# Mineração de Dados em Biologia Molecular

## Análise de associação

Docente: André C. P. L. F. de Carvalho
PAE: Victor Hugo Barella

## Principais tópicos

- Análise de associação
- Itens frequentes
- Conjunto de itens frequentes
- Regras de associação
- Avaliação de regras de associação



## Análise de associação

- Várias empresas acumulam grandes quantidades de dados de transações diárias
  - Cesta de transações
  - Ex.: compras em supermercados
- Descoberta de relacionamentos interessantes escondidos em grandes conjuntos de dados
  - Análise dos dados pode revelar associações entre atributos preditivos
  - Ex.: Produtos comprados em conjunto



## Itens frequentes

- Um dos temas mais pesquisados em MD
- Trabalho pioneiro: análise de cestas de compras
  - Encontrar grupos de itens frequentemente comprados juntos
  - Utilizar conhecimento extraído para inferir outros itens que podem ser comprados
    - Podem ser exibidos em locais próximos



## Itens frequentes

- Seja A = $\{a_1, ..., a_m\}$ um conjunto de *m* itens
  - Produtos, peças de equipamentos, genes ...
  - Qualquer subconjunto $I \subseteq A$ é denominado um conjunto de itens ou *itemset*
    - Ex.: qualquer conjunto de produtos que podem ser comprados juntos
    - Itemset com k-itens é chamado k-itemset
    - Conjunto nulo ou vazio: itemset sem itens



## Conjuntos de itens frequentes

- Seja T = $(t_1, ,t_2, ..., t_n\}$ um conjunto de *n* transações
  - Forma um banco de dados de transações
    - Ex.: conjunto de compras feiras em um supermercado em um dia
  - Cada transação $t_i$ contém um k-itemset$_i$
    - K = largura da transação

## Exemplo

| Transação | Itens comprados |
|---|---|
| 1 | pão, queijo, manteiga, massa |
| 2 | pão, geléia, suco |
| 3 | queijo, arroz, massa |
| 4 | massa, queijo |
| 5 | massa, queijo, pão |



## Exemplo

| Transação | Itens comprados |
|---|---|
| 1 | pão, queijo, manteiga, massa |
| 2 | pão, geléia, suco |
| 3 | queijo, arroz, massa |
| 4 | massa, queijo |
| 5 | massa, queijo, pão |

pão, queijo, manteiga, massa é um 4-itemset
Transação $T_2$ contém um 3-itemset$_2$



## Conjuntos de itens frequentes

- Contagem de suporte
  - Número de transações que contêm um dado itemset
  - Contagem de suporte para um itemset I

$$\sigma(I) = \left|\{t_i / I \subseteq t_i, t_i \in T\}\right|$$



## Conjuntos de itens frequentes

| Transação | Itens comprados |
|---|---|
| 1 | pão, queijo, manteiga, massa |
| 2 | pão, geléia, suco |
| 3 | queijo, arroz, massa |
| 4 | queijo, vinho |
| 5 | massa, queijo, pão |

σ(pão e queijo) =
σ(suco) =



## Conjuntos de itens frequentes

| Transação | Itens comprados |
|---|---|
| 1 | pão, queijo, manteiga, massa |
| 2 | pão, geléia, suco |
| 3 | queijo, arroz, massa |
| 4 | queijo, vinho |
| 5 | massa, queijo, pão |

σ(pão e queijo) = 2
σ(suco) = 1



## Conjuntos de itens frequentes

- Dados um conjunto de itens frequentes e um conjunto de dados de transações
  - É possível encontrar regras de associação entre itens
    - Regras de associação inferem relações entre itens

## Regras de associação

- Uma regra de associação tem o formato de uma regra *Se A Então B*

Antecedente Consequente

  - Antecedente e consequente são itemsets disjuntos ($A \cap B = \varnothing$)
  - Ex.: se cliente compra pão e queijo então ele compra manteiga e café



## Conjunto de dados

- Cada transação é um objeto
  - Ex.: compras em um supermercado
    - Cada item (produto) é um atributo

| Transação | Itens comprados |
|---|---|
| 1 | pão, queijo, manteiga, massa |
| 2 | pão, geléia, suco |
| 3 | queijo, arroz, massa |
| 4 | massa, queijo |
| 5 | massa, queijo, pão |



