Data Science and Big Data

Taconeli, C.A.

23 de novembro, 2018

Mineração de regras de associação

Mineração de regras de associação (*association rules*)

- Mineração de bases de dados de transações comerciais;
- Identificação de padrões de compra e produtos (itens) adquiridos conjuntamente;
- Tem como principais objetivos definir estratégias de marketing, criar promoções personalizadas e recomendar novos produtos de acordo com o perfil do cliente;
- Aplicações em problemas gerais, em que se deseja descobrir relações ocultas em grandes bases de dados.

Mineração de regras de associação (*association rules*)

- Na mineração de regras de associação, a base de dados corresponde a um conjunto de transações (compras);

Tabela 1: Exemplo de transações em uma loja de conveniência

Transação	Itens
1	{Pão, Leite}
2	{Pão, Fraldas, Cerveja, Ovos}
3	{Leite, Fraldas, Cerveja, Refrigerante}
4	{Pão, Leite, Fraldas, Cerveja}
5	{Pão, Leite, Fraldas, Refrigerante}

Mineração de regras de associação (*association rules*)

- Uma regra de associação representa a relação entre itens, ou subconjuntos (cestas) de itens. Assim:

$$\{\text{Pão}\} \rightarrow \{\text{Leite}\}$$

é uma regra que sugere relação entre a compra de lpão e a compra de leite (muitos clientes que compram pão também compram leite).

- Regras de associação podem envolver múltiplos itens:

$$\{\text{Pão, Queijo}\} \rightarrow \{\text{Presunto}\}$$

$$\{\text{Pão, Queijo, Presunto}\} \rightarrow \{\text{Cerveja, Refrigerante}\}, \ldots$$

Mineração de regras de associação (*association rules*)

- Aplicações em outras áreas:
  - Na Medicina, em que as transações seriam pacientes, e os itens poderiam ser sintomas clínicos;
  - Na Ecologia, em que as transações seriam regiões de uma floresta, e os itens seriam espécies (animais, vegetais) observadas;
  - Nas Ciências Políticas, em que as transações seriam deputados ou senadores, e os itens seriam projetos que eles votariam como contrários ou favoráveis.

Mineração de regras de associação (*association rules*)

- É comum, originalmente, se dispor de uma base de dados binária sobre as transações.

- Neste caso, um item não adquirido em uma transação seria marcado por 0, enquanto um item adquirido por 1.

Tabela 2: Representação binária das transações em uma loja de conveniência

Transação	Pão	Leite	Fraldas	Cerveja	Ovos	Refrigerante
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

Cesta de itens e suporte

- Uma **cesta de itens** corresponde ao conjunto de produtos adquiridos em uma transação;
- Vamos denotar a cesta composta por todos os itens disponíveis (digamos d), por:

$$I = \{i_1, i_2, ..., i_d\}.$$

- Adicionalmente, vamos considerar um conjunto de N transações, denotado por:

$$T = \{t_1, t_2, ..., t_N\}.$$

Cesta de itens e suporte

- Cada transação t_i contém um subconjunto dos itens em I. Por exemplo, para $d = 10$ seguem relacionadas algumas transações:

Transação 1: $\{i_1, i_5, i_9\}$;

Transação 2: $\{i_2, i_5, i_6, i_9, i_{10}\}$;

Transação 3: $\{i_2\}$;

$\vdots$

Transação 10.000: $\{i_1, i_5, i_9\}$.

Cesta de itens e suporte

- Dizemos que uma transação t_j contém um conjunto de itens X se X é um subconjunto de t_j, ou seja, se todos os itens que compõem X aparecem na transação t_j;

- Voltando ao exemplo das transações na loja de conveniência (Tabelas 1 e 2), temos:

$$X = \{\text{Pão}\} \subseteq t_1;$$

$$X = \{\text{Pão, Ovos}\} \subseteq t_2;$$

$$X = \{\text{Pão, Ovos}\} \not\subseteq t_1.$$

Cesta de itens e suporte

- O **suporte** de um conjunto de itens X, aqui denotado por $S(X)$ é a frequência com X é verificado no conjunto de N transações:

$$S(X) = \left|\{t_i | X \subseteq t_i, t_i \in T\}\right|,$$

em que $|\cdot|$ denota o número de elementos no conjunto.

