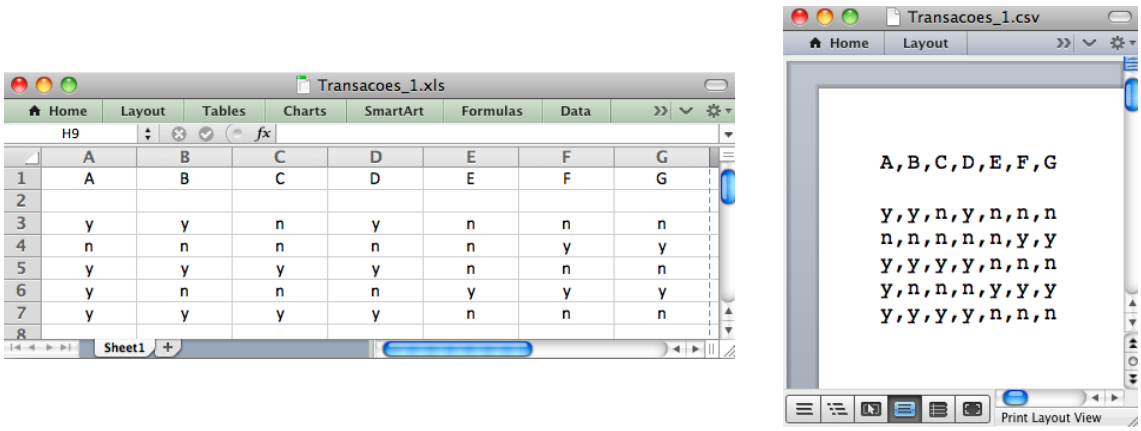
Sistemas Inteligentes – Aula 02

Mineração de Dados e Regras de Associação

Como Gerar Regras de Associação Usando a Ferramenta Weka

Nesta seção será apresentado um pequeno tutorial sobre a geração de Regras de Associação usando o algoritmo "Apriori" implementado na ferramenta de Aprendizado de Máquina para tarefas de Mineração de Dados Weka (Weka, 2013). A versão utilizada é a 3.6.7. Para fazer uma simulação no Weka, a Base de Dados terá de ser escrita ou no formato CSV *(Comma-Separated Value)* (".csv") ou no formato "ARFF" *(Attribute-Relation File Format)*, um formato bastante simples e intuitivo dessa ferramenta. Com o arquivo ".arff" carregado, podemos ajustar os parâmetros Suporte e Confiança e rodar o algoritmo Apriori.

Passo 1 - Vamos supor que nossa Base de Dados tenha sido retirada de uma planilha eletrônica (".xls") e salva no formato ".csv", como mostra a Fig. 2.1.



