

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



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

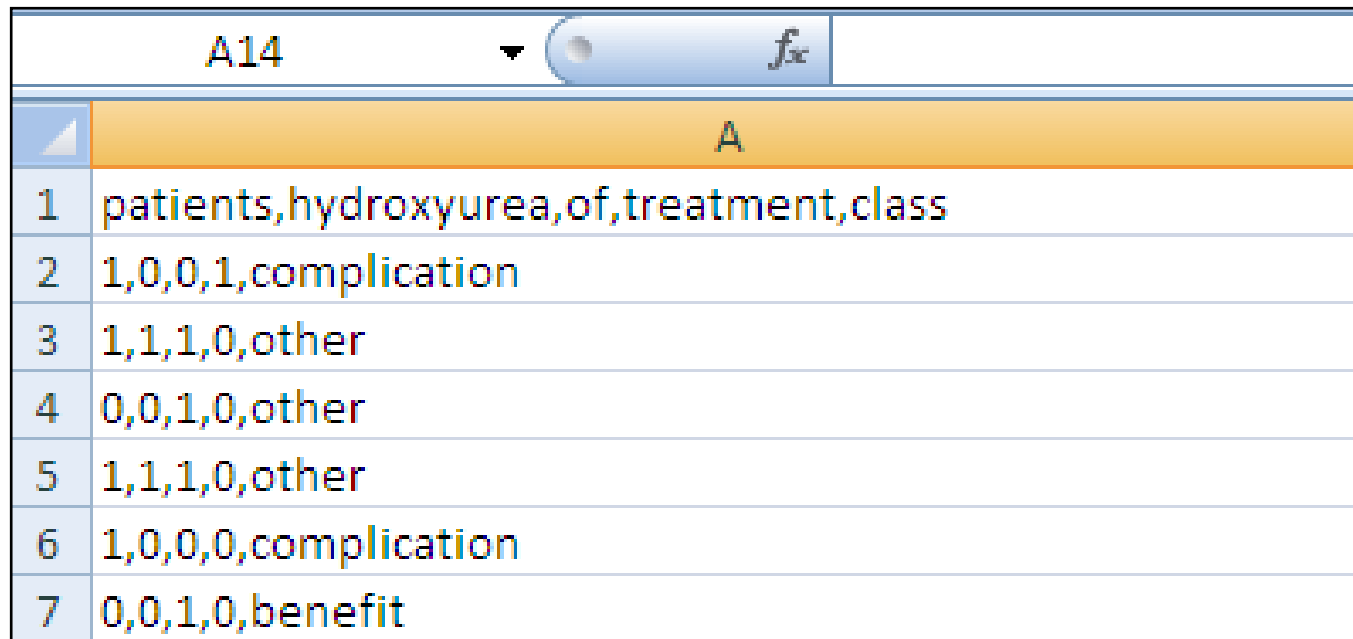
@relation AnemiaFalciforme

@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)*



The screenshot shows a Weka interface window titled 'A14'. The window contains a table with a single column labeled 'A' and 7 rows of data. The data is as follows:

	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

The screenshot shows the Weka Explorer application window. The 'Select attributes' tab is active, and the 'patients' attribute is selected. The interface displays the current relation 'mover' with 15 instances and 11 attributes. A table lists the attributes, with 'patients' selected. The 'Selected attribute' section shows the distribution of the 'patients' attribute, with a bar chart visualizing the counts for labels 0 and 1.

Current relation
Relation: mover
Instances: 15
Attributes: 11

Attributes

No.	Name
1	<input checked="" type="checkbox"/> patients
2	<input type="checkbox"/> hydroxyurea
3	<input type="checkbox"/> of
4	<input type="checkbox"/> treatment
5	<input type="checkbox"/> to
6	<input type="checkbox"/> was
7	<input type="checkbox"/> in
8	<input type="checkbox"/> 2
9	<input type="checkbox"/> the
10	<input type="checkbox"/> a
11	<input type="checkbox"/> class

Selected attribute
Name: patients
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

Label	Count
0	10
1	5

Class: class (Nom) Visualize All

Status
OK

Log x 0

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**

Weka Explorer

Preprocess Classify Cluster Associate **Select attributes** Visualize

Attribute Evaluator

Choose **InfoGainAttributeEval**

Search Method

Choose **Ranker -T -1.7976931348623157E308 -N -1**

Attribute Selection Mode

Use full training set

Cross-validation Folds Seed

(Nom) class

Start Stop

Result list (right-click for options)

21:35:39 - Ranker + InfoGainAttributeEv
21:35:47 - Ranker + InfoGainAttributeEv

Attribute selection output

Attribute Evaluator (supervised, Class (no
Information Gain Ranking Filter

Ranked attributes:

