

Fundamentos de Sistemas Inteligentes: Atividade com o Weka

Prof. Arnaldo Candido Junior
UTFPR – Medianeira

Instalação

- <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>
 - Download no link “Click here to download a self-extracting executable without the Java VM”



Arquivos adicionais

- <http://saci-devel.ufscar.br/weka/>
 - bag.arff
 - complexidade.arff
 - corpus.zip
 - olimpiadas.arff

Mineração de textos ₍₂₎

- Estratégia comum ***bag of words*** (saco de palavras – conjunto de palavras): cada palavra virá um atributo
- Problema de exemplo: classificar textos entre os gêneros esportivo, biografia e humorístico

nasceu	jogador	papagaio	(...)	classes
1	0	0	(...)	??
0	2	0	(...)	??
1	0	3	(...)	??

Arquivos arff

- Arquivos de entrada do WEKA
- Tem duas partes:
 - Cabeçalho
 - Nome da relação
 - Lista de atributos com os tipos
 - Dados
 - Dados separados por vírgula e seguindo a ordem em que os atributos são definidos no cabeçalho

Arquivos arff ₍₂₎

Cabeçalho

@RELATION olimpiadas

@ATTRIBUTE tamanho REAL

@ATTRIBUTE peso REAL

@ATTRIBUTE class {basquete,levantamento}

Dados

@DATA

1.79,88,basquete

1.86,94,levantamento

1.56,56,levantamento

2.05,106,basquete

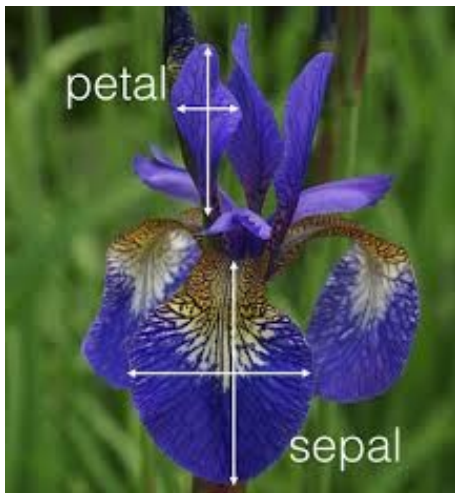
1.83,145,levantamento

1.95,89,basquete

Arquivos arff₍₃₎

- Os atributos podem ser dos tipos:
 - Numeric (inteiros ou reais)
 - <nominal-specification> (classe → entre { })
 - String
 - Date [<date-format>]

