

Data Mining

Regras de Associação

Prof. Dr. Joaquim Assunção

DEPARTAMENTO DE COMPUTAÇÃO APLICADA
CENTRO DE TECNOLOGIA

UFSM

2023

Fair user agreement

Este material foi criado para a disciplina de Mineração de Dados - Centro de Tecnologia da UFSM.

Você pode usar este material livremente*; porém, caso seja usado em outra instituição, **me envie um e-mail** avisando o nome da instituição e a disciplina.

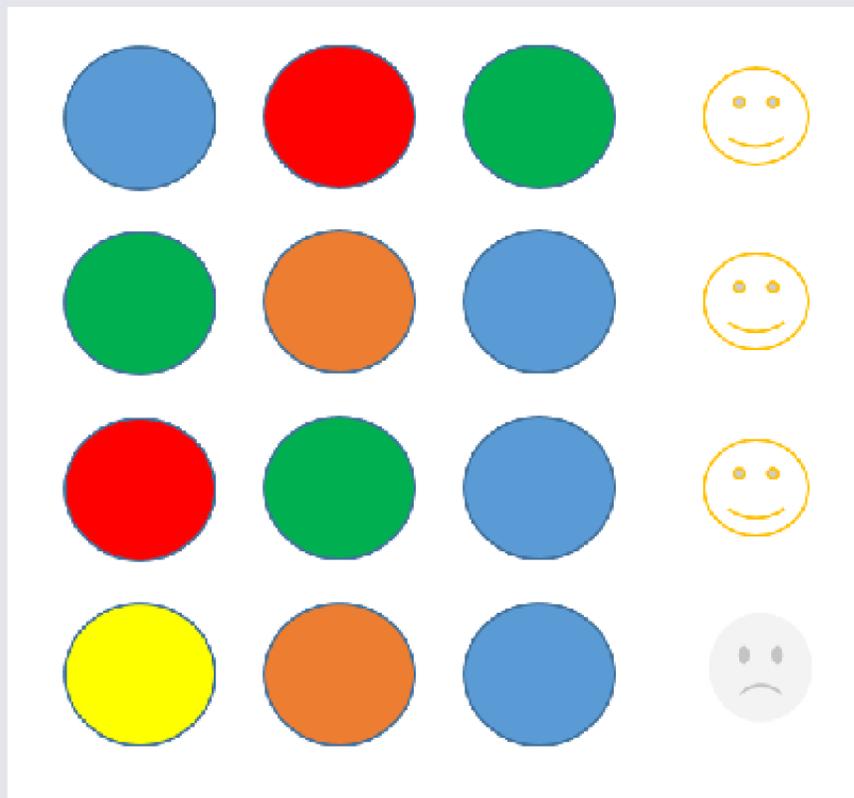
*A maior parte deste material foi retirado do livro: “**Joaquim V. C. Assunção. Uma Breve Introdução à Mineração de Dados: Bases Para a Ciência de Dados, com Exemplos em R.** 192 páginas. Novatec. 2021. ISBN-10 : 6586057507.”

Prof. Dr. Joaquim Assunção.

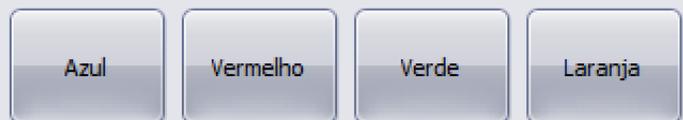
joaquim@inf.ufsm.br

Prequel – Association Game

Lv0 - Zoólogos estudam um primata que sorri quando uma determinada combinação de cores aparece. A julgar pela combinação, qual a cor preferida do primata?



Score
000



Nível 1. Uma bactéria tem se espalhado por vitaminas vendidas em uma banca de sucos. Dada a seguinte combinação de frutas em cada vitamina, qual a fruta que está espalhando a bactéria?



Nível 2. Uma nova praga se espalhou, dessa vez sabemos que não é uma única fruta, mas uma combinação que prolifera a bactéria qual a combinação em questão?



