Regras de associação

Gabriel Rosa

O que são as Regras de Associação?

- Utilizado para identificar possíveis relações entre diferentes itens dentro de um banco de dados, ou padrões existentes nas **transações** de um banco de dados;
- Têm seu formato no formato parecido com os de regras de classificação: *IF* A *THEN* B;
- Os itens que irão formar a regra de associação não contém nada em comum, ou seja, são coisas completamente distintas: $atributos(A) \cap atributos(B) = \emptyset$

Exemplo de regra de associação

- Dado esse dataset, com 5 transações de uma pessoa em uma padaria local fictícia, quais regras podem ser criadas? (Considerando que 1 é a presença do produto na compra e 0 a falta dele)

ID	Leite	Pão	Ovos
1	1	0	1
2	1	1	0
3	1	1	1
4	1	1	1
5	0	0	1

Exemplo de regra de associação

- Dado esse dataset, com 5 transações de uma pessoa em uma padaria local fictícia, quais regras podem ser criadas? (Considerando que 1 é a presença do produto na compra e 0 a falta dele)

ID	Leite	Pão	Ovos
1	1	0	1
2	1	1	0
3	1	1	1
4	1	1	1
5	0	0	1

Uma possível regra seria a do *ID 2*: **SE comprou leite ENTÃO comprará pão.**

Definição de conceitos e especificações

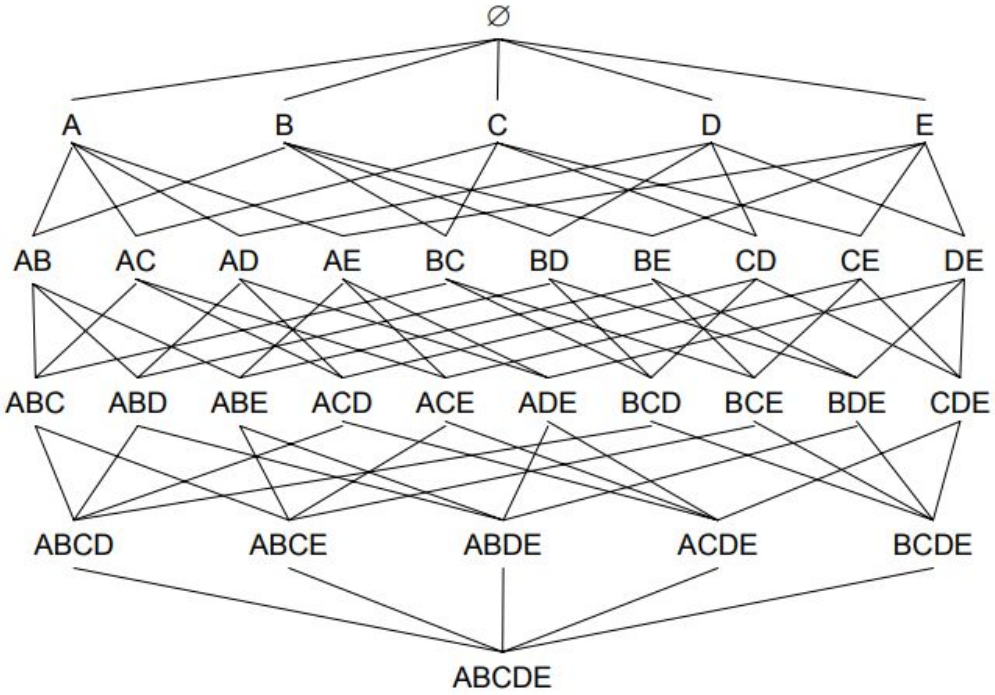
- Item: É o atributo de determinado dado a ser analisado; Ex.: Leite;
- Itemset: É a conjunção desses atributos para formar as regras de associação, geralmente representado por **e,** ∧ ou um espaço em branco. Ex.: Leite **e** Ovos;
- Os itemsets podem ser de variados tamanhos, dependendo de quantos atributos estão sendo utilizados. Por exemplo, 1-itemset, realiza o teste com um único atributo, 2-itemset, dois atributos, e assim sucessivamente. Generalizando temos que um **r**-itemset, contém **r** atributos;
- Transações: É o dado a ser analisado. Ex.: A lista de compras realizada;
- O lado esquerdo da regra é a premissa e o direito é a conclusão;
- Na regra exemplo *IF* A *THEN* B, A e B sempre são subconjuntos do conjunto de itens e não tem nenhuma relação entre si;