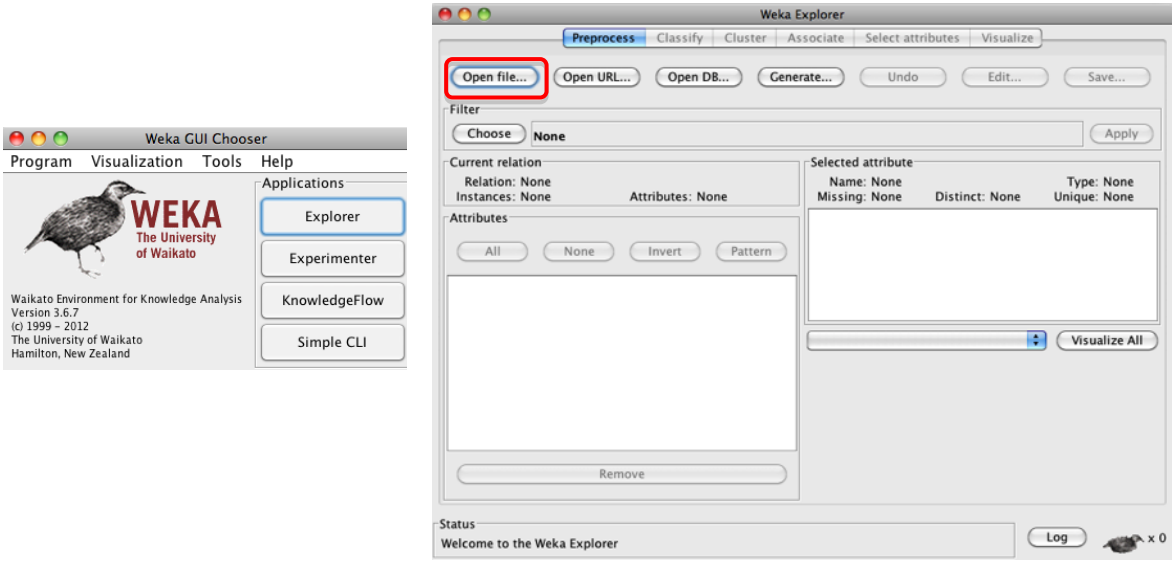
(a) **(b)**

Fig. 2.1 – A Base de Dados Transacoes_1 na Forma (a) Planilha ".xls" e (b) ".csv".

Como o Weka tem um conversor interno do formato ".csv" para ".arff", vamos primeiramente usar este recurso. Depois vamos mostrar como transformar manualmente o arquivo ".csv" para ".arff".

Obs.: Certifique-se que em seu arquivo ".csv" o separador de células seja efetivamente a vírgula "," e não ";". Se o arquivo ".csv" gerado pela sua planilha utilizar ";", faça a substituição para ", ". Caso contrário, ocorrerá um erro de leitura no Weka e o arquivo será interpretado de forma completamente diferente do esperado.

Passo 2 – Dispare o Weka ("GUI Chooser) e tome a opção "Explorer", que corresponde à versão com recursos gráficos e ícones (em vez de linha de comando). Veja Fig. 2.2.



(a) **(b)**

Fig. 2.2 – Telas Iniciais do Weka (a) GUI Chooser e (b) Explorer.

Passo 3 – Com a aba superior "Preprocess" escolhida, dê um clique em "Open file...". Uma janela denominada "Open" deve se abrir. Ajuste a opção de "File Format:" para ".csv", e escolha o arquivo "Transacoes_1.csv", conforme mostra a Fig. 2.3.

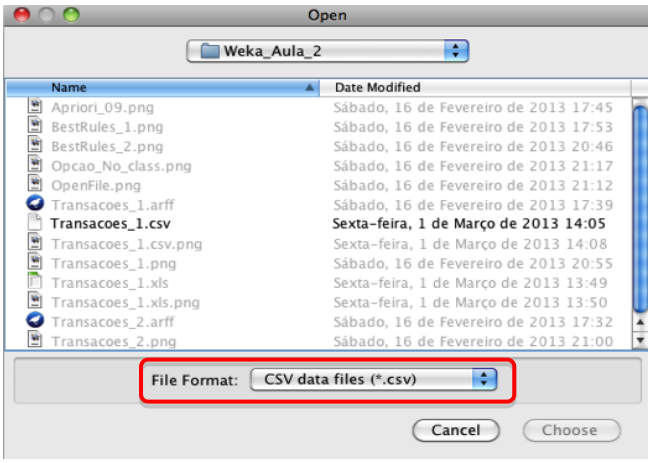


Fig. 2.3 – Janela "Open" com a Opção "File Format:" em ".csv".

Passo 4 – A tela do Weka Explorer deve apresentar os sete atributo do arquivo "Transacoes_1", como mostra a Fig. 2.4

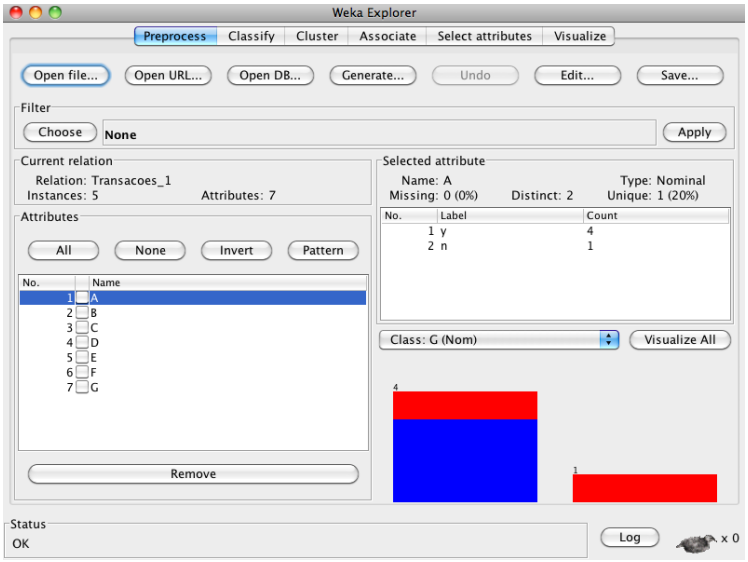


Fig. 2.4 – Os Sete Atributos do Arquivo "Transacoes_1" São Mostrados.