## Conjunto de dados

- Representa objetos por vetores binários
- Cada item é um atributo

| Trans. | Itens comprados |
|---|---|
| 1 | pão, queijo, manteiga, massa |
| 2 | pão, geléia, suco |
| 3 | queijo, arroz, massa |
| 4 | massa, queijo |
| 5 | massa, queijo, pão |

| Arroz | Geléia | Manteiga | Massa | Pão | Queijo | Suco |
|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 | 1 | 1 | 0 |



## Regras de Associação

- A qualidade de uma regra pode ser medida por seu suporte e pela sua confiança
  - Suporte: frequência com que uma regra é aplicável a um conjunto de dados

$$Sup(A \rightarrow B) = \frac{\sigma(A \cup B)}{N}$$

  - Confiança: quão frequentemente itens de B aparecem em transações que contêm A

$$Conf(A \rightarrow B) = \frac{\sigma(A \cup B)}{\sigma(A)}$$



## Regras de Associação

- Suporte
  - Quanto menor o suporte de uma regra, maior a chance de ela ocorrer por acaso
  - Permite eliminar regras pouco interessantes
- Confiança
  - Mede a confiabilidade de uma inferência feita por uma regra
  - Fornece uma estimativa da probabilidade condicional de B dado A



## Exemplo

- Definir o suporte e a confiança para a regra **Se queijo Então massa** para os dados

| Transação | Itens comprados |
|---|---|
| 1 | pão, queijo, manteiga, massa |
| 2 | pão, geléia, suco |
| 3 | queijo, arroz, massa |
| 4 | queijo, vinho |
| 5 | massa, queijo, pão, suco |

## Exemplo

$$Sup(A \rightarrow B) = \frac{\sigma(A \cup B)}{N}$$

$$Conf(A \rightarrow B) = \frac{\sigma(A \cup B)}{\sigma(A)}$$

- Se queijo Então massa
- Resultado
  - σ (queijo e massa) =
  - σ (queijo) =
  - N =
  - Sup (queijo → massa) =
  - Conf (queijo → massa) =



## Exemplo

$$Sup(A \rightarrow B) = \frac{\sigma(A \cup B)}{N}$$

$$Conf(A \rightarrow B) = \frac{\sigma(A \cup B)}{\sigma(A)}$$

- Se queijo Então massa
- Resultado
  - σ (queijo e massa) = 3
  - σ (queijo) = 4
  - N = 5
  - Sup (queijo → massa) = 3/5 = 0,6
  - Conf (queijo → massa) = 3/4 = 0,75



## Exercício

- Definir o suporte e a confiança para as regras a seguir:
  - Se queijo então vinho
  - Se massa então pão
  - Se queijo e pão então massa
  - Se geléia então suco
  - Se suco então manteiga



## Observação

- Itemsets e regras de associação geralmente trabalham com valores qualitativos
  - Conversão de quantitativos para qualitativos
  - Pode tambem incluir valores quantitativos
    - Quantidade de cada item em uma transação



## Algoritmo para achar regras

- Método de força bruta
  - Pesquisa todas as possibilidades com suporte e confiança maior que um limiar

*Definir $sup_{min}$ e $conf_{min}$*
*Encontrar itemsets frequentes*
*Todas as regras com sup ≥ $sup_{min}$*
*Descobrir regras com confiança elevada*
*Usar itemsets frequentes com conf ≥ $conf_{min}$*



## Encontrar itemsets frequentes

- Calcular sup para cada itemset com:
  - 1 item
    - Descartar itemsets com suporte $< sup_{min}$
  - 2 itens
    - Descartar itemsets com suporte $< sup_{min}$
  - ...
  - N itens
    - Descartar itemsets com suporte $< sup_{min}$

## Encontrar itemsets frequentes

- Cálculo do suporte
  - Número de possíveis itemsets para conjunto de dados com $d$ itens: $2^d-1$ (conjunto nulo)
  - Para cada itemset candidato, varrer todas as transações
- Usar força bruta tem um custo computacional elevado



## Descobrir regras com conf elevada

- Para cada itemset
  - Gerar todas as possíveis regras com pelo menos um item em A e um item em B
    - Todas as variações de A e B para o itemset
  - Calcular a confiança para cada regra
  - Manter apenas as regras com confiança $\geq conf_{min}$
- Número de possíveis regras extraídas de um conjunto de dados com $d$ itens: $3^d - 2^{d+1} + 1$
  - Para d = 6, é possível encontrar 602 regras



## Exemplo

- Encontrar regras de associação para o conjunto de dados abaixo usando força bruta
- Com $sup_{min}$ = 0,2 e $conf_{min}$ = 0,4

| Transação | Itens comprados |
|---|---|
| 1 | pão, queijo |
| 2 | pão |



## Exemplo (itemsets)

| Transação | Itens comprados |
|---|---|
| 1 | pão, queijo |
| 2 | pão |

- 1 item:
- 2 item:
- ...