Espécies de formigas, classificadas de "A" a "N", são responsáveis por destruir uma determinada planta. Pedacos de terra foram isolados de acordo com o tipo de formiga. Qual a espécie mais danosa para a plantação?

		Inseto			Destruido
A	B	A	B	C	SIM
		E	G	F	SIM
		C	L	M	NAO
		C	G	H	NAO
		K	C	N	NAO
		C	B	H	SIM
		A	I	B	SIM
		J	K	I	SIM

Final... Sem pensar muito, associe as imagens.

A



?



B



?



C



?



Considerações:

- Você provavelmente levou poucos segundos para responder corretamente a primeira.
- Você provavelmente levou o dobro do tempo para responder a segunda (em relação a primeira), embora a quantidade de elementos pouco tenha aumentado.
- Você provavelmente precisou prestar bem mais atenção na terceira.
- A última foi fácil e intuitiva.

Considerações:

- Você conseguiria resolver o segundo problema se, por exemplo, fosse composto por 20 frutas, 30 linhas e 5 colunas?
 - E o terceiro problema se fosse composto por dezenas de espécies?
- Ao passo que último problema representa algo que humanos fazem bem, os problemas intermediários ilustram o quão pequeno é o limite da capacidade humana em obter padrões. Porém, o aumento da quantidade de dados não elimina padrões, frequentemente, apenas os esconde da capacidade humana.

Considerações

- Há conhecimento em meio aos dados. Saber quais as combinações implicam um resultado é um conhecimento a ser obtido. Tal tipo de conhecimentos muitas vezes está implícito em um conjunto de dados.

Introdução

Em determinados casos uma ação está associada a outra com probabilidade p .

Exemplo: Quem compra queijo possivelmente comprará presunto.

Conjunto de dados

Os conjuntos de dados devem ser compostos por:

- Conjunto de transações **T**;
- Conjunto **I** com **M** atributos binários.
- Conjunto **T** é composto por **N** transações.

$$I = i_1, i_2, \dots, i_M$$

$$T = t_1, t_2, \dots, t_N$$

Problema

Dado um conjunto de transações, encontrar a probabilidade de ocorrência de um conjunto de itens dada a ocorrência de um outro conjunto de itens.

Mineração de dados no Wal-Mart

Dadas incontáveis notas de compras, quais itens tendem a ser comprados simultaneamente?

Compra 1: pão, ovos, queijo

Compra 2: arroz, feijão, sal

...

Compra N: pão, queijo, presunto

Mineração de dados no Wal-Mart

Lembre-se da definição: “*A extração não trivial de informação implícita, previamente desconhecida e potencialmente útil*” W. Frawley, G. Piatetsky-Shapiro e C. Matheus.

pão => queijo

...

pão, presunto => queijo, margarina

Mineração de dados no Wal-Mart

Fraldas e Cerveja.