Passo 5 – Como nossa "Base de Dados" é muito pequena, a conversão manual do arquivo ".csv" para ".arff" pode ser feita muito rapidamente.

Digite no arquivo "Transacoes_1.csv" as palavras-chave "@relation", "@attribute" e "@data", de acordo com a Fig. 2.5, salve e feche o arquivo ".csv". Mude a terminação do arquivo de ".csv" para ".arff". Há ainda outras alternativas: Crie um arquivo "Transacoes_1.txt" com o conteúdo mostrado abaixo na Fig. 2.5 (certifique-se de que se trata efetivamente de arquivo tipo ".txt" e não, por ex., "Transacoes_1.txt.doc" ou "Transacoes_1.txt.rtf"). Feche o arquivo e mude a terminação para ".arff", ou seja, para "Transacoes_1.arff".

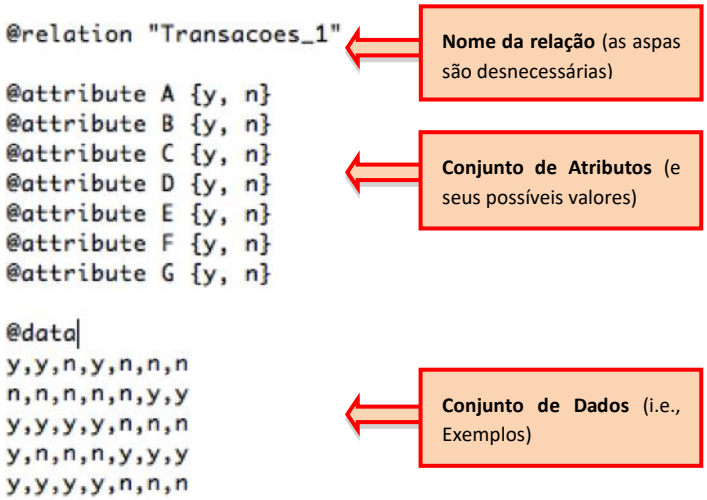


Fig. 2.5 – Arquivo ARFF (Transacoes_1.arff), com Itens Ausentes Representados por "n".

Passo 6 – Com o arquivo "Transacoes_1.arff" pronto, disparar o Weka, selecionar a aba "Preprocess", depois clicar na opção "Open file..." e escolher o arquivo "Transacoes1_.arff", conforme mostra a Fig. 2.6.

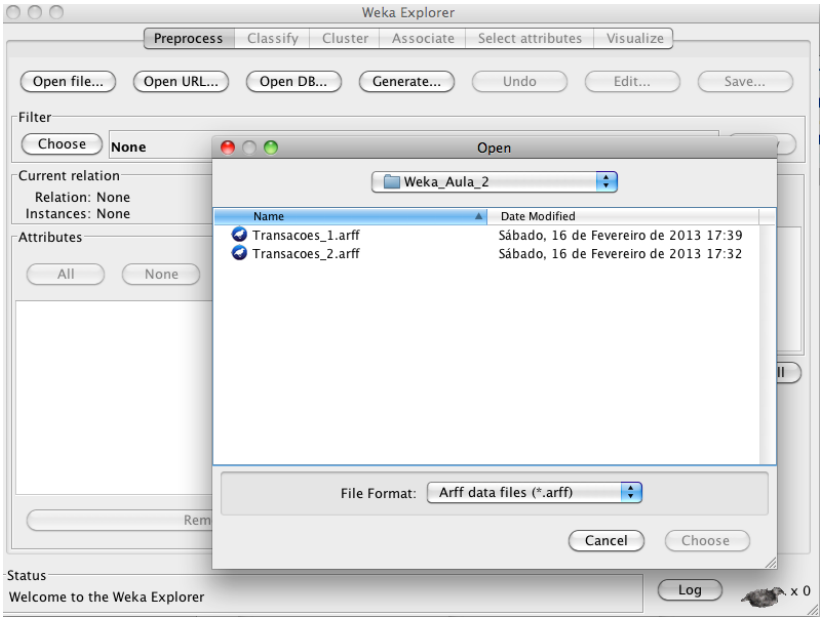


Fig. 2.6 – Aba "Preprocess" + "Open file..." para Escolha do Arquivo ARFF.