- Mais frequentemente o suporte de um conjunto de itens X é definido como a proporção de transações que contém X:

$$S(X) = \frac{\left|\{t_i | X \subseteq t_i, t_i \in T\}\right|}{N}.$$

Cesta de itens e suporte

- Voltando aos dados sobre transações da loja de conveniência, temos:

$$S(\{\text{Pão}\}) = \frac{4}{5} = 0.8;$$

$$S(\{\text{Ovos}\}) = \frac{1}{5} = 0.2;$$

$$S(\{\text{Leite, Pão}\}) = \frac{3}{5} = 0.6;$$

$$S(\{\text{Leite, Pão, Fraldas}\}) = \frac{2}{5} = 0.4.$$

Regras de associação

- uma regra de associação é uma implicação do tipo:

$$X \rightarrow Y,$$

onde X e Y são conjuntos de itens disjuntos, ou seja, $X \cap Y = \emptyset$.

- O **suporte de uma regra de associação** corresponde à frequência com que X e Y aparecem conjuntamente no total de transações:

$$S(X \rightarrow Y) = S(X \cup Y).$$

- A **confiança de uma regra de associação** corresponde à frequência com que Y aparece nas transações que contém X:

$$C(X \rightarrow Y) = \frac{S(X \cup Y)}{S(X)}.$$

Regras de associação

- Voltando aos dados sobre transações da loja de conveniência:

$$S(\{\text{Leite, Pão}\} \rightarrow \{\text{Fraldas}\}) = S(\{\text{Leite, Pão, Fraldas}\}) = \frac{2}{5} = 0.4,$$

ou seja, 40% das transações apresentam o conjunto de itens {Leite, Pão, Fraldas}.

$$C(\{\text{Leite, Pão}\} \rightarrow \{\text{Fraldas}\}) = \frac{S(\{\text{Leite, Pão, Fraldas}\})}{S(\{\text{Leite, Pão}\})} = \frac{2/5}{3/5} = 0.67,$$

ou seja, 67% das transações que apresentam os itens {Leite, Pão} também apresentam o item {Fraldas}.

Regras de associação

- O suporte de uma regra de associação é importante por que regras com baixo suporte podem ter sido verificadas nas transações por mero acaso;
- Além disso, regras com maior suporte correspondem a conjuntos de itens comercializados com maior frequência, sendo, em geral, de maior interesse;
- A confiança de uma regra de decisão é importante para se quantificar o *poder preditivo* de uma regra;
- Além disso, a confiança corresponde à probabilidade condicional de Y dado X.

Regras de associação

- Outra medida importante para se qualificar uma regra de decisão é o *lift*, definido por:

$$L(X \to Y) = \frac{S(X \cup Y)}{S(X)S(Y)}.$$

- O *lift* de uma regra de decisão pode ser interpretado como seu desvio em relação ao esperado se os conjuntos de itens X e Y ocorressem nas transações de forma independente.

- Quanto maior o valor do *lift*, maior a associação entre X e Y.

Regras de associação

- Voltando ao exemplo da loja de conveniência:

$$L(\{\text{Leite, Pão}\} \rightarrow \{\text{Fraldas}\}) = \frac{S(\{\text{Leite, Pão, Fraldas}\})}{S(\{\text{Leite, Pão}\})S(\{\text{Fraldas}\})}$$

$$= \frac{2/5}{3/5 \times 4/5} = 0.83;$$

$$L(\{\text{Leite, Pão}\} \rightarrow \{\text{Cerveja}\}) = \frac{S(\{\text{Leite, Pão, Cerveja}\})}{S(\{\text{Leite, Pão}\})S(\{\text{Fraldas}\})}$$

$$= \frac{1/5}{3/5 \times 3/5} = 0.56;$$

Regras de associação

$$L(\{\text{Fraldas, Pão}\} \rightarrow \{\text{Cerveja}\}) = \frac{S(\{\text{Fraldas, Pão, Cerveja}\})}{S(\{\text{Fraldas, Pão}\})S(\{\text{Cerveja}\})}$$

$$= \frac{2/5}{2/5 \times 3/5} = 1.67.$$

- Assim, a regra $\{\text{Fraldas, Pão}\} \rightarrow \{\text{Cerveja}\}$ apresenta maior associação que as regras $\{\text{Leite, Pão}\} \rightarrow \{\text{Fraldas}\}$ e $\{\text{Leite, Pão}\} \rightarrow \{\text{Cerveja}\}$.

