}(H_m)$.
    for each $h_{m+1} \in H_{m+1}$ **do**
        $conf = \sigma(f_k)/\sigma(f_k - h_{m+1})$.
        if $conf \geq minconf$ **then**
            output the rule $(f_k - h_{m+1}) \longrightarrow h_{m+1}$.
        else
            delete h_{m+1} from H_{m+1}.
        end if
    end for
    call ap-genrules(f_k, H_{m+1}.)
end if

Retirado de Algoritmo 6.3 de Tan et al.

Medidas de interesse

- Apesar do suporte e confiança fornecerem um bom mecanismo para filtragem de regras desinteressantes, elas também apresentam falhas
- Como foi discutido anteriormente, o filtro por suporte pode excluir da análise itens altamente lucrativos porém raros
  - Diminuir o limiar de suporte mínimo inviabiliza a mineração dos padrões (tanto computacionalmente quanto no número de padrões retornados)
- Os problemas relacionados à confiança são mais sutis

Medidas de interesse

- Considere uma base de dados com a avaliação da preferência de bebidas quentes de um grupo de pessoas
- A regra $chá \rightarrow café$ parece uma regra interessante, já que tem suporte=15% e confiança=75%
  - A interpretação dessa regra é: 15% das pessoas bebem ambas as bebidas e 75% das que bebem chá também bebem café
- Essa interpretação induz a conclusão de que o fato da pessoa beber chá influencia ela a beber café
  - Porém 80% das pessoas bebem café, logo o fato delas beberem chá influenciam negativamente o consumo de café

	Café	!Café	
Chá	150	50	200
!Chá	650	150	800
	800	200	1000

Medidas de interesse

- O problema anterior nos leva a conclusão de que padrões envolvendo itens mutuamente independentes ou que cubram poucas transações são desinteressantes
- Assim, outras medidas de interesse podem ser utilizadas para complementar a informação trazida pelo arcabouço suporte-confiança
- O mais comum na literatura é usar métricas de correlação como medidas de interesse

Lift

- O problema identificado com a confiança é que ela sugeria uma correlação positiva entre os itemsets, enquanto, na verdade, existia uma correlação negativa
- O *lift* é uma medida que avalia a independência entre os itemsets ou o quanto a ocorrência de um itemset 'eleva' a ocorrência de outro
- Ela é definida por
  - $lift(X,Y) = \frac{P(X \cup Y)}{P(X)P(Y)} = \frac{conf(X \rightarrow Y)}{\sup(Y)}$
- Veja que:
  - Se o lift(X,Y)=1, então X e Y são independentes;
  - Se o lift(X,Y)<1, então X e Y são negativamente correlacionados; e
  - Se o lift(X,Y)>1, então X e Y são positivamente correlacionados
- Para o exemplo das bebidas, $lift(chá, café) = \frac{0.15}{0.2 \times 0.8} = 0.9375$, indicando a correlação negativa observada anteriormente

Limitações do Lift

- Embora o lift seja bastante útil na prática, ele também apresenta limitações
- Considere uma situação de mineração de textos em que palavras sejam itens. Podemos avaliar a co-ocorrência de *data mining* e *compiler mining* nos textos através da métrica
  - Espera-se que as duas primeiras estejam positivamente correlacionadas enquanto as duas últimas não
- Supondo as tabelas ao lado, temos que o lift de data mining é 1.02 e de compiler mining é 4.08 para nossa surpresa
  - Mas o primeiro tem suporte 88% contra 2% do segundo
  - A confiança das regras $data \rightarrow mining$ e $compiler \rightarrow mining$ são 94.6% e 28.5%, o que parece mais razoável

	p	$\overline{p}$	
q	880	50	930
$\overline{q}$	50	20	70
	930	70	1000

	r	$\overline{r}$	
s	20	50	70
$\overline{s}$	50	880	930
	70	930	1000

Mathews' Correlation Coefficient

- A correlação de Pearson é amplamente usada para variáveis contínuas, porém ela não se aplica a dados categóricos
- A correlação de dados categóricos é medida pelo coeficiente phi ou Mathews' Correlation Coefficient (MCC)
- O MCC é computado através da tabela de contingência da seguinte forma
  - $phi(X,Y) = (N_{XY} \times N_{!X!Y}) - (N_{!XY} \times N_{X!Y}) / (N_X \times N_Y \times N_{!X} \times N_{!Y})$
- O coeficiente varia entre –1 e +1, sendo
  - -1 indicativo de correlação negativa
  - +1 correlação positiva
  - 0 independência
- No caso do exemplo das bebidas, o valor do coeficiente é -0.0625 (ligeiramente negativamente correlacionado)