Passo 7 – Depois de abrir o arquivo "Transacoes_1.arff", ainda com a aba "Preprocess" selecionada, escolha "No class" (ao lado de "Visualize all"), conforme ilustra a Fig. 2.7. (Como vamos gerar Regras de Associação, qualquer um dos atributos pode funcionar como "classe". Este conceito vai ser melhor explicado quando formos estudar Regras de Classificação.)

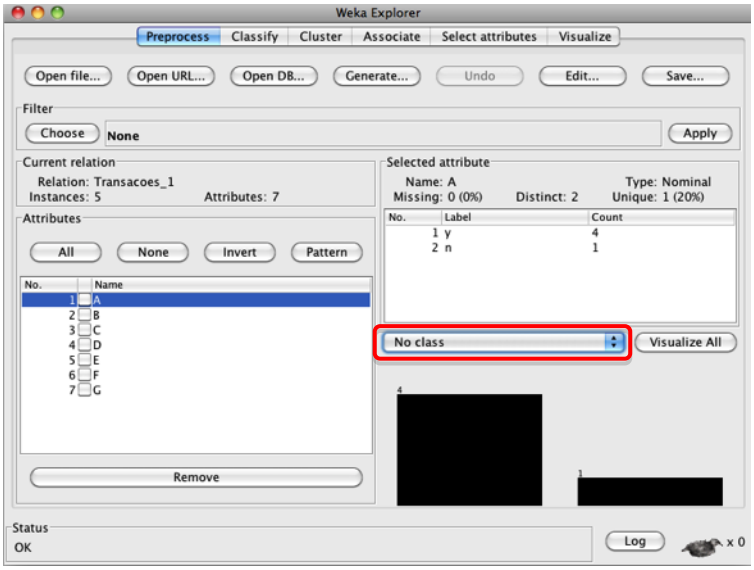


Fig. 2.7 – Seleção da Opção "No class" para Regras de Associação.

Passo 8 – Na aba superior do Weka, escolher “Associate” e ao lado de “Choose” clicar duas vezes sobre o algoritmo “Apriori”, conforme mostra a Fig. 2.8.