## Exemplo (itemsets)

| Transação | Itens comprados |
|---|---|
| 1 | pão, queijo |
| 2 | pão |

- 1 item:
  - pão
  - queijo
- 2 item:
  - pão, queijo



## Exemplo (regras)

| Transação | Itens comprados |
|---|---|
| 1 | pão, queijo |
| 2 | pão |

- Regras
  - pão → queijo
  - queijo → pão

## Exemplo (contagem de suporte)

| Transação | Itens comprados |
|---|---|
| 1 | pão, queijo |
| 2 | pão |

- 1 item:
  - σ (pão) = 2
  - σ (queijo) = 1
- 2 item:
  - σ (pão, queijo) = 1



## Exemplo (suporte de regras)

| Transação | Itens comprados |
|---|---|
| 1 | pão, queijo |
| 2 | pão |

- Regras
  - sup (pão → queijo) / 2 = 0,5
  - sup ( queijo → pão / 2 = 0,5
  - Suporte das duas regras ≥ 0,2



## Exemplo (confiança de regras)

| Transação | Itens comprados |
|---|---|
| 1 | pão, queijo |
| 2 | pão |

- Regras
  - conf (pão → queijo) / 2 = 0,5/ σ (pão)
  - conf ( queijo → pão / 2 = 0,5/ σ (queijo)



## Exemplo (confiança de regras)

| Transação | Itens comprados |
|---|---|
| 1 | pão, queijo |
| 2 | pão |

- Regras
  - conf (pão → queijo) / 2 = 0,5/2 = 0,25
  - conf ( queijo → pão / 2 = 0,5/1 = 0,5
  - Confiança de uma das duas regras ≥ 0,4



## Algoritmo Apriori

- Reduz conjunto de itemsets por meio de poda baseada em suporte
  - Elimina alguns itemsets candidatos antes de calcular seu suporte
  - Suporte de um itemset nunca é maior que o suporte de seus subconjunto
    - Se um itemset é frequente, todos os seus subconjuntos devem ser frequentes
    - Se um itemset é pouco frequente, todos os seus superconjuntos devem ser também



## Encontrar itemsets frequentes

- Computar sup para cada itemset com:
  - 1 item
    - Descartar itemsets com suporte $< sup_{min}$
  - 2 itens
    - Descartar itemsets com suporte $< sup_{min}$
  - ...
  - N itens
    - Descartar itemsets com suporte $< sup_{min}$
  - Itemsets com um subconjunto anteriormente descartado são ignorados

## Em bioinformática

- Regras entre atributos preditivos de dados biológicos
- Exemplo: dados de expressão gênica
  - **Se** genes A e C têm alta expressão e F tem baixa expressão **Então** genes MN e V têm baixa expressão



## Conclusão

- Descoberta de regras em geral tem um custo computacional elevado
- Algumas regras podem ser óbvias ou não ser válidas
  - Descobertas por acaso
- Benefícios:
  - Marketing
  - Ferramentas de bioinformática
  - Relacionamento com clientes
  - Promoções



## Conclusão

- Inferência encontrada não implica casualidade nem poder preditivo
  - Mas sim co-ocorrência entre itemsets
- Regras encontradas devem ser avaliadas por especialistas no domínio
  - Para definir utilidade
  - Usuário pode sugerir padrões de regras úteis e inúteis



## Exercício

- Encontrar 4 relações com Sup > 0.1 e Conf > 0.2 para os dados:

| Arroz | Geléia | Manteiga | Massa | Pão | Queijo | Suco |
|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| 1 | 1 | 1 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| 0 | 1 | 1 | 1 | 0 | 0 | 0 |



## Perguntas?



## Exercício

- Avaliar o desempenho do algoritmo Apriori
- Usar conjuntos de dados glass e iris
  - Variar suporte mínimo e confiança mínima
  - Comentar resultados