Weka

 Weka GUI Chooser



Program Visualization Tools Help

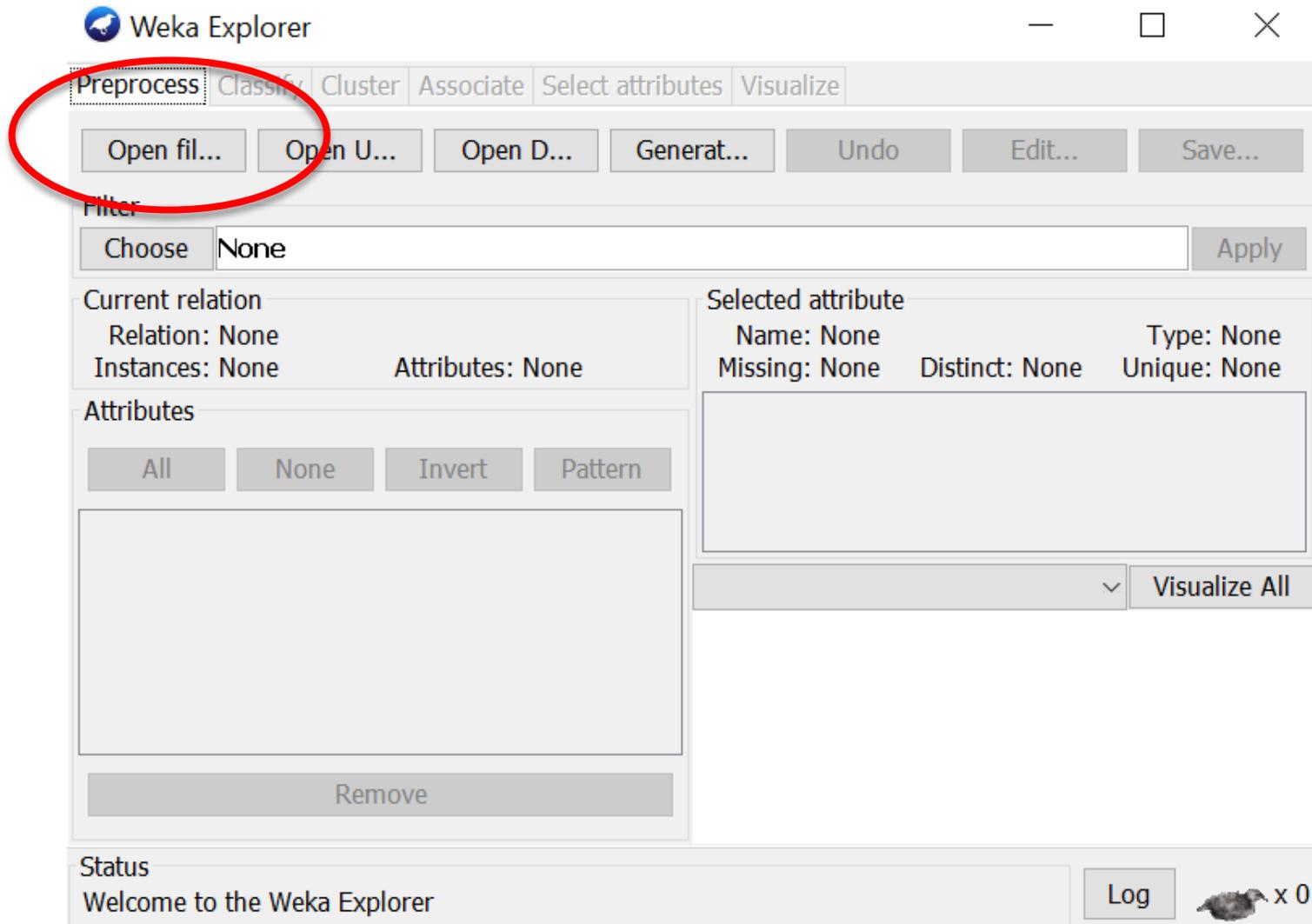


Waikato Environment for Knowledge Analysis
Version 3.6.13
(c) 1999 - 2015
The University of Waikato
Hamilton, New Zealand

