Universidade Federal de São Carlos - UFSCar
Departamento de Computação - DC
Programa de Pós-Graduação em Ciência da Computação - PPGCC

Ambiente Weka

Waikato Environment for Knowledge Analysis

Classificação Textual

Aluno: Pablo Freire Matos
Orientador: Dr. Ricardo Rodrigues Ciferri
Coorientador: Dr. Thiago Alexandre S. Pardo
Área: Banco de Dados

O que é Weka?

- Coleção de algoritmo de AM para tarefas de MD
  - Implementado em Java
- Algoritmos de classificação:
  - Naive Bayes, SVM, Árvore de Decisão, Regras de Classificação....
- 3 modos de execução
  - GUI
  - Linha de Comando
  - Java API

Arquivo ARFF (*Attribute-Relation File Format*)

- **@relation** <nome-relação>

  @relation AnemiaFalciforme

- **@attribute** <nome-atributo> <tipo-de-dados>
  - Tipo de dados pode ser *numeric, nominal, string* ou *date*

  @attribute “termos” numeric

  @attribute class {complication,benefit,other}

- **@data**
  - Valores perdidos são representados por **?**

  @data

  1,0,0,complication

Formato .ARFF

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Formato .ARFF – Atributo Numérico

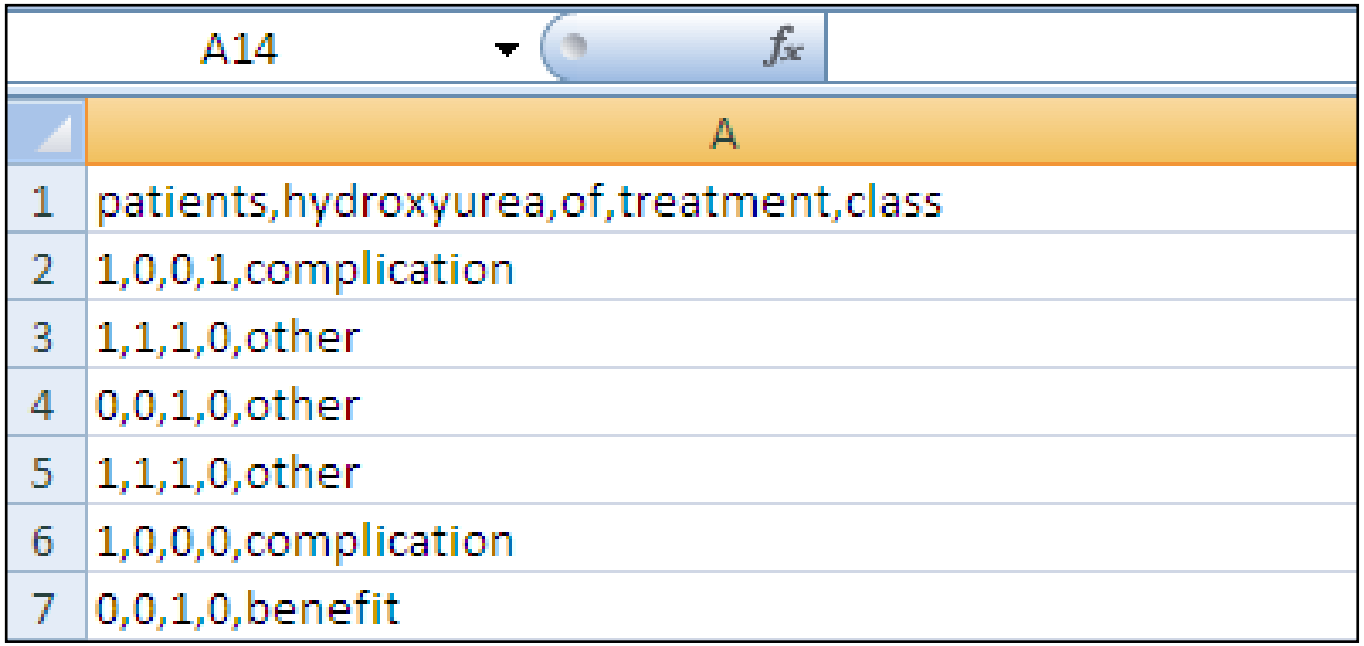
```
% Number of Instances: 6
% Number of Attributes: 5
% 0 - atributo NÃO pertence à sentença
% 1 - atributo pertence à sentença

@relation AnemiaFalciforme

@attribute patients numeric
@attribute hydroxyurea numeric
@attribute of numeric
@attribute treatment numeric
@attribute class {complication,benefit,other}

@data
1,0,0,1,complication
1,1,1,0,other
0,0,1,0,other
1,1,1,0,other
1,0,0,0,complication
0,0,1,0,benefit
```