iris.arff



- Setosa



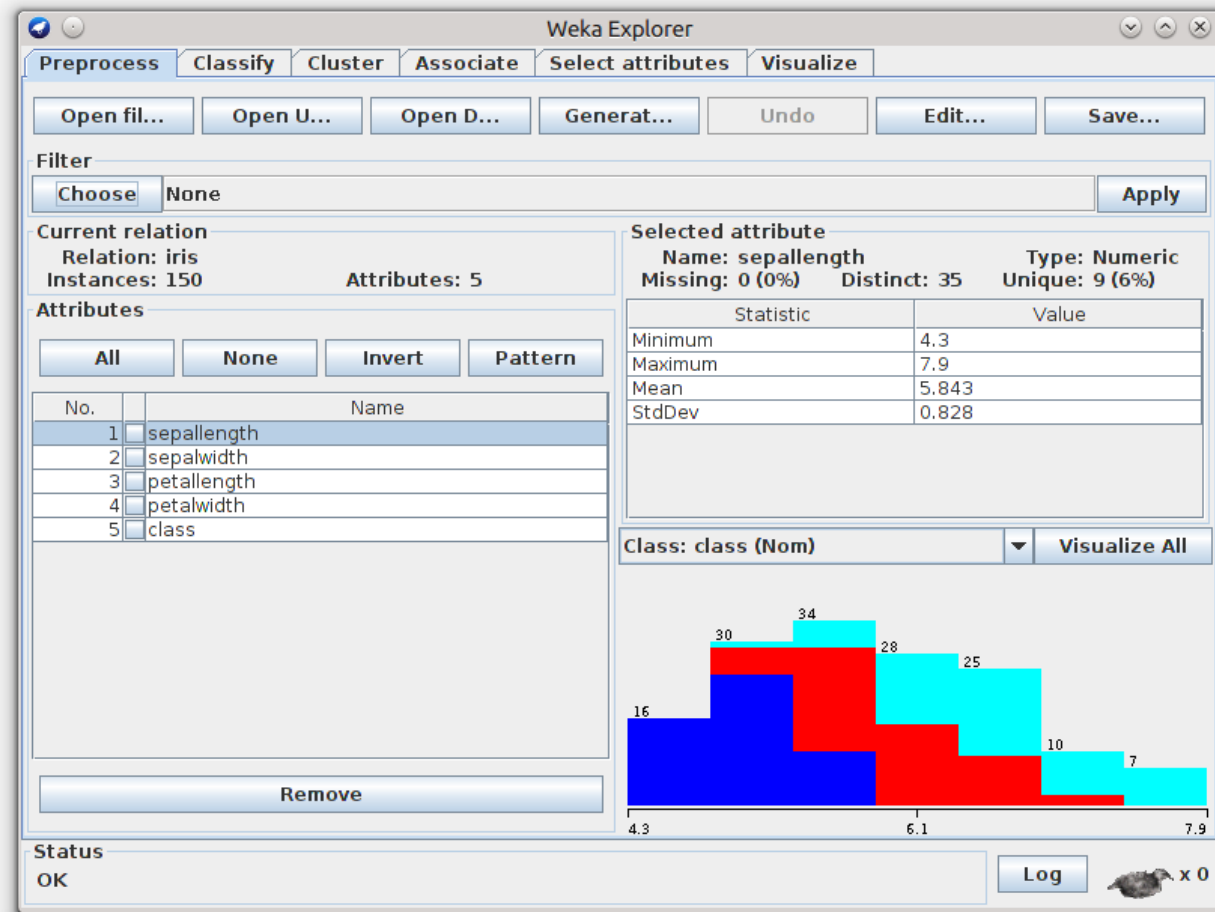
- Versicolor



- Virginica

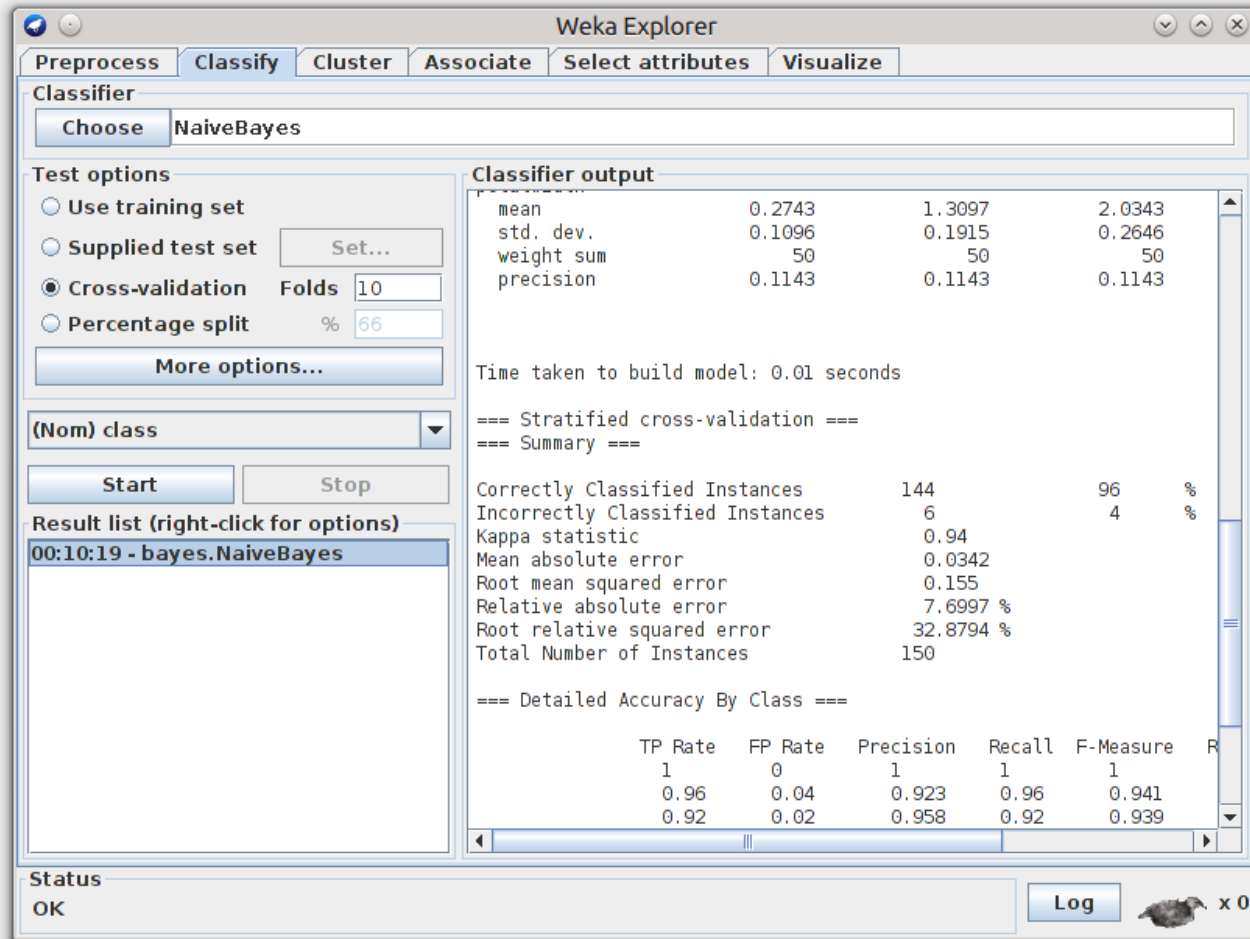
iris.arff (2)