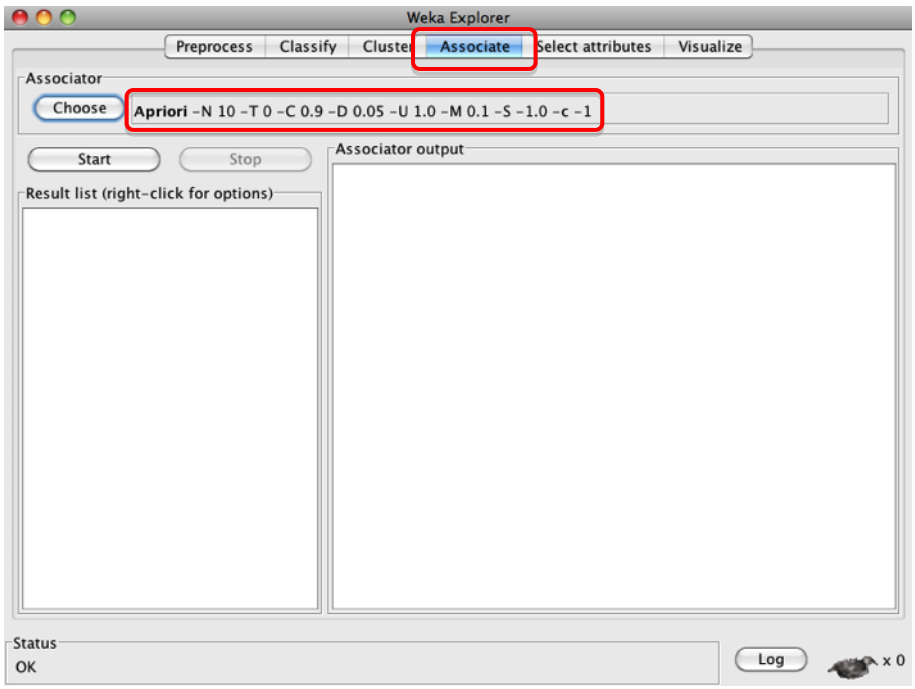


Fig. 2.8 – Ajuste dos Parâmetros de Entrada do Algoritmo Apriori.

Passo 9 – Na janela que se abre, ajustar o SupMin (lowerBoundMinSupport) para 0.4, a ConfMin (minMetric) para 0.9 e o número de regras mostradas (numRules) para 1000, conforme mostra a Fig. 2.9. Clicar em “OK”.

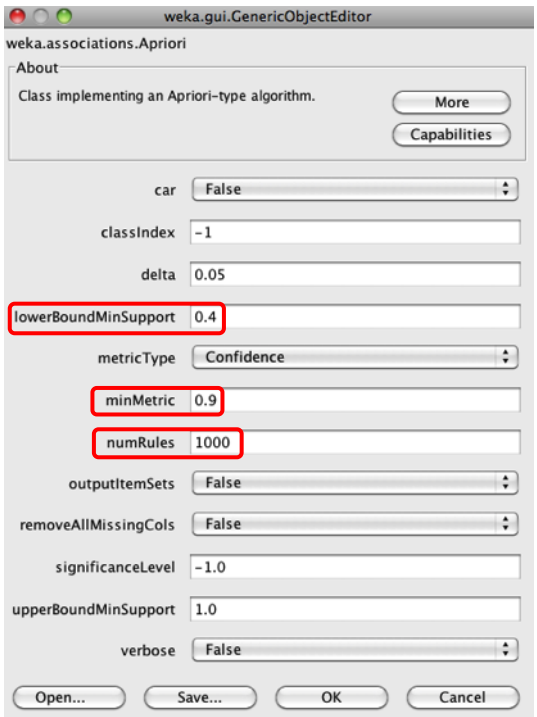


Fig. 2.9 – Ajuste dos Parâmetros SupMin e ConfMin.

Passo 10 – Ao clicar em "Start" centenas de Regras de Associação serão geradas, a maioria delas sem qualquer interesse, conforme ilustra a Fig. 2.10. Um dos riscos da geração de Regras de Associação é que muitas delas podem não ter qualquer significado prático. Para contornar este tipo de problema, é possível introduzir pequenas mudanças na forma como os atributos são declarados e reduzir significativamente o número de regras geradas.

```
Best rules found:

 1. B=y 3 ==> A=y 3    conf:(1)
 2. D=y 3 ==> A=y 3    conf:(1)
 3. F=n 3 ==> A=y 3    conf:(1)
 4. G=n 3 ==> A=y 3    conf:(1)
 5. D=y 3 ==> B=y 3    conf:(1)
 6. B=y 3 ==> D=y 3    conf:(1)
 7. B=y 3 ==> E=n 3    conf:(1)
 8. F=n 3 ==> B=y 3    conf:(1)
 9. B=y 3 ==> F=n 3    conf:(1)
10. G=n 3 ==> B=y 3    conf:(1)
11. B=y 3 ==> G=n 3    conf:(1)
12. D=y 3 ==> E=n 3    conf:(1)
13. F=n 3 ==> D=y 3    conf:(1)
14. D=y 3 ==> F=n 3    conf:(1)
15. G=n 3 ==> D=y 3    conf:(1)
16. D=y 3 ==> G=n 3    conf:(1)
17. F=n 3 ==> E=n 3    conf:(1)
18. G=n 3 ==> E=n 3    conf:(1)
19. G=n 3 ==> F=n 3    conf:(1)
20. F=n 3 ==> G=n 3    conf:(1)
```

Fig. 2.10 – Algumas Regras de Associação Geradas com o Arquivo "Transacoes_1.arff".