Formato .ARFF – Atributo Nominal

```
% Number of Instances: 6
% Number of Attributes: 5
% 0 - atributo NÃO pertence à sentença
% 1 - atributo pertence à sentença

@relation AnemiaFalciforme

@attribute patients {0,1}
@attribute hydroxyurea {0,1}
@attribute of {0,1}
@attribute treatment {0,1}
@attribute class {complication,benefit,other}

@data
1,0,0,1,complication
1,1,1,0,other
0,0,1,0,other
1,1,1,0,other
1,0,0,0,complication
0,0,1,0,benefit
```

Formato .CSV

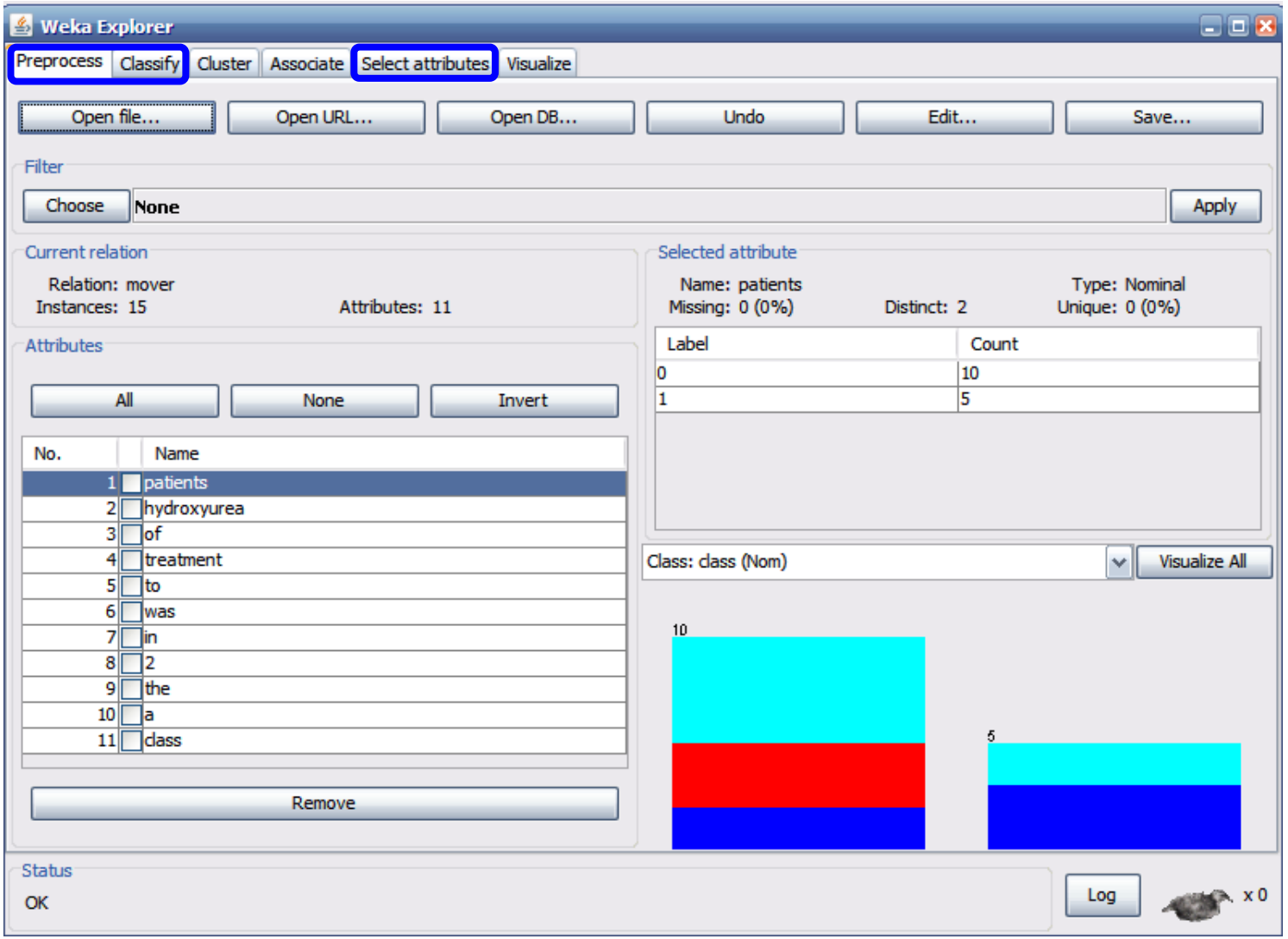
- .CSV (*Comma-separated values*)