Problemáticas das regras de associação

- O número de regras de associação utilizadas para um determinado dataset tende a crescer exponencialmente, afinal, seguindo a ideia de indução para cada transação iria gerar muitas regras. Para cada **m** itens únicos, existem **2^m-1** itemsets, desconsiderando o conjunto vazio;
- E para cada itemset, temos um conjunto de regras.
- A fórmula que sinaliza o número de regras de associação conforme o número de itemsets é 3^m - 2^m, sendo **m** o número de itens únicos no dataset.

Problemáticas das regras de associação

m	Itemsets	Número de Regras
1	2	1
2	4	5
3	8	19
4	16	65
5	32	211
6	64	665
7	128	2059
8	256	6305
9	512	19171
10	1024	58025



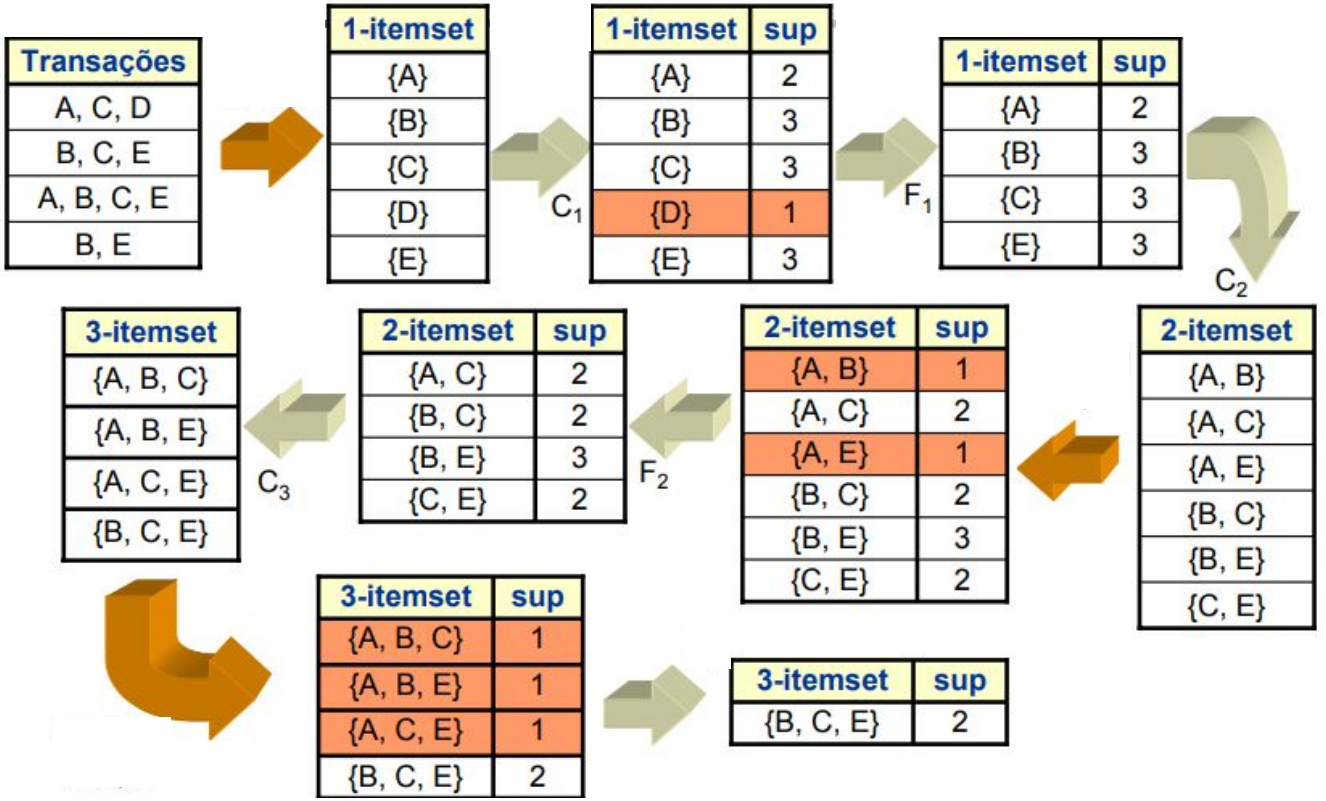
Métricas para utilização das regras de associação