- Abra o Weka
- Clique em “Explorer”
- Open file
- Data
- Iris.arff
- Clique nos campos e compare com o gráfico
 - Qual atributo é um bom separador?



iris.arff: classificação

- Treine os algoritmos a seguir
 - J48 (árvore de decisão) – visualize a árvore
 - MLP (mude os parâmetros da rede)
 - SMO (SVM)
 - Naive Bayes



The screenshot shows the Weka Explorer window with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'NaiveBayes'. Under 'Test options', 'Cross-validation' is selected with 'Folds' set to 10. The 'Result list' shows '00:10:19 - bayes.NaiveBayes' selected. The 'Classifier output' pane displays the following statistics:

mean	0.2743	1.3097	2.0343
std. dev.	0.1096	0.1915	0.2646
weight sum	50	50	50
precision	0.1143	0.1143	0.1143

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	144	96	%
Incorrectly Classified Instances	6	4	%
Kappa statistic	0.94		
Mean absolute error	0.0342		
Root mean squared error	0.155		
Relative absolute error	7.6997 %		
Root relative squared error	32.8794 %		
Total Number of Instances	150		

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	R
1	1	0	1	1	1	
0.96	0.04	0.923	0.96	0.941		
0.92	0.02	0.958	0.92	0.939		

Status: OK

Log x 0

iris.arff: classificação ₍₂₎

- Analise as medidas de desempenho
- Visualiza a árvore de decisão
- Mude os parâmetros da rede neural
 - Veja a topologia gerada

iris.arff: classificação ₍₃₎

- Compare
 - Cross validation
 - Holdout
 - Avaliação usando conjunto de treinamento