A14

	A
1	patients,hydroxyurea,of,treatment,class
2	1,0,0,1,complication
3	1,1,1,0,other
4	0,0,1,0,other
5	1,1,1,0,other
6	1,0,0,0,complication
7	0,0,1,0,benefit

Classificação Textual



Classificação Textual

1ª Fase
Pré-Processamento

1. Exclusão das *stops words*
   A. import weka.core.Stopwords

2. Padronização: Aplicar algoritmo de *stemming* de Porter (1980)
   A. import org.tartarus.snowball.*

3. Cálculo da frequência das palavras (1 até 5 gramas)

Classificação Textual

1ª Fase
Pré-Processamento

4. Criação da Tabela Atributo-Valor
   A. *Frequência das n-gramas*
   B. *Bag of words*

	Atributos					
Sentença		t_1	t_2	…	t_m	Classe
	s_1	a_{11}	a_{12}	…	a_{1m}	c_1
	s_2	a_{21}	a_{22}	…	a_{2m}	c_2
	s_3	a_{31}	a_{32}	…	a_{3m}	c_3
	…	…	…	…	…	…
	s_n	a_{n1}	a_{n2}	…	a_{nm}	c_n

Classificação Textual

1ª Fase
Pré-Processamento

4. Criação da Tabela Atributo-Valor

Exemplo

	Atributos					
Sentenças		parvovirus	infection	crises	hu	Classe
	Sentença1	1	1	0	0	Complication
	Sentença2	0	0	1	1	Benefit
	Sentença3	0	0	0	1	Other
1 = contém; 0 = não contém						

Classificação Textual

1ª Fase
Pré-Processamento

4. Criação da Tabela Atributo-Valor

Exemplo: Sentenças e Atributos

- **Complication**
  - In six patients **parvovirus** B19 **infection** developed during treatment;
- **Benefit**
  - **HU** therapy can ameliorate the clinical course of sickle cell anemia in some adults with three or more painful **crises** per year;
- **Other**
  - Mean daily doses of **HU** as well as hematologic response to **HU** were similar to that of the whole cohort.

Classificação Textual

1ª Fase
Pré-Processamento

5. Criar Arquivo ARFF

```
% Number of Instances: 3
% Number of Attributes: 5
% 0 - atributo NÃO pertence à sentença
% 1 - atributo pertence à sentença

@relation AnemiaFalciforme

@attribute parvovirus {0,1}
@attribute infection {0,1}
@attribute crises {0,1}
@attribute hu {0,1}
@attribute class {complication,benefit,other}

@data
1,1,0,0,complication
0,0,1,1,benefit
0,0,0,1,other
```