0.19186	1	patients
0.11272	2	hydroxyurea
0.06085	4	treatment
0.06085	3	of
0.03422	5	to
0.03422	6	was
0.02971	8	2
0.02971	7	in
0.0082	10	a
0.0082	9	the

Selected attributes: 1,2,4,3,5,6,8,7,10,9

Status

OK Log x 0

Gerado pelo Mover

patients 0.191856037238505
hydroxyurea 0.112716636725634
of 0.0608516308277867
treatment 0.0608516308277866
to 0.0342208387984868
was 0.0342208387984868
in 0.0297091368698651
2 0.0297091368698651
the 0.00819687042760753
a 0.00819687042760753

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

The screenshot shows the Weka Explorer interface with the NaiveBayes classifier selected. The 'Classifier output' pane displays the following results:

```
Correctly Classified Instances      3          20  %
Incorrectly Classified Instances    12          80  %
Kappa statistic                    -0.3235
Mean absolute error                 0.5104
Root mean squared error             0.569
Relative absolute error            115.7782 %
Root relative squared error        119.2909 %
Total Number of Instances          15
```

Below this, a table shows 'Detailed Accuracy By Class':

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.2	0.3	0.25	0.2	0.222	complication
0	0.167	0	0	0	benefit
0.286	0.875	0.222	0.286	0.25	other

Finally, the 'Confusion Matrix' is shown:

```
=== Confusion Matrix ===
 a b c  <-- classified as
 1 0 4 | a = complication
 0 0 3 | b = benefit
 3 2 2 | c = other
```

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

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Undo | Edit... | Save...

Filter: Choose **None** Apply

Current relation: Relation: AnemiaFalciforme, Instances: 15, Attributes: 11

Attributes: All | None | Invert

No.	Name
1	<input checked="" type="checkbox"/> patients
2	<input type="checkbox"/> hydroxyurea
3	<input type="checkbox"/> of
4	<input type="checkbox"/> treatment
5	<input type="checkbox"/> to
6	<input type="checkbox"/> was
7	<input type="checkbox"/> in
8	<input type="checkbox"/> 2
9	<input type="checkbox"/> the
10	<input type="checkbox"/> a
11	<input type="checkbox"/> class

Remove

Selected attribute: Name: patients, Missing: 0 (0%), Distinct: 2, Type: Numeric, Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.333
StdDev	0.488

Class: class (Nom) Visualize All

15

0 0.5 1

15 [0, 1]

Status: OK Log x 0

Classificação Textual

3ª Fase Classificação

1. Classificação: Atributo Numérico

The screenshot shows the Weka Explorer interface with the NaiveBayes classifier selected. The 'Test options' section is set to 'Cross-validation' with 10 folds and a 66% split. The 'Classifier output' section displays the following results:

```
Correctly Classified Instances      3          20  %
Incorrectly Classified Instances   12          80  %
Kappa statistic                    -0.3043
Mean absolute error                 0.5322
Root mean squared error             0.6197
Relative absolute error             120.7324 %
Root relative squared error         129.9266 %
Total Number of Instances          15
```

Below this, a table shows 'Detailed Accuracy By Class':

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.2	0.4	0.2	0.2	0.2	complication
0	0.167	0	0	0	benefit
0.286	0.75	0.25	0.286	0.267	other

Finally, the 'Confusion Matrix' is shown:

```
=== Confusion Matrix ===
 a b c  <-- classified as
1 0 4 | a = complication
1 0 2 | b = benefit
3 2 2 | c = other
```


Classificação Textual

3ª Fase Classificação

1. Classificação: Atributo Nominal

The screenshot shows the Weka Explorer interface with the NaiveBayes classifier selected. The 'Classifier output' pane displays the following performance metrics:

Correctly Classified Instances	3	20	%
Incorrectly Classified Instances	12	80	%
Kappa statistic	-0.3235		
Mean absolute error	0.5104		
Root mean squared error	0.569		
Relative absolute error	115.7782 %		
Root relative squared error	119.2909 %		
Total Number of Instances	15		

Below the metrics, the 'Detailed Accuracy By Class' is shown:

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.2	0.3	0.25	0.2	0.222	complication
0	0.167	0	0	0	benefit
0.286	0.875	0.222	0.286	0.25	other

The 'Confusion Matrix' is also displayed:

```
=== Confusion Matrix ===
 a b c  <-- classified as
1 0 4 | a = complication
0 0 3 | b = benefit
3 2 2 | c = other
```

The 'Test options' pane shows 'Cross-validation' selected with 10 folds. The 'Result list' shows the classifier '21:39:54 - bayes.NaiveBayes'.