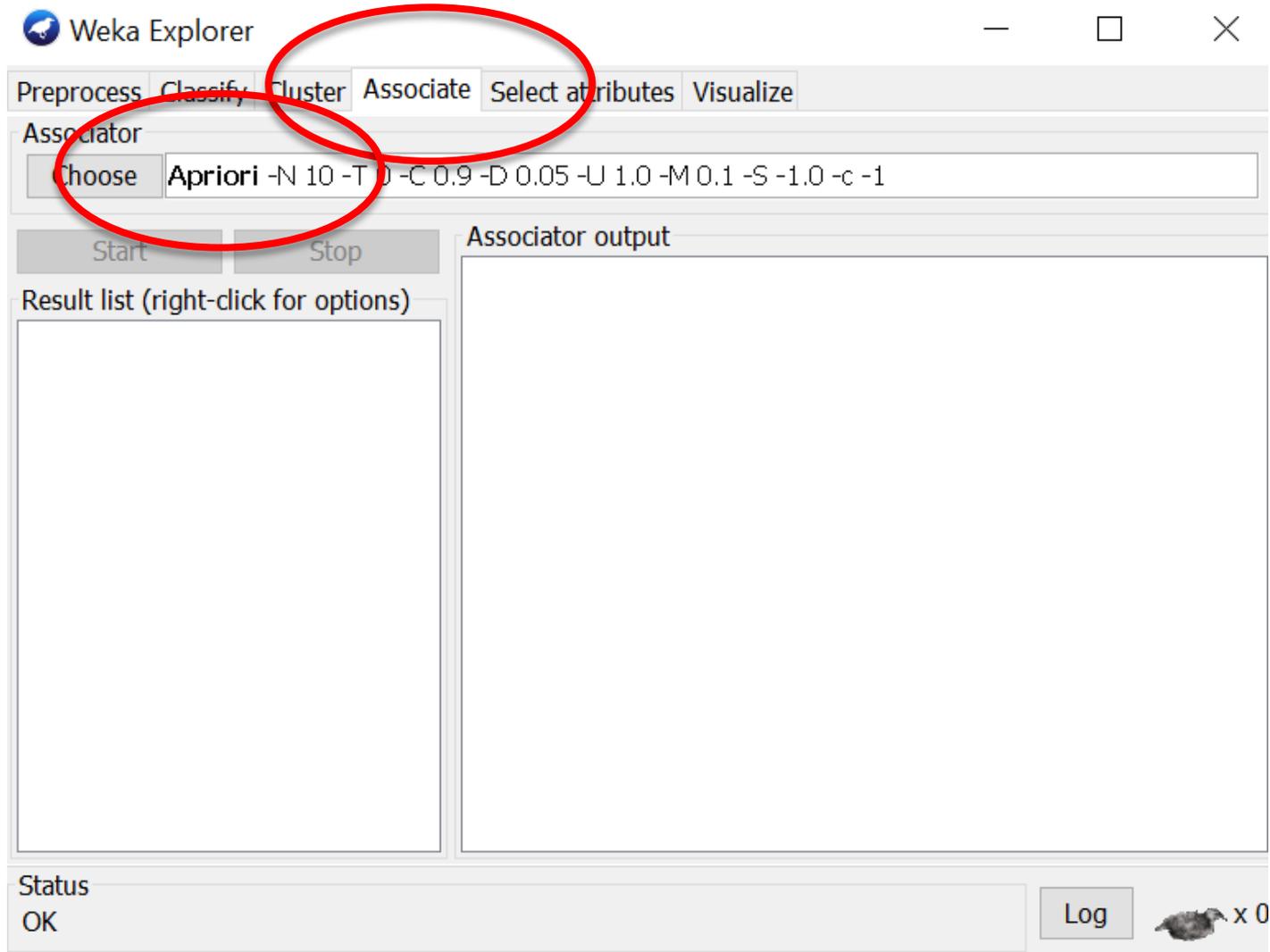
Applications

- Explorer
- Experimenter
- KnowledgeFlow
- Simple CLI

Weka – *Open File*



Weka – *Associate*



Hands On!

- Abra o arquivo: `_ASSOC00.csv`
- 1. Rode o APRIORI. Encontrou algum padrão. Qual?
- 2. Altere o parâmetro da confiança para 0.6. O que aconteceu? Quantos padrões foram encontrados?

Suporte e Confiança

Suporte, ou cobertura, é o primeiro filtro para se obter as regras. Ao definirmos um suporte mínimo, o algoritmo verifica, logo de início, se o suporte é satisfeito. Somente se verdadeiro o item é adicionado ao subconjunto de itens frequentes.

$$\textit{suporte}(A \Rightarrow B) = P(B|A) = \frac{A \cup B}{\textit{total}(T)}$$

Suporte e Confiança

A confiança demonstra o quanto podemos confiar em uma implicação de acordo com o suporte dos termos.