Classificação Textual

1ª Fase
Pré-Processamento

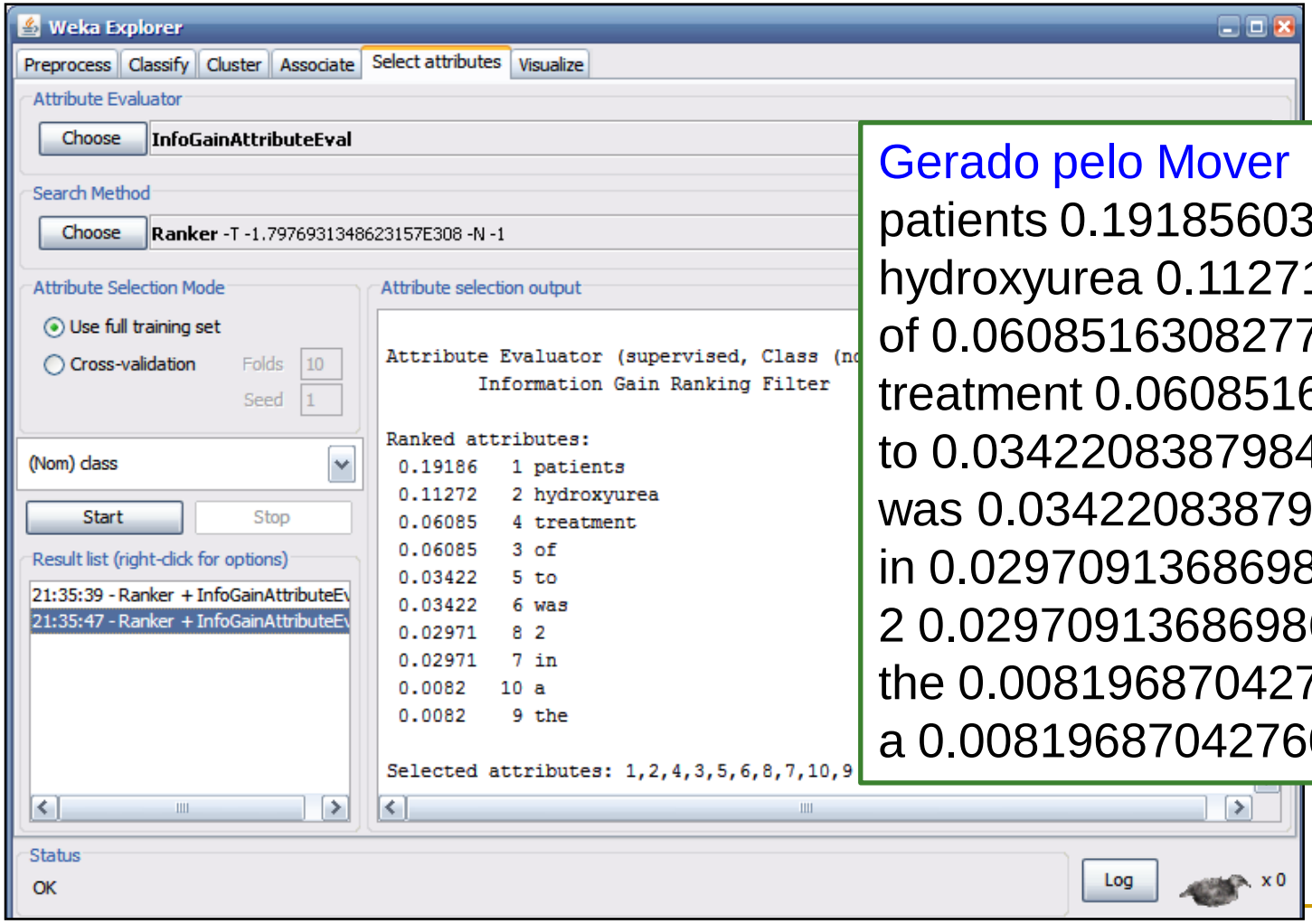
5. Criar Arquivo ARFF

```
10  import weka.core.Attribute;
11  import weka.core.FastVector;
12  import weka.core.Instance;
13  import weka.core.Instances;
14
15  public class CriaARFF
16  {
17      public static void main(String[] args)
18      {
19          // First let's create the attributes.
20          Attribute x = new Attribute("x");
21          Attribute y = new Attribute("y");
22          // Third attribute is nominal.
23          FastVector classesLabels = new FastVector(5);
24          classesLabels.addElement("A");
25          classesLabels.addElement("B");
26          classesLabels.addElement("C");
27          classesLabels.addElement("D");
28          classesLabels.addElement("E");
29          Attribute classes = new Attribute("classes",classesLabels);
30          // Create a vector of attributes information.
31          FastVector attributes = new FastVector(3);
32          attributes.addElement(x);
```