iris.arff: classificação ₍₄₎

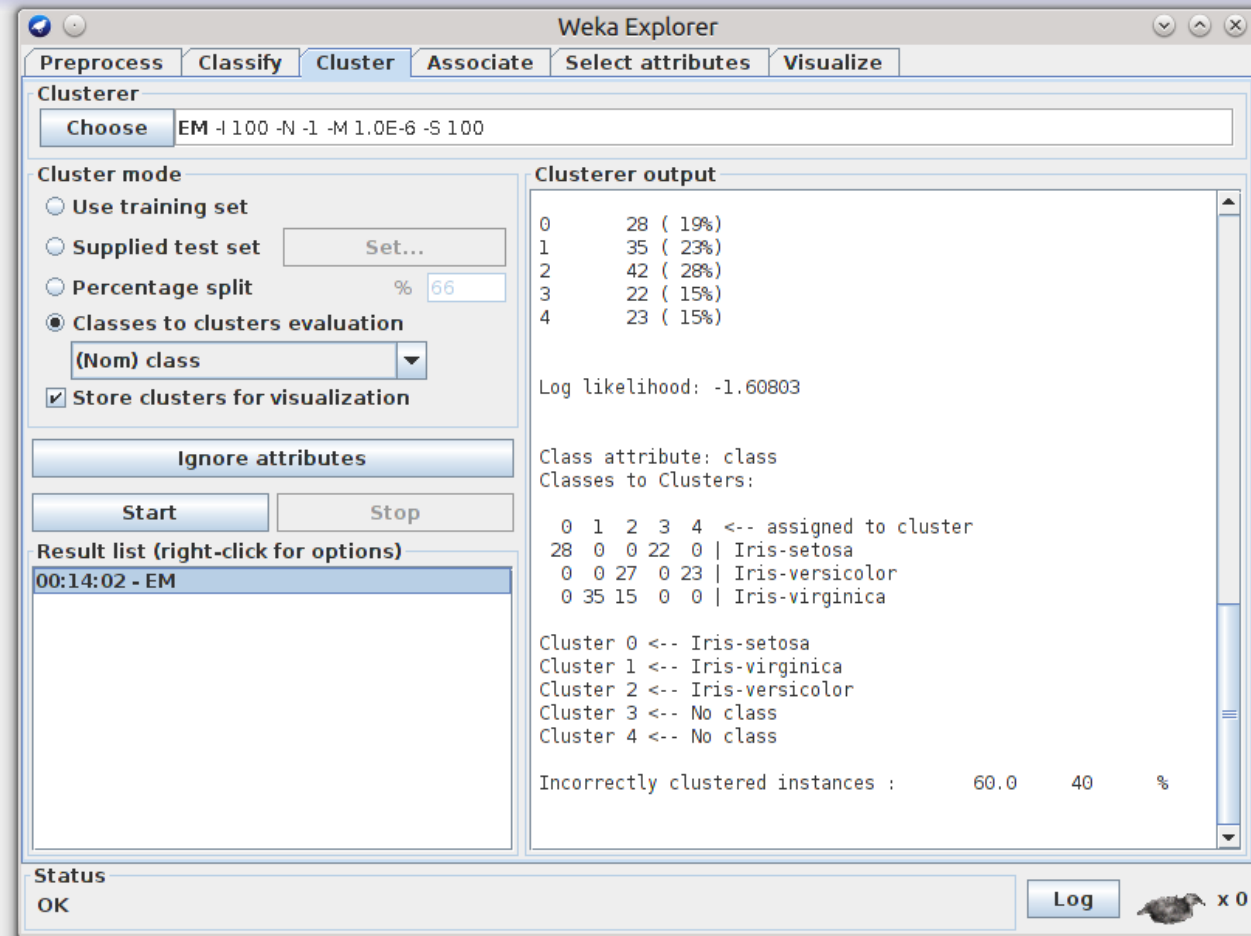
- Outros algoritmos interessantes:
 - Rules.Part
 - Function.Logistic
 - Lazy.IBK (KNN)
 - Meta.Adabost com SMO/IBK/MLP

iris.arff: regressão

- Remova o atributo classe
- Escolha sepalwidth como classe
- O algoritmo SimpleLinearRegression ficará disponível
- Execute
- Ao fim, abra de novo iris.arff