A confiança é um bom indicador de quão forte é a associação entre os itens.

$$\text{confiança}(A \Rightarrow B) = P(B|A) =$$

$$\frac{\text{suporte}(A \Rightarrow B)}{\text{suporte}(A)} = \frac{A \cup B}{\text{suporte}(A)}$$

Entendendo ...

- Dado o seguinte conjunto de dados

A {0,1,0,1,0,0,1,0,0,0}

B {0,0,1,1,1,1,0,0,1,1}

C {1,0,0,0,1,1,1,1,0,0}

D {1,1,1,0,1,0,1,1,1,0}

E {0,0,1,1,0,1,1,0,1,1}

Objetivo <- c(0,1,1,1,0,0,1,0,1,1)

- Poderíamos gerar uma regra: **A --> Objetivo 0.3 1.00**

→0.3 é o suporte e 1 é a confiança.

→Todas as vezes que A apareceu, o objetivo foi cumprido (100%), mas isso só ocorreu em 3 dos 10 casos (30%)

Lift

Um segundo parâmetro é o lift que determina a dependência e a correlação das regras. Acima de 1 indica correlação positiva, abaixo negativa.

$$\text{Lift}(A \Rightarrow B) = (B|A) = \frac{A \cup B}{A * B}$$

Para a regra anterior temos $0.3 / (0.3 * 0.6) = 1.666$

APRIORI com R

- Para termos um pouco mais de flexibilidade, podemos usar uma linguagem de programação.
- Podemos unir o pré-processamento com a execução de algoritmos.

Vamos a um exemplo...

Technical demo

- Rode o seguinte trecho de código:

```
a <- c(1,1,0,0,1)
b <- c(0,1,0,1,1)
c <- c(0,1,1,0,1)
goal <- c(1,0,1,0,1)
```

- No R, podemos usar `cbind()` “colar” colunas.

```
> myDF <- cbind(a,b,c,goal)
> myDF
      a b c goal
[1,] 1 0 0    1
[2,] 1 1 1    0
[3,] 0 0 1    1
[4,] 0 1 0    0
[5,] 1 1 1    1
```

Technical help

- Podemos usar a biblioteca `arules` para gerar regras de associação.

`library(arules)`

- Esta biblioteca possui uma implementação do APRIORI.
Os escopo é:
`apriori(data, parameter = NULL,
appearance = NULL, control = NULL)`

Technical help `apriori(data, parameter = NULL, appearance = NULL, control = NULL)`

- Podemos usar a biblioteca `arules` para gerar regras de associação.

`library(arules)`

- Dentro de `parameter` podemos usar o parâmetro `target` para especificar o tipo de associação a ser minerado. O APRIORI pode usar “`rules`”. `parameter` sempre recebe uma lista. ...Use `list(target = “rules”)`

Technical help

- A matriz de entrada também pode ser gerada através da função `read.transactions()`.
- Por exemplo, a linha a seguir carrega um *csv* contendo n colunas, onde `cols` é o número do ID da transação. `format="basket"` altera para o formato de entrada do algoritmo (APRIORI da biblioteca `arules`).

```
read.transactions(file="meuArquivo.csv", format="basket", sep=";", cols=1);
```

Hands On!

1. Rode o APRIORI para obter regras de associação no objeto myDF. A saída é um objeto que contém as especificações dos parâmetros e algumas estatísticas. Use `inspect()` para obter as regras dentro desse objeto.

Technical help

- A saída das regras possui 2 partes. A primeira (*lhs*) são os itens que implicam na saída (*rhs*).

Ex: Pao, queijo → presunto. Pão e queijo é a parte lhs e presunto é rhs

Controle a saída desejada usando os comandos `subset()` e `"%in%"`.

Ex: para o problema anterior podemos fazer:

```
Meu_subconjunto <- subset(regras, (lhs %in%  
"goal"))
```

Hands On!

2. Abra o arquivo “_ASSOC00.csv” no R. Use a biblioteca `arules` para gerar regras de associação.