Classificação Textual

2ª Fase
Seleção de Atributo

1. Seleção de atributo: Ganho de Informação, Qui-Quadrado e Log-likelihood. Atributo Nominal



Gerado pelo Mover
patients 0.191856037238505
hydroxyurea 0.112716636725634
of 0.060851630827786 7
treatment 0.060851630827786 6
to 0.034220838798486 8
was 0.034220838798486 8
in 0.029709136869865 1
2 0.029709136869865 1
the 0.008196870427607 53
a 0.008196870427607 53

Classificação Textual

2ª Fase
Seleção de Atributo

2. Criação da Tabela Atributo-Valor
   A. *Ganho de Informação, Qui-Quadrado,...*
   B. *Bag of words*

	Atributos					
Sentenças		parvovirus	infection	crises	hu	Classe
	Sentença1	1	1	0	0	Complication
	Sentença2	0	0	1	1	Benefit
	Sentença3	0	0	0	1	Other
1 = contém; 0 = não contém						

Classificação Textual

2ª Fase
Seleção de Atributo

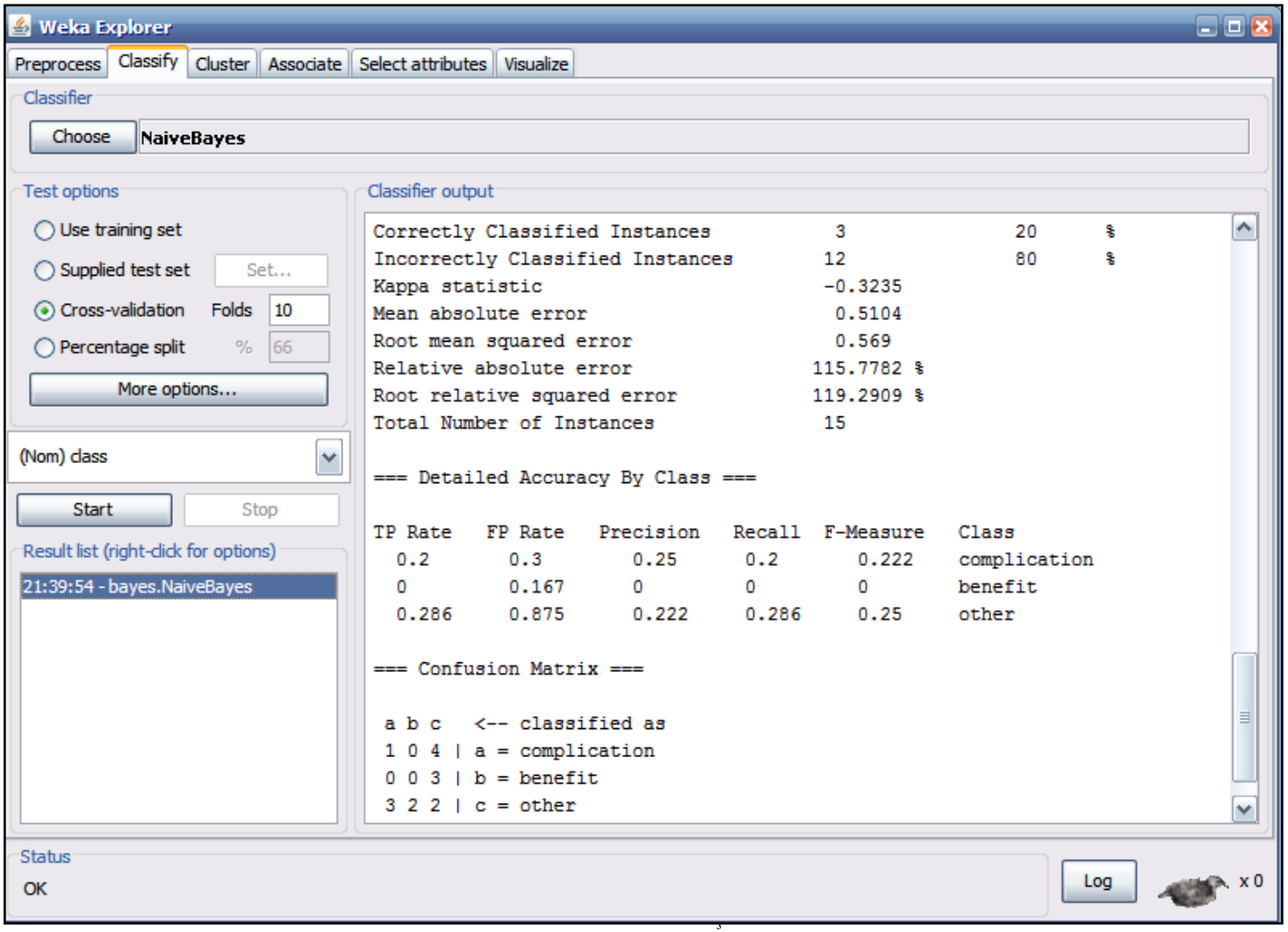
3. Criar Arquivo ARFF

```
% Number of Instances: 3
% Number of Attributes: 5
% 0 - atributo NÃO pertence à sentença
% 1 - atributo pertence à sentença

@relation AnemiaFalciforme

@attribute parvovirus {0,1}
@attribute infection {0,1}
@attribute crises {0,1}
@attribute hu {0,1}
@attribute class {complication,benefit,other}

@data
1,1,0,0,complication
0,0,1,1,benefit
0,0,0,1,other
```