Passo 11 – Uma forma de diminuir o número de regras é substituir os valores ausentes de atributo "n" por "?". Crie um arquivo "Transacoes_2.arff" conforme mostra a Fig. 2.11.

```
@relation "Transacoes_2"

@attribute A {y, n}
@attribute B {y, n}
@attribute C {y, n}
@attribute D {y, n}
@attribute E {y, n}
@attribute F {y, n}
@attribute G {y, n}

@data
y,y,?,y,?,?,?
?,?,?,?,?,y,y
y,y,y,y,?,?,?
y,?,?,?,y,y,y
y,y,y,y,?,?,?
```

Fig. 2.11 – Arquivo "Transacoes_2.arff" com Itens Ausentes Representados por "?".

Isso vai evitar que o Weka crie regras sem qualquer significado prático envolvendo itens ausentes, como por exemplo, {F=n} ⇒ {G=n} (Regra 20 na Fig. 2.10). Embora a regra {F=y} ⇒ {G=y} (i.e., "quem compra queijo também costuma comprar vinho") possa ser de interesse, a regra de que "quem não compra queijo também não compra vinho") dificilmente trará alguma informação prática. Numa Base de Dados muito grande, regras desse tipo podem aparecer em quantidades proibitivamente grandes.

Com o arquivo "Transacoes_2.arff" foram geradas 30 Regras de Associação (Fig. 2.12), sendo que as regras ilustrativas do texto de teoria da Aula 2 envolvendo o CF = {A, B, C} aparecem na Fig. 2.12 como as regras 15, 16 e 17.

```
Minimum support: 0.4 (2 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 12

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 7

Size of set of large itemsets L(3): 4

Size of set of large itemsets L(4): 1

Best rules found:

 1. B=y 3 ==> A=y 3    conf:(1)
 2. D=y 3 ==> A=y 3    conf:(1)
 3. D=y 3 ==> B=y 3    conf:(1)
 4. B=y 3 ==> D=y 3    conf:(1)
 5. B=y D=y 3 ==> A=y 3    conf:(1)
 6. A=y D=y 3 ==> B=y 3    conf:(1)
 7. A=y B=y 3 ==> D=y 3    conf:(1)
 8. D=y 3 ==> A=y B=y 3    conf:(1)
 9. B=y 3 ==> A=y D=y 3    conf:(1)
10. C=y 2 ==> A=y 2    conf:(1)
11. C=y 2 ==> B=y 2    conf:(1)
12. C=y 2 ==> D=y 2    conf:(1)
13. G=y 2 ==> F=y 2    conf:(1)
14. F=y 2 ==> G=y 2    conf:(1)
15. B=y C=y 2 ==> A=y 2    conf:(1)
16. A=y C=y 2 ==> B=y 2    conf:(1)
17. C=y 2 ==> A=y B=y 2    conf:(1)
18. C=y D=y 2 ==> A=y 2    conf:(1)
19. A=y C=y 2 ==> D=y 2    conf:(1)
20. C=y 2 ==> A=y D=y 2    conf:(1)
21. C=y D=y 2 ==> B=y 2    conf:(1)
22. B=y C=y 2 ==> D=y 2    conf:(1)
23. C=y 2 ==> B=y D=y 2    conf:(1)
24. B=y C=y D=y 2 ==> A=y 2    conf:(1)
25. A=y C=y D=y 2 ==> B=y 2    conf:(1)
26. A=y B=y C=y 2 ==> D=y 2    conf:(1)
27. C=y D=y 2 ==> A=y B=y 2    conf:(1)
28. B=y C=y 2 ==> A=y D=y 2    conf:(1)
29. A=y C=y 2 ==> B=y D=y 2    conf:(1)
30. C=y 2 ==> A=y B=y D=y 2    conf:(1)
```

Fig. 2.12 – As 30 Regras de Associação Geradas com o Arquivo "Transacoes_2.arff".

Há outras formas de melhorar a qualidade dos resultados e controlar o número de regras geradas, por exemplo, através do parâmetro ***Lift***, cujo significado fica como lição de casa.

Referência Bibliográfica

Weka. The Waikato University. In http://www.cs.waikato.ac.nz/ml/weka/ . Acessado em 03.03.13.

WITTEN, I. H. & FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques.** Second Edition. Amsterdam: Morgan Kaufmann Publishers, 2005.