Mineração de regras de associação

- A mineração de regras de associação corresponde à identificação de todas as regras do tipo $X \rightarrow Y$ tais que:

$$S(X \rightarrow Y) \geq s_p;$$

$$C(X \rightarrow Y) \geq c_p,$$

em que s_p e c_p referem-se ao suporte mínimo e à confiança mínima das regras a serem descobertas.

Mineração de regras de associação

- A derivação de todas as possíveis regras de associação é inviável, uma vez que o número de regras aumenta muito rapidamente conforme o número de itens.

- Pode-se mostrar que o número total de regras que podem ser extraídas de um conjunto de d itens é dado por:

$$R = 3^d - 2^{d+1} + 1.$$

- A título de ilustração, para $d = 5$ temos $R = 180$ regras; para $d = 10$, $R = 57.002$ regras e, para $d = 20$, $R = 3.484.687.250$ regras!

Mineração de regras de associação

- Uma estratégia para diminuir o esforço computacional é avaliar os requisitos de suporte e confiança separadamente;

- Considere o conjunto de itens $\{A, B, C\}$ e as seguintes regras de associação:

$$\{A, B\} \rightarrow \{C\}, \quad \{A, C\} \rightarrow \{B\},$$
$$\{B, C\} \rightarrow \{A\}, \quad \{A\} \rightarrow \{B, C\},$$
$$\{B\} \rightarrow \{A, C\}, \quad \{C\} \rightarrow \{A, B\}.$$

- Como as seis regras dependem de um mesmo suporte ($S(\{A, B, C\})$), então nenhuma delas precisará ser avaliada se o conjunto $\{A, B, C\}$ for pouco frquente.

Mineração de regras de associação

- Assim, a mineração de regras de associação fica decomposta nas seguintes etapas:

1. **Geração de conjuntos de itens frequentes**- objetivo é identificar todos os conjuntos de itens com suporte superior ao ponto de corte (s_p);

2. **Geração das regras**- objetivo é identificar todas as regras com elevada confiança produzidas pelos conjuntos de itens selecionados no passo 1.

Geração de conjuntos de itens frequentes

- A geração de todos os conjuntos de itens também é inviável, uma vez que o número total de conjuntos (M) para k itens é dado por:

$$M = 2^k - 1,$$

descontando-se o conjunto vazio.

- Para N transações, usando a *força bruta* seria necessário avaliar a ocorrência de cada conjunto em cada transação, o que, evidentemente, seria proibitivo.

Geração de conjuntos de itens frequentes

- Há diferentes formas de reduzir a dimensão desse problema, dentre elas:

1. **Reduzir o número de conjuntos de itens a serem avaliados** - Neste caso, o número de conjuntos para os quais precisamos calcular os respectivos suportes diminuiria. É a base do princípio *Apriori*;

2. **Reduzir o número de comparações** - O número de avaliações dos conjuntos de itens nas N transações pode ser reduzido alterando a estrutura da base de dados.

Geração de conjuntos de itens frequentes - o princípio *Apriori*

- O princípio *Apriori* se baseia no seguinte par de teoremas:

1. Se um conjunto de itens (X) é frequente, então todos os seus subconjuntos (todo Y tal que $Y \subset X$) também são frequentes;

2. Se um conjunto de itens (X) não é frequente, então todos os seus superconjuntos (todo Y tal que $X \subset Y$) também não são frequentes.

- Para melhor entendimento, vamos usar uma rede de itens.