Classificação Textual

3ª Fase
Classificação

1. Classificação: Atributo Nominal



Referências

- PORTER, M. F. An algorithm for suffix stripping. **Program**, v. 14, n. 3, p. 130-137, 1980. Disponível em: <http://www.mis.yuntech.edu.tw/~huangcm/research/porter.pdf>. Acesso em: 24 ago. 2009.

- SANTOS, R. **Weka na munheca.** 2005. Disponível em: <www.lac.inpe.br/~rafael.santos/Docs/CAP359/2005/weka.pdf>. Acesso em: 25 ago. 2009.

- WEKA. **Data mining with open source machine learning software in Java.** Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 10 fev. 2009.

- WITTEN, I. H.; FRANK, E. **Data mining:** practical machine learning tools and techniques with Java implementations. 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005. 525 p.

Universidade Federal de São Carlos - UFSCar
Departamento de Computação - DC
Programa de Pós-Graduação em Ciência da Computação - PPGCC

Ambiente Weka

Waikato Environment for Knowledge Analysis

Aluno: Pablo Freire Matos
Orientador: Dr. Ricardo Rodrigues Ciferri
Coorientador: Dr. Thiago Alexandre S. Pardo
Área: Banco de Dados

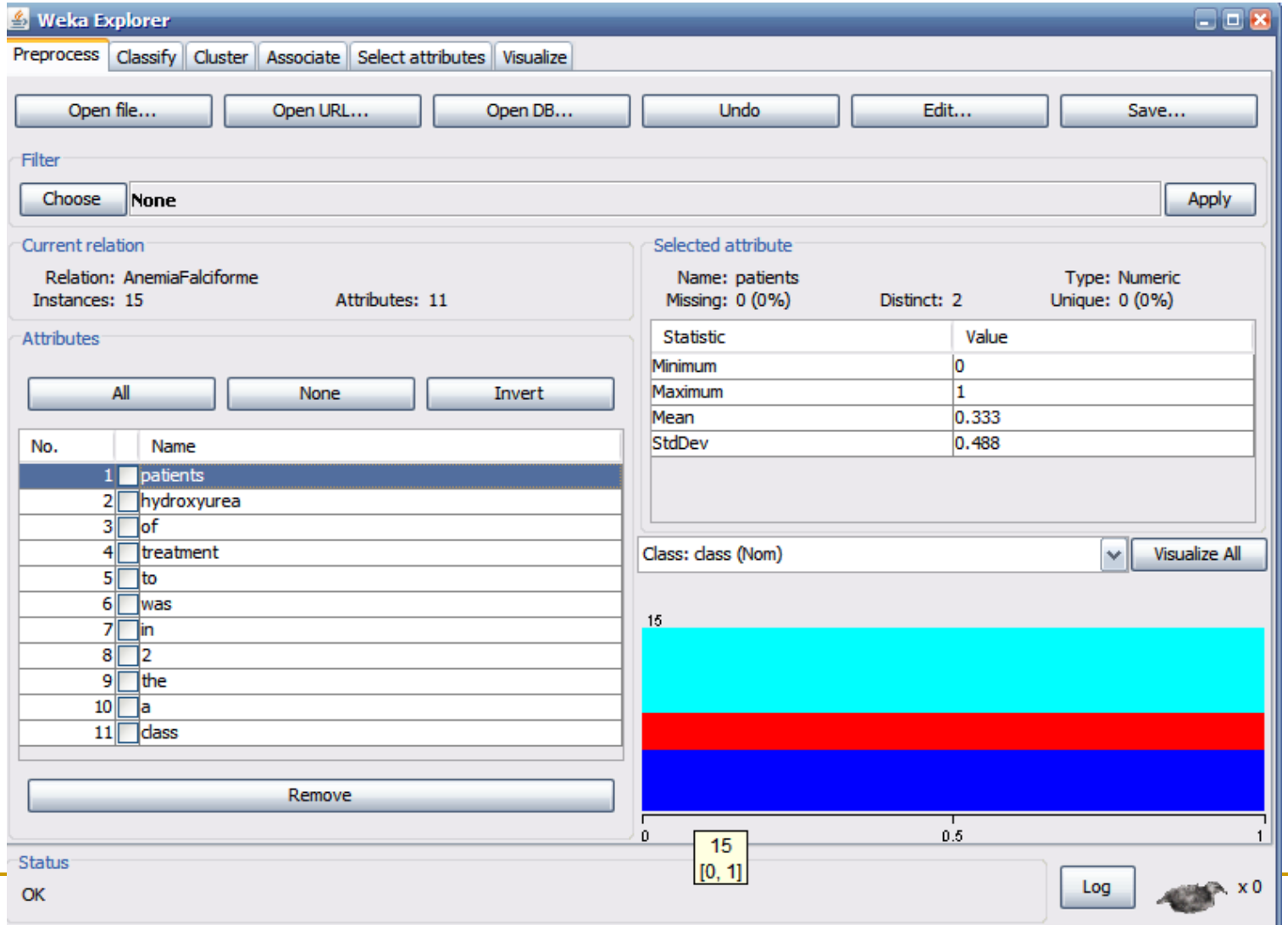
Linha de comando

- Classificação
  - java -**cp** weka.jar weka.classifiers.trees.J48 **-t** data/weather.arff
- Discretizar valores
  - java **-cp** weka.jar weka.filters.unsupervised.attribute.Discretize **-R** first-last **-B** 5 **-i** data/iris.arff **-o** data/iris-disc.arff
    - **-R** indica quais os índices dos atributos que devem ser discretizados
    - **-B** máximo de valores discretos
- Ajuda
  - java –cp weka.jar weka.classifiers.bayes.NaiveBayes -?