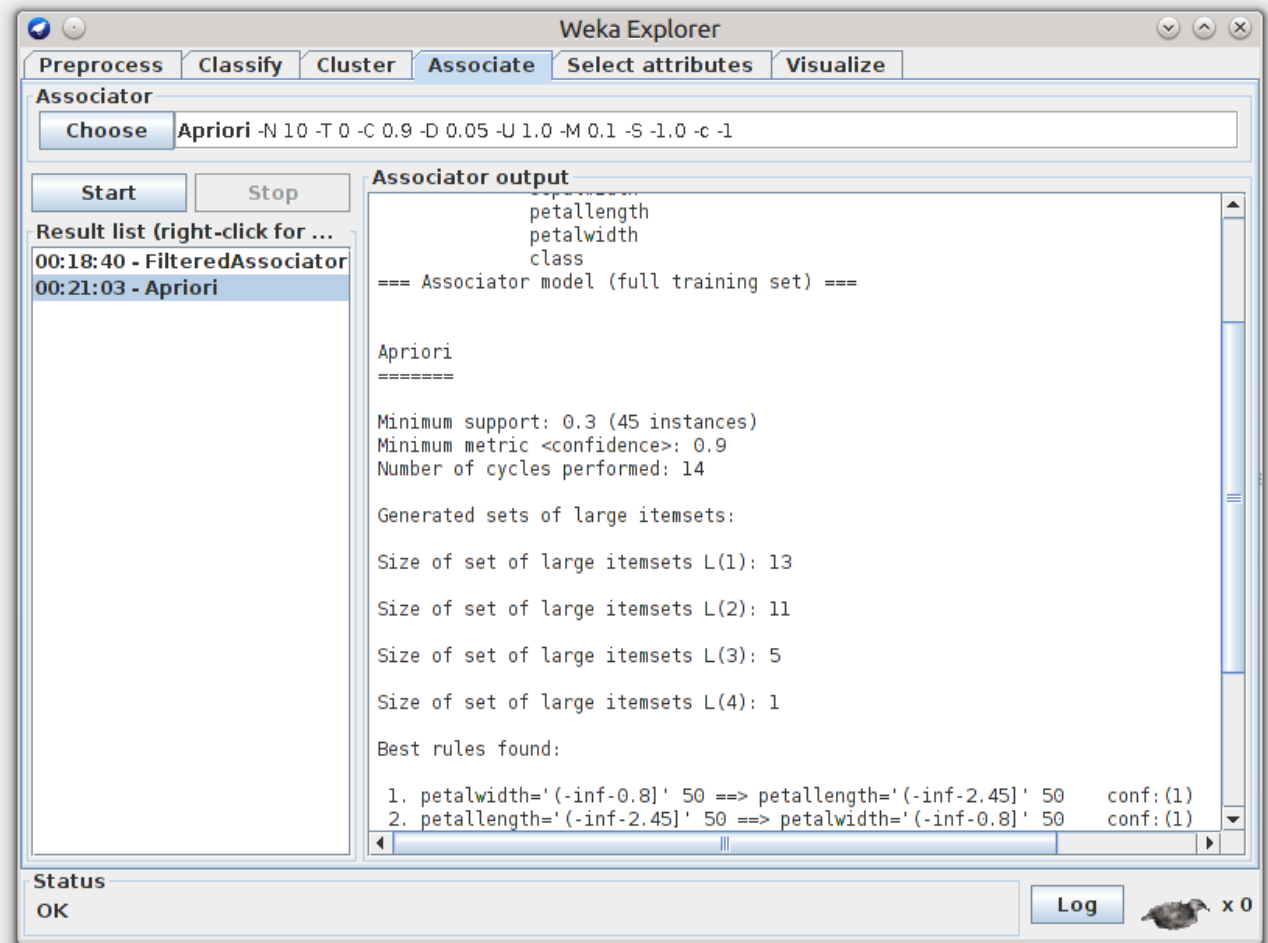
iris.arff: agrupamento

- Teste
 - EM
 - Hierarquical clusterer
- Para EM
 - Avalie clusteres com classes verdadeiras