Geração de conjuntos de itens frequentes - o princípio *Apriori*

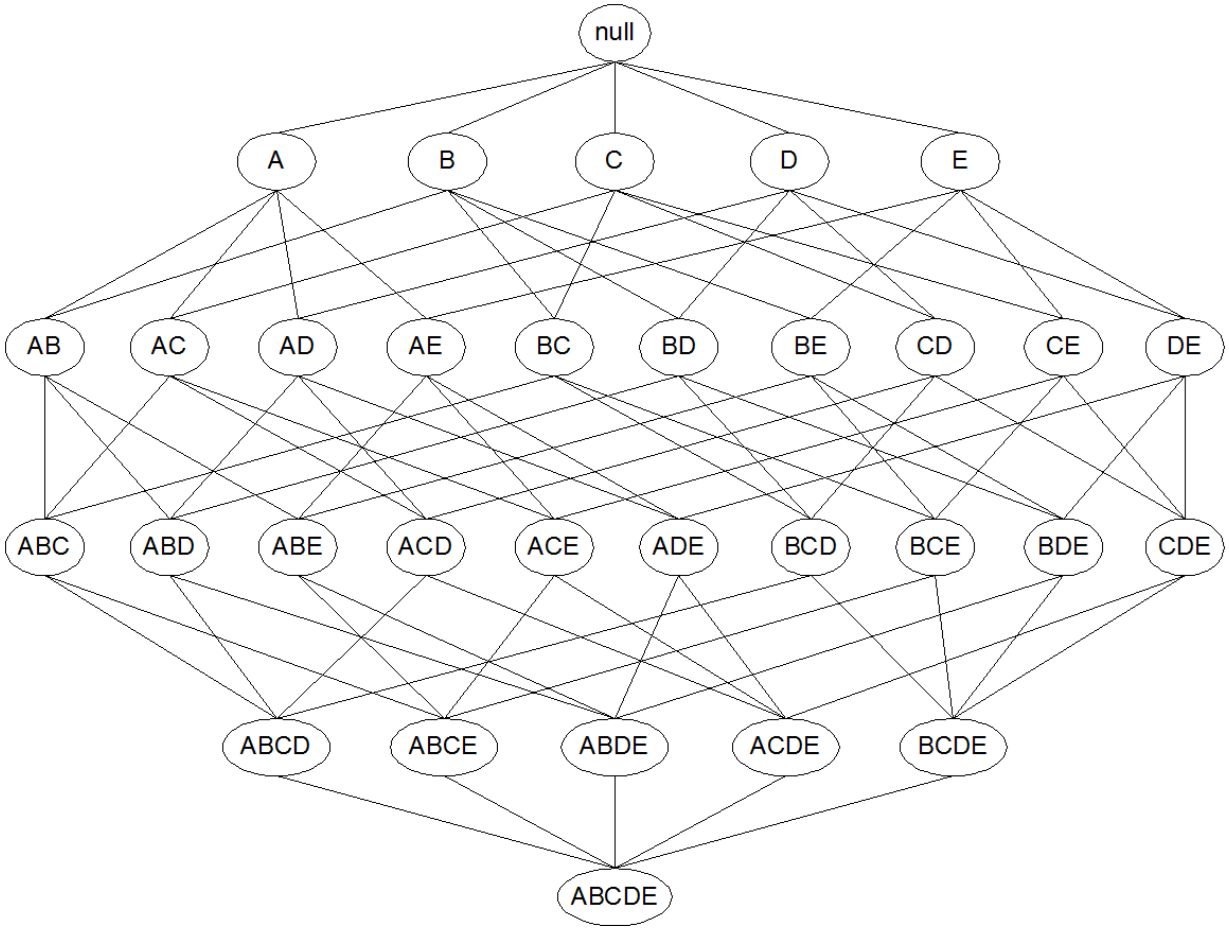

Figura 1: Rede de itens.