- Suporte: É a proporção de vezes que um conjunto X aparece no dataset, ou no total de transações. Têm como fórmula **supp(X)** = ocorrências de X / total de transações;
- Confiança: Calculada em cima das regras de associação. Por base da regra X=>Y, a confiança é a quantidade de ocorrências de Y em todas as ocorrências de X. Tem como fórmula **conf(X=>Y)** = supp(X U Y) / supp(X);
- Lift: É a chance de Y, caso o X ocorra, considerando a popularidade de Y, caso seu valor seja maior que 1 então o Y tem chance de ocorre quando X ocorrer, se menor ou igual a 1 então Y **não** tem chance de ocorrer se X ocorrer. tem como fórmula **lift(X=>Y)** = supp(X U Y) / (supp(X) * supp(Y)).
- Convicção: É a chance de Y não ocorrer quando X ocorre, ou seja, as escessões da regra. Tem como fórmula conv(X=>Y) = (1-supp(Y)) / (1 - (conf(X=>Y));

Algoritmo Apriori

- Utiliza do suporte da regra para estipular e reduzir o número total de regras de associação a serem utilizadas em um dataset;
- Se baseia na ideia de que *Qualquer subconjunto de um itemset frequente também é frequente;*
- Os passos tomados pelo algoritmo são:
    - Encontrar os r-itemsets que tenham uma frequência maior ou igual ao **min_sup** (valor mínimo para suporte das regras);
    - Dos r-itemsets que sobraram, combine eles para formar **r+1-itemsets** e compare os valores resultantes do suporte para todos com o min_sup;
    - Repita até que nenhum valor esteja acima do min_sup;

Exemplo (com o valor de min_sup = 2)



Exemplo (com o valor de min_sup = 2)