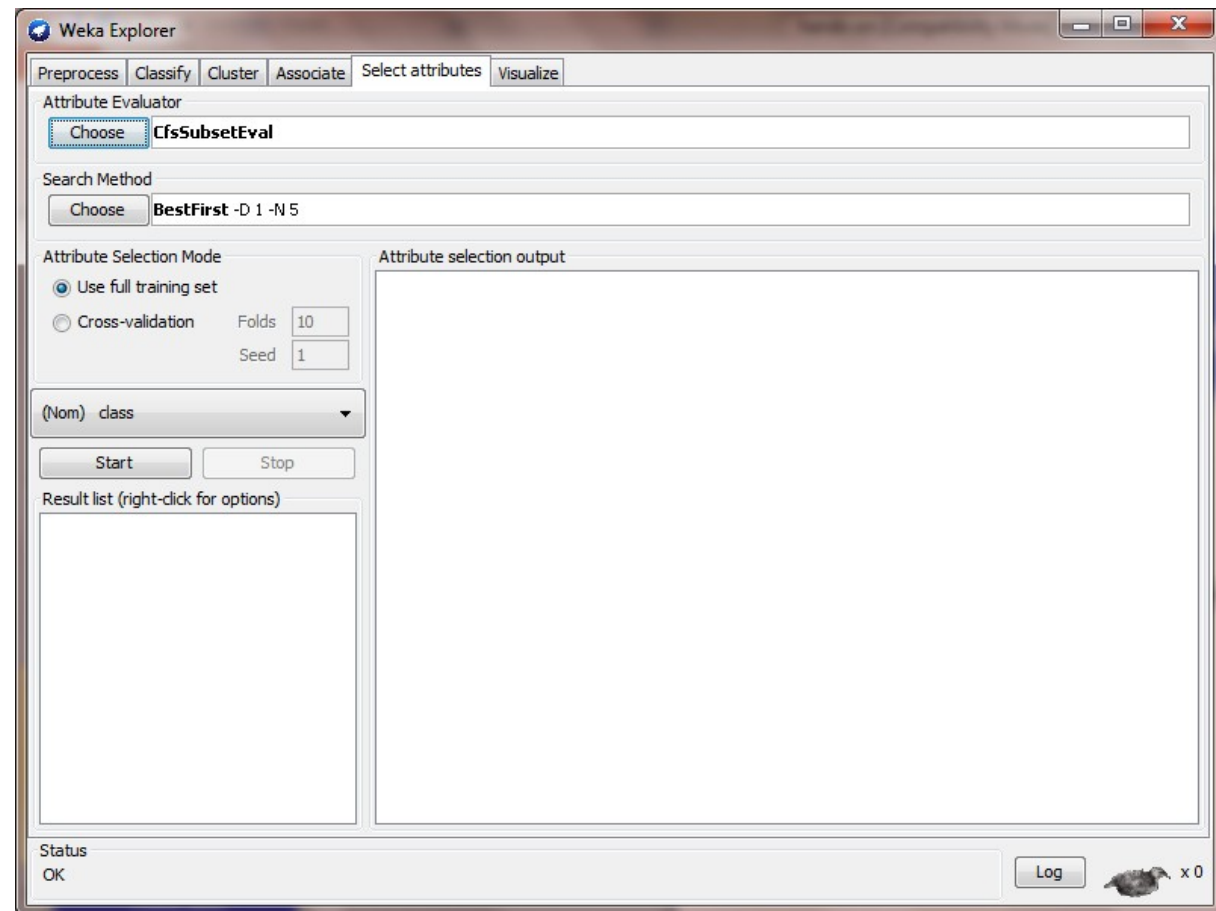
iris.arff (5)

- Aba preprocess → aplicar filtro → discretize
- Aba associate → apriori



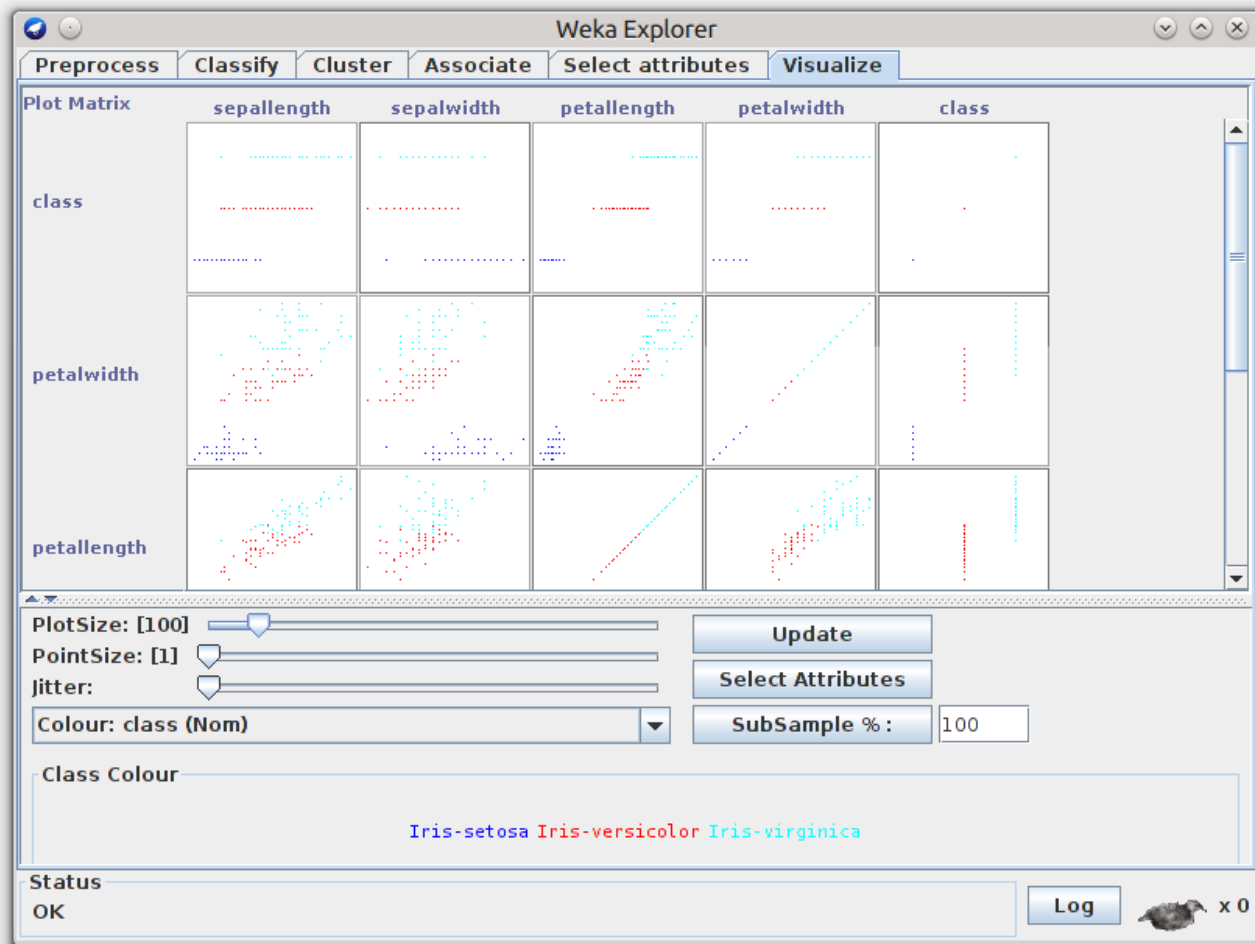
iris.arff (6)