Limitações do MCC

- No caso do exemplo das palavras, se computarmos os coeficientes para ambos os casos, chegamos ao mesmo valor 0.232
- A razão é que o coeficiente dá o mesmo peso para co-ocorrências e co-ausências. Nesse caso, a medida é mais adequada para variáveis simétrica
- A medida também é sensível mesmo a mudanças proporcionais no tamanho da amostra

	p	$\overline{p}$	
q	880	50	930
$\overline{q}$	50	20	70
	930	70	1000

	r	$\overline{r}$	
s	20	50	70
$\overline{s}$	50	880	930
	70	930	1000

Chi-quadrado

- O teste do χ^2(chi-quadrado) é muito usado para avaliar a dependência entre variáveis categóricas
- O teste é computado a partir da tabela de contingência das variáveis
  - Faz-se uma tabulação cruzada e computa-se a número de ocorrências de cada combinação da variáveis

	B	!B	
A	N_{AB}	$N_{A!B}$	N_{A}
!A	$N_{!AB}$	$N_{!A!B}$	$N_{!A}$
	N_{B}	$N_{!B}$	Total

Chi-quadrado

- O valor da estatística χ^2 é obtido computando-se a proporção do desvio entre o valor observado e esperado em cada célula da matriz
  - $\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}$, onde O_{ij} e E_{ij} são respectivamente os valores observados e esperados para cada célula
- O valor esperado de cada célula é obtido assumindo-se a independência das variáveis
  - $E_{ij} = \frac{N_i * N_j}{Total}$

Chi-quadrado

- Uma vez computado o valor da estatística, pode-se usar uma tabela de valores críticos para estimar a probabilidade de se observar valores tão extremos quanto o calculado
  - Caso essa probabilidade seja relativamente pequena, rejeitamos a hipótese nula de independência entre as variáveis
  - No caso das tabelas 2x2, o grau de liberdade é 1 e o valor crítico para uma probabilidade de 0.05 é 3.84 ($P(\chi^2 \geq 3.84) = 0.05$)
- Finalmente, descartada a independência das variáveis, compara-se o valor observado com o esperado para concluir o tipo de correlação

Chi-quadrado

- A tabela do exemplo anterior foi atualizada com os valores esperados
- O valor da estatística é $\chi^2 = 3.906$, o que é maior que o valor crítico
- Portanto, as variáveis são dependentes
- Como o valor observado para chá e café é ligeiramente inferior, conclui-se que elas são negativamente correlacionadas

	Café	!Café	
Chá	150 (160)	50 (40)	200
!Chá	650 (640)	150 (160)	800
	800	200	1000

Cosseno

- A similaridade de cosseno é amplamente usada em aprendizado de máquina, sobretudo em aplicações envolvendo texto
- A medida pode ser adaptada para itemsets se visualizarmos suas coberturas como vetores binários
- Ela é formalmente definida por:
    - Cosseno(X,Y) = sup(X U Y)/sqrt(sup(X) x sup(Y))
    - Ela também pode ser vista como a média geométrica das confianças das regras X->Y e Y->X
- Os valores do cosseno variam entre 0 e 1, e indicam maior relação entre os itemsets quando estiver mais próximo de 1
- A vantagem é que ela depende somente das proporções das ocorrências de X, Y e X U Y (a ausência não é levada em conta como no MCC)

Comparação das medidas

Data Set	mc	$\overline{m}c$	$m\overline{c}$	$\overline{mc}$	χ^2	*lift*	*cosine*
D_1	10,000	1000	1000	100,000	90557	9.26	0.91
D_2	10,000	1000	1000	100	0	1	0.91
D_3	100	1000	1000	100,000	670	8.44	0.09
D_4	1000	1000	1000	100,000	24740	25.75	0.5
D_5	1000	100	10,000	100,000	8173	9.18	0.29
D_6	1000	10	100,000	100,000	965	1.97	0.10

Leitura

- Seção 6.7 Tan et al.
- Seção 6.4 Han et al.
- Capítulo 12 Zaki e Meira

Aprendizado Descritivo

Aula 08 – Regras de associação e métricas de interesse

Professor Renato Vimieiro

DCC/ICEx/UFMG