Geração de conjuntos de itens frequentes - o princípio *Apriori*

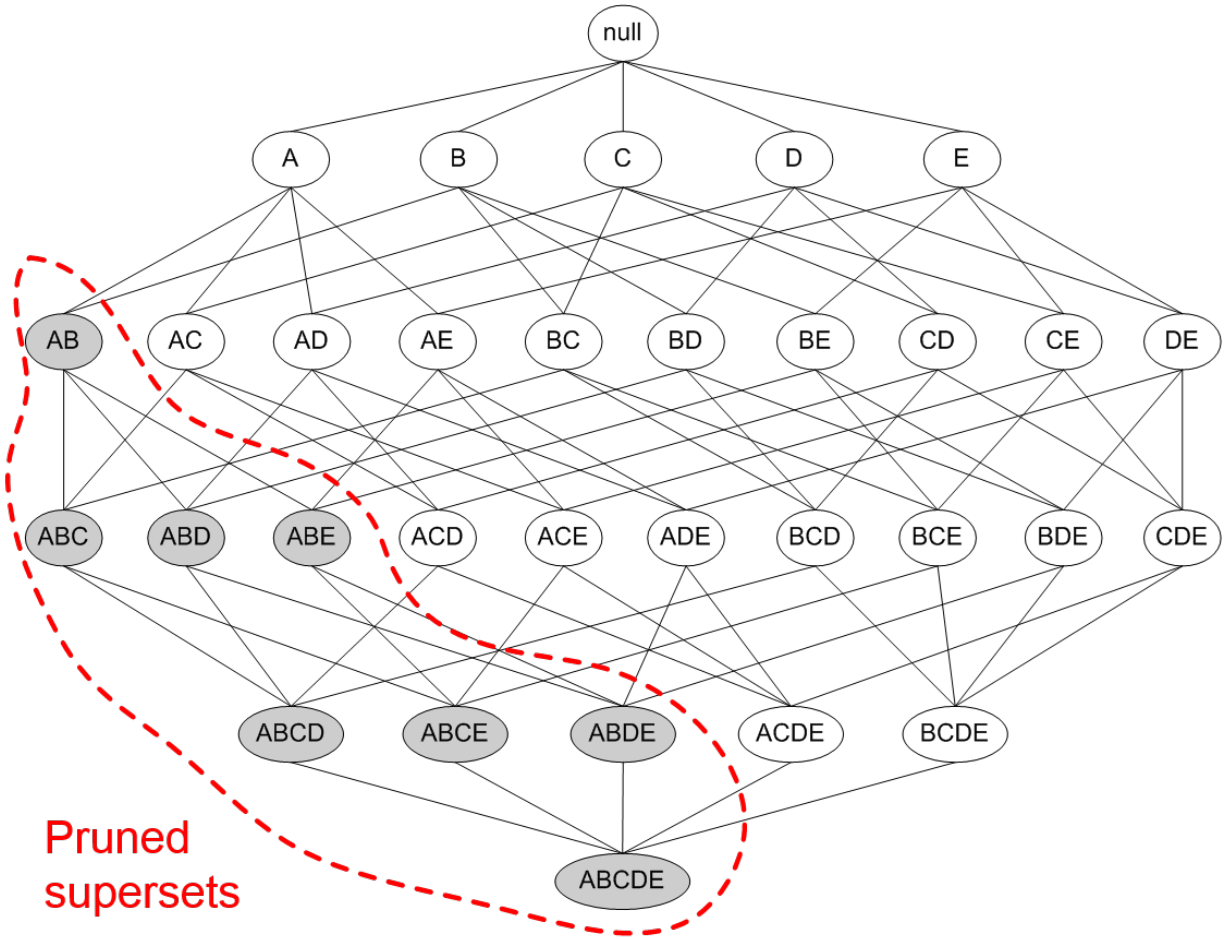

Figura 2: Ilustração do princípio Apriori - se {a,b} não é frequente, todos os seus superconjuntos também não são.

Geração de conjuntos de itens frequentes - o princípio *Apriori*

- Usando a Figura 2, se o conjunto $\{a, b\}$ não é frequente, então a rede pode ser podada de maneira que os respectivos superconjuntos não sejam avaliados.

- Podemos iniciar avaliando o suporte dos itens individualmente e eliminando aqueles com baixa frequência;

- Na segunda etapa, avaliamos apenas o suporte de pares de itens dentre aqueles que não foram eliminados no passo anterior. Novamente, conjuntos (pares) de itens com baixa frequência são eliminados;

- O processo continua, fazendo o *merge* dos conjuntos remanescentes e eliminando os novos conjuntos com baixa frequência, até que nenhum conjunto apresente o suporte mínimo.

Geração das regras

- Cada conjunto frequente Y com k itens pode originar $2^k - 2$ regras de associação, do tipo $X \rightarrow Y - X$, com $X \subset Y$.

- Todas essas regras satisfazem à restrição de suporte, porque são geradas a partir de um conjunto frequente.

- Calcular as confianças para essas regras de associação não requer avaliar novamente o conjunto de transações, uma vez que os suportes necessários já foram calculados.

Geração das regras

- A seleção (poda) das regras de associação baseia-se no seguinte teorema:
  - Se a regra $X \rightarrow Y - X$ não atinge o ponto de corte c_p, então qualquer regra do tipo $X' \rightarrow Y - X'$, onde $X' \subset X$, também não deve atingir o ponto de corte.

- O algoritmo *Apriori* gera regras de associação de forma hierárquica, onde cada etapa (nível) corresponde ao número de itens no conjunto que está a direita na regra.