- Aba select attributes
- Choose → cfs sub set eval
- Start
- Quais são os atributos mais relevantes?



iris.arff₍₇₎

- Aba visualize
- Compare atributos dois a dois
- Teste com diferentes plot-sizes e gitters



complexidade.arff

- Arquivo: complexidade.arff

Córpus	Número de textos	Número de palavras	Média de palavras por textos	Classe
ZH – jornalístico	166	63996	385,518	Complexo
CH – divulgação científica	130	81139	624,146	Complexo
PSFL – jornalístico	166	19257	116,006	Simple
CHC – divulgação científica	127	56096	441,701	Simple

complexidade.arff₍₂₎

- Abra o arquivo complexidade.arff
- Gere uma árvore de decisão visualize-a (J48)
- Gere uma máquina de vetor suporte (SVM)

wikipedia.arff

- Artigos de capa do ano de 2012
- 13 domínios: arte, biografias, ciências exatas, ciências da natureza, ciências sociais, cultura e sociedade, desporto, geografia, história, literatura, musica, religião, tecnologia
- Cada classe com 12 textos (total: 156)

wikipedia.arff₍₂₎

- Descompactar corpus.zip
- Weka → knowledge flow
- Data Sources → Text Directory Loader (clique duplo) → espaço em branco (clique duplo)
 - Clique duplo no campo que aparecer
 - Selecionar diretório de entrada (corpus). Obs: pode ser necessário digitar manualmente

wikipedia.arff₍₃₎

- Data Sink → Arff Saver
 - Mesmo processo
 - Apontar arquivo de saída wikipedia.arff
- Botão direito no TextDirectoryLoader criado → dataset
- Arrastar setting até o ArffSaver
- **Importante:** TextDirectoryLoader está com erro em em diversas versões do Weka. Em caso de problemas, usa a linha de comando (a seguir).

wikipedia.arff ₍₄₎

- Prompt de comandos:
 - `cd corpus`
 - `java -cp $diretorio_weka/weka.jar weka.core.converters.TextDirectoryLoader -dir text_example > wikipedia.arff`
 - `java -Xms1000m -Xmx1000m weka.jar`
- Após conversão: aplicar filtro unsupervised → attribute → string to word vector
 - Salvar arff como bag.arff

bag.arff

- Vamos gerar a partir de wikipedia.arff
- Abrir wikipedia.arff
- Na aba preprocess
- Filter → Choose → weka/filters/unsupervised/string to word vector
- Abaixo de “selected attribute” manter:
Class: @@class@@ (noun)
- Clicar em “apply”

bag.arff₍₂₎

- Remover campos 2 até 49
- Salvar resultado como bag.arff
- Aba classify,
- Abaixo de “More Options”:
@@class@@ (noun)
- Rodar J48 e SVM
 - Visualizar árvore gerada
 - Ela faz sentido?