Linha de comando

- Running an *N-fold cross* validation experiment
  - `java -cp ~cs4705/bin/weka.jar weka.classifiers.bayes.NaiveBayes -t trainingdata.arff -x N -i`
- Using a predefined test set
  - `java -cp ~cs4705/bin/weka.jar weka.classifiers.bayes.NaiveBayes -t trainingdata.arff -T testingdata.arff`

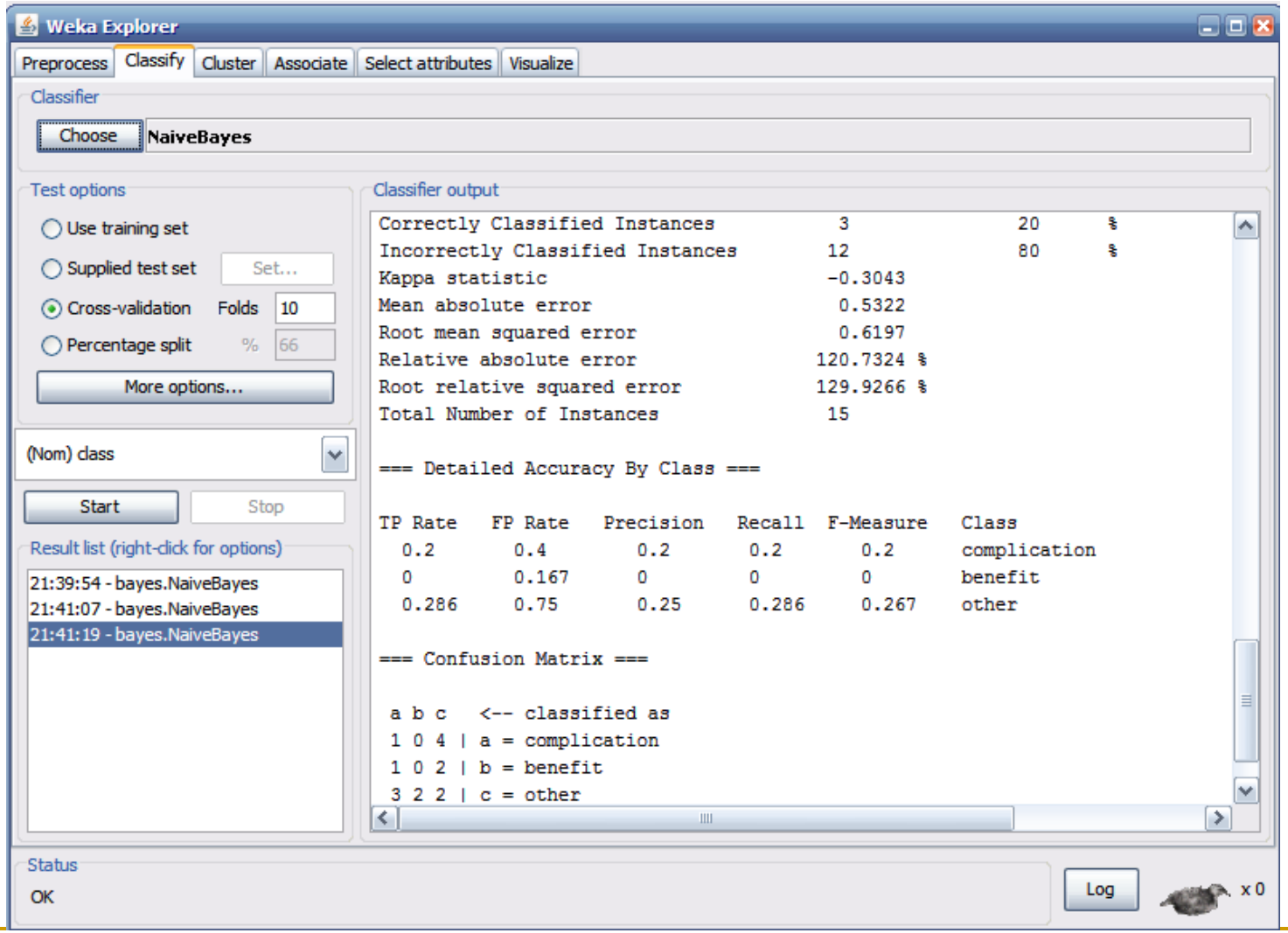
Weka Explorer – Pré-processamento Atributo Numérico



Classificação Textual

3ª Fase
Classificação

1. Classificação: Atributo Numérico



Classificação Textual

3ª Fase
Classificação

1. Classificação: Atributo Nominal