- Desta forma, se a regra $\{a, b, c, d, e\} \rightarrow \{f\}$ não atinge c_p, emtão as regras $\{a, b, c, d\} \rightarrow \{e, f\}$; $\{a, b, d\} \rightarrow \{c, e, f\}$; $\{d, f\} \rightarrow \{a, b, c, f\}$; $\{d, e\} \rightarrow \{a, b, c, f\}$... podem ser imediatamente descartadas.

Geração de conjuntos de itens frequentes - o princípio *Apriori*

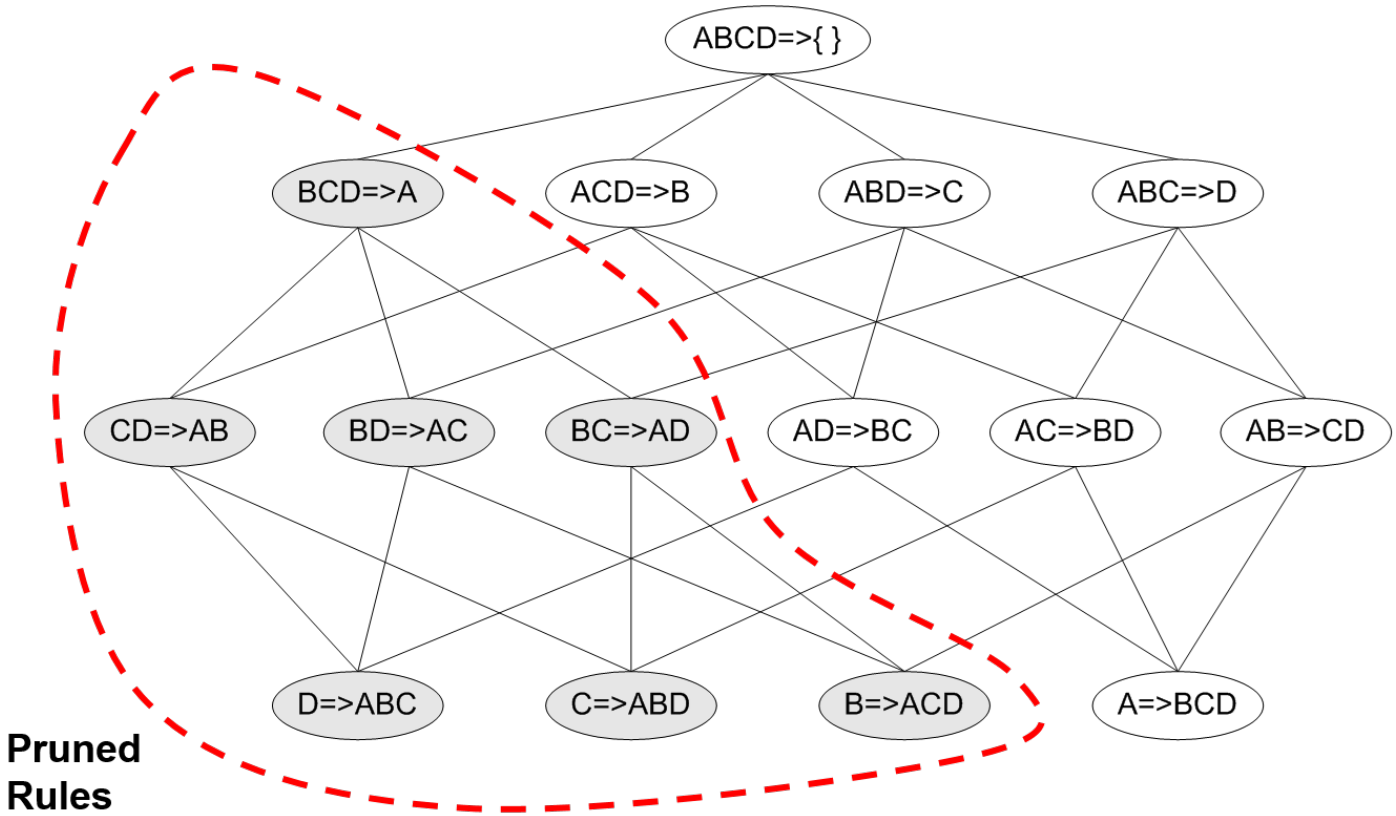

Figura 3: Rede de regras de associação e ilustração de poda. A regra $\{B, C, D\} \rightarrow \{A\}$ não atinge o ponto de corte.