INTRODUÇÃO AO WEKA

Estagiário PAE: Pablo Andretta Jaskowiak
Professor: Ricardo J. G. B. Campello

SCC0173 | Mineração de Dados Biológicos

Créditos

- O material a seguir consiste de adaptações e extensões dos originais cedidos gentilmente por:
  - Prof. Dr. André C. P. L. F. Carvalho
  - Thiago F. Covões

Sumário

Introdução

- Mineração de Dados
  - Escrever código do zero
  - Reaproveitar código existente
  - Utilizar um ambiente dedicado
    - Ferramentas apropriadas para cada etapa do processo
    - Menor esforço por parte do usuário
    - Agilidade na realização de experimentos

Introdução

- Pássaro típico da Nova Zelândia

Introdução

- **W**aikato **E**nvironment for **K**nowledge **A**nalysis
- Desenvolvido na Universidade de Waikato, NZ
- Mineração de dados e Aprendizado de Máquina
- Escrito na linguagem JAVA
  - Distribuído sob a licença GPL (GNU Public License)

Introdução

- Versão atual
  - WEKA 3.7 (instável)
  - WEKA 3.6 (estável)
- Versão do livro
  - WEKA 3.4

Witten, I. H., Frank E., Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, 2005.

Introdução

- Características
  - Diversos algoritmos disponíveis
  - Software livre
  - Independente de plataforma
  - Fácil utilização
  - Atualizado freqüentemente
    - Adição de novos algoritmos e funcionalidades

Funcionalidades

- Possui módulos para:
  - Pré-processamento
  - Uso de algoritmos de AM/MD
  - Visualização de dados, resultados e modelos
  - Comparação de modelos e algoritmos

Interfaces

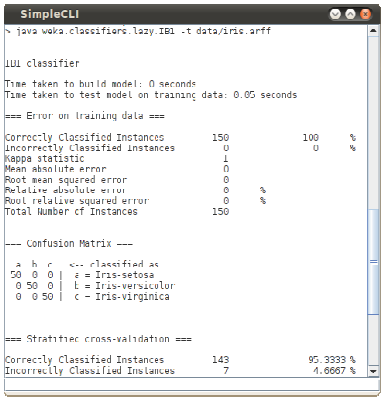
- Quatro principais

Interfaces

- Simple CLI (*Command Line Interface*)
  - Antiga
  - Utilização em sistemas sem interface gráfica
  - Simples e útil
  - Requer maior intimidade

Interfaces

- Simple CLI

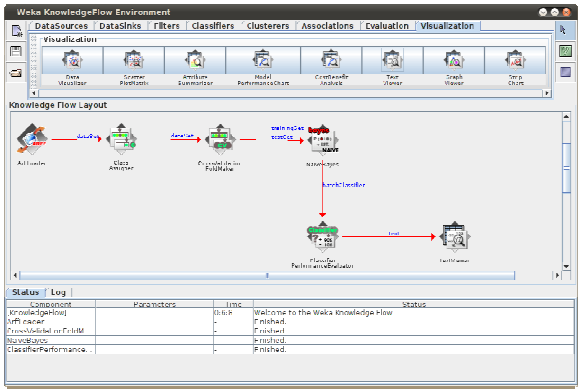


Interfaces

- KnowledgeFlow
  - Interface *drag-and-drop*
  - Trabalha com fluxo de dados
  - Descrição visual do processo
  - Tem sido aprimorada recentemente

Interfaces

- KnowledgeFlow

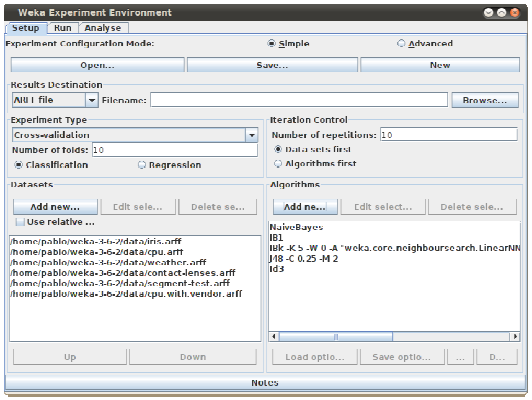


Interfaces

- Experimenter
  - Realização dos experimentos em modo *batch*
  - Comparação de algoritmos
  - Distribuição dos experimentos
  - Prático quando são utilizados
    - Diversos algoritmos
    - Diversas bases de dados

Interfaces

- Experimenter

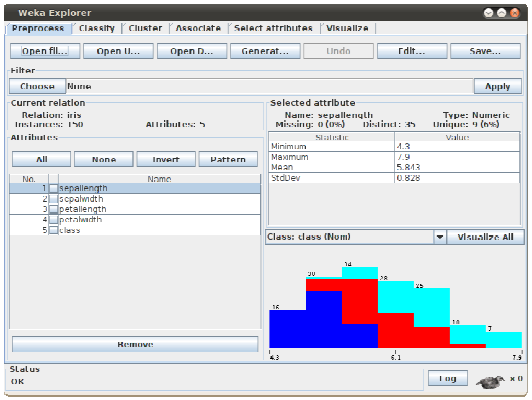


Interfaces

- Explorer
  - Exploração de dados
  - Rápido e prático
  - Mesmas funcionalidades do KnowledgeFlow
    - Sem descrição visual do processo

Interfaces

- Explorer



Formato dos Dados

- Atributos
  - Numéricos
  - Nominais
- Formato padrão
  - ARFF – *Attribute-Relation File Format*
- Suporta também alguns outros formatos
  - CSV
- Maiores informações
  - http://weka.wikispaces.com/ARFF

Formato dos Dados

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall
%      (c) Date: July, 1988

@RELATION iris

@ATTRIBUTE sepallength  REAL
@ATTRIBUTE sepalwidth   REAL
@ATTRIBUTE petallength  REAL
@ATTRIBUTE petalwidth   REAL
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
6.2,2.9,4.3,1.3,Iris-versicolor
5.1,2.5,3.0,1.1,Iris-versicolor
5.7,2.8,4.1,1.3,Iris-versicolor
6.3,3.3,6.0,2.5,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
7.1,3.0,5.9,2.1,Iris-virginica
```

Documentação

- Documentação *online*
  - Site Oficial
    - http://www.cs.waikato.ac.nz/ml/weka/
  - Javadoc
    - http://www.opendocs.net/javadoc/weka/
  - Wiki
    - http://weka.wikispaces.com/

Noções Básicas de Uso

- Utilização da interface **Explorer**
  - Visualizando bases ARFF
  - Carregando bases de dados
  - Pré-processamento
    - Discretização de atributos
    - Normalização de atributos
    - Amostragem de dados
  - Visualizando os dados
    - Avaliação visual de atributos

Noções Básicas de Uso

- Utilização da interface **Explorer**
  - Aplicando algoritmos de classificação
    - 1Rule
    - kNN
      - Ponderação da distância
    - Naïve Bayes
    - Árvores de decisão
      - Visualização das árvores resultantes
  - Gerando curvas ROC

Noções Básicas de Uso

- Utilização da interface **KnowledgeFlow**

Dúvidas?