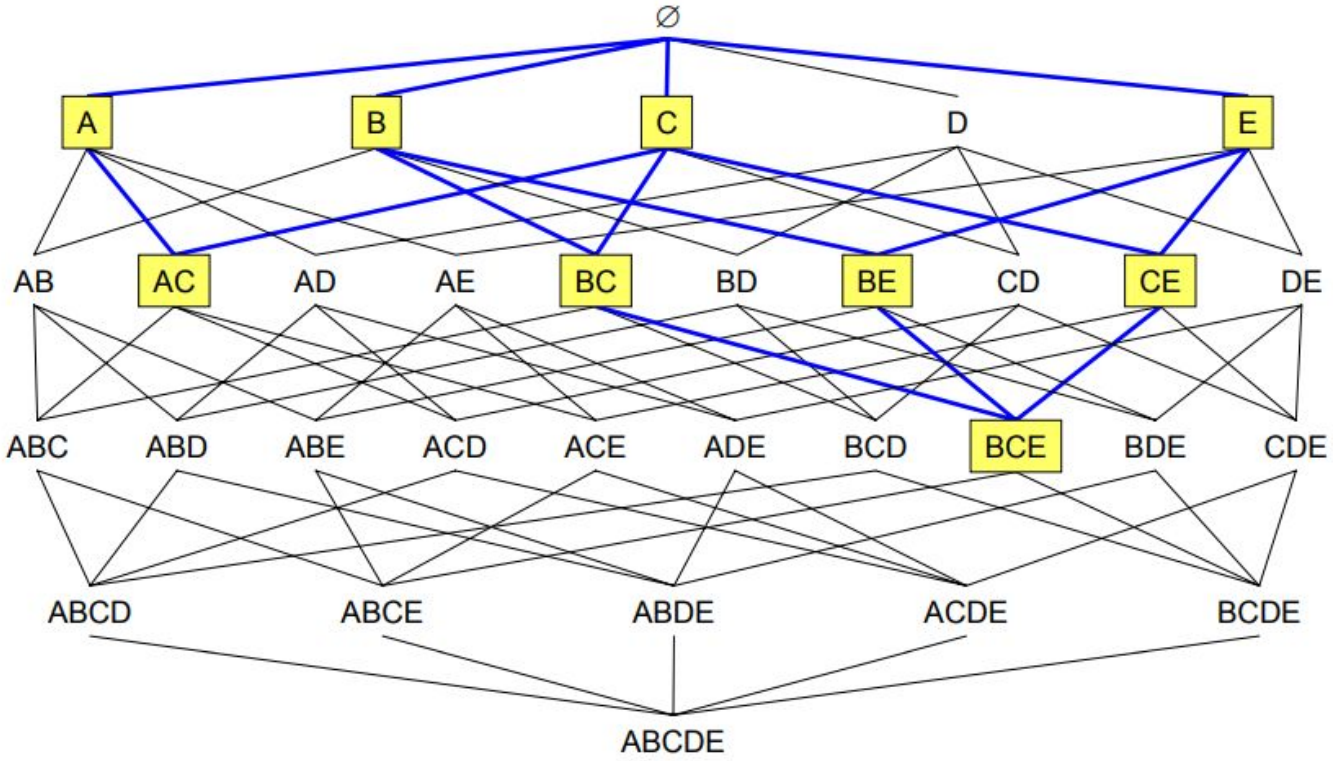
- A partir do processo realizado no slide anterior, o nosso conjunto de regras de associação utilizará de apenas esses três conjuntos de itemsets

1-itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

2-itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

3-itemset	sup
{B, C, E}	2

Exemplo (com o valor de min_sup = 2)



Algoritmo de RIPPER

- Vem da sigla: Repeated Incremental Pruning to Produce Error Reduction;
- Tem o objetivo de lidar com datasets com muito ruído e dataset com número de itens de cada classe desiguais, uma classe muito dominante sobre as outras;
- Funciona definindo um conjunto de regras a partir do conjunto de treinamento;
- Abrange dois casos para lidar com cada situação:
  - Caso 1 - Quando, dentre duas classes, há uma única muito predominante. Nesse caso o algoritmo irá considerá-la como a classe padrão e aprenderá/derivará regras que consigam identificar a outra classe.
  - Caso 2 - Quando há mais de duas classes. Nesse caso o algoritmo irá identificar a frequência de cada classe e a que tiver a maior será a classe padrão.
- A derivação das regras ocorre da seguinte forma:
  - Para cada conjunto é definido exemplos positivos (EP) e exemplos negativos (EN), sendo eles os que pertencem a classe padrão e os das outras classe, respectivamente.
  - A partir desses exemplos é utilizado o Algoritmo de Cobertura Sequencial para formulação das regras.

Algoritmo de RIPPER

- Utiliza de regras gerais para uma estratégia específica, ou seja, ele começa com uma regra vazia e vai adicionando o melhor conjunto, dentro de suas métricas, no antecedente.
- Para avaliação do melhor conjunto a métrica utilizada é o Ganho de Informação FOIL;
- Para de formar regras quando a regra a ser adicionada abrange mais os conjuntos dos exemplos negativos (EN) do que os exemplos positivos (EP);
- A construção do conjunto de regras do algoritmo é definido por:
    - Depois da regra ser derivada, os EP e os EN cobertos pela regra são eliminados;
    - A regra é adicionada caso não entre nos critérios de parada: *Princípio do comprimento mínimo da descrição* ou pela *Taxa de erro.*