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

Resultados com o WEKA

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

Roteiro

- **Testes Realizados na Classificação de Sentenças**
  - **MOVER x WEKA**

Configuração Usada no Mover

- Tipo de treinamento
  - *frequency measure (orig)*
  - *chi-squared (chi)*
  - *information gain (ig)*

- Todas *features* utilizadas
  - *all #used features*
  - *all #features output cut*
  - *all #orig features used*

- Otimização e n-gramas
  - *2 #max stored limit for optimize flow*
  - *5 #max cluster size used to generate features*

Sentenças

- Para cada sentença de treinamento foi excluído:
  - Ponto final
  - Vírgula
  - Parênteses

Todas as palavras de cada sentença de treinamento devem ser **Minúsculas**

Sentença

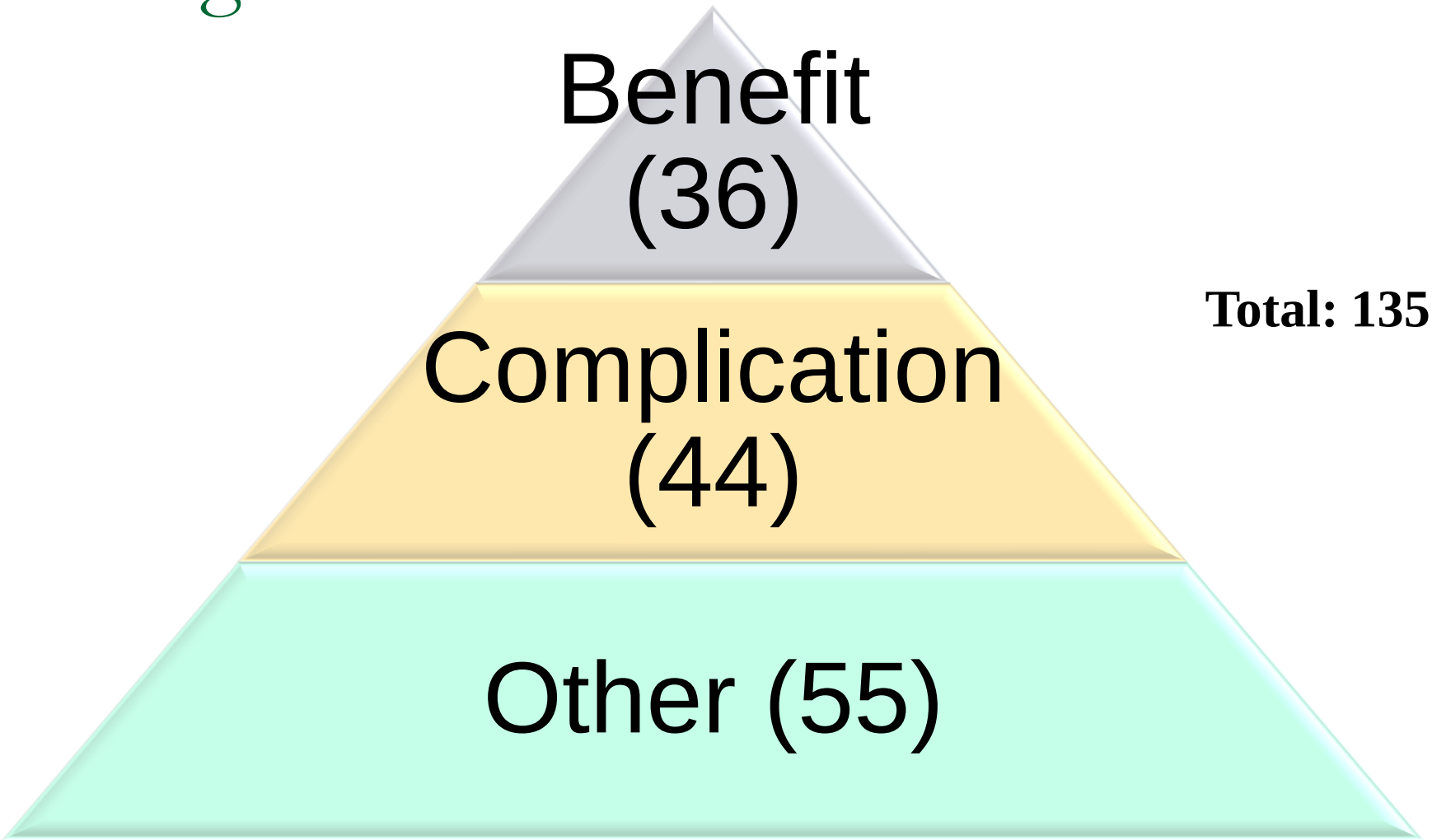
- Cada sentença tem um número informando a que artigo ele pertence, e.g.:
  - “8 the storage of frozen sperm was offered to mature boys”
  - Significa que a sentença “the storage of frozen…” foi extraída do artigo 8

Categorias de Treinamento

- A pasta "*training*" contém 3 categorias para treinamento:
  - 1_complication
  - 2_benefit
  - *5_other*

Cada categoria contém várias sentenças identificadas manualmente

Quantidade de sentença por classe com 4 artigos



Total: 135

Medida de Utilidade no Mover

Stratified Cross-Validation	Precisão do Teste Aleatório	Desvio Padrão	Medida de Utilidade
5 Folds	65,93%	+ ou - 8,45%	IG (Ganho de Informação)
5 Folds	65,93%	+ ou - 8,45%	QUI (Qui-Quadrado)
5 Folds	65,93%	+ ou - 8,45%	ORIG (Frequência)

- WEKA:
  - Ganho de Informação e Qui-Quadrado **também** geram o mesmo resultado

Passos para a Classificação no WEKA

1. Carregar os dados
2. Aplicar pré-processamento
    1. Gerar matriz atributo-valor
        1. Frequência, n-gramas, stopwords, stemmer ...
    2. Converter numérico para nominal
    3. RemoveMisclassifiedTest
    4. Randomize
    5. Seleção de atributo
3. Selecionar Classificador

Resultado Mover x Weka

Perda de 5,59%

- Naïve Bayes
- Ganho de Informação (com ou sem)
- Matriz atributo-valor
  - Frequência mínima = 1
  - 1-grama

Cross-Validation	Mover (Stratified CV)	Weka (CV)
3 Folds	64,24%	56,29%
5 Folds	68,15%	61,48%
10 Folds	70,77%	63,70% 59,25% (com **stopword**) 62,22% (com stemmer **IteratedLovins**) 62,96% (com stemmer **Lovins**) 63,70% (com stemmer **snowball**) **65,18%** (stopword e snowball)

Resultado Mover x Weka

Ganho de 3,3%

- Naïve Bayes
- Ganho de Informação (com ou sem)
- Matriz atributo-valor
  - Frequência mínima = 2 e 3
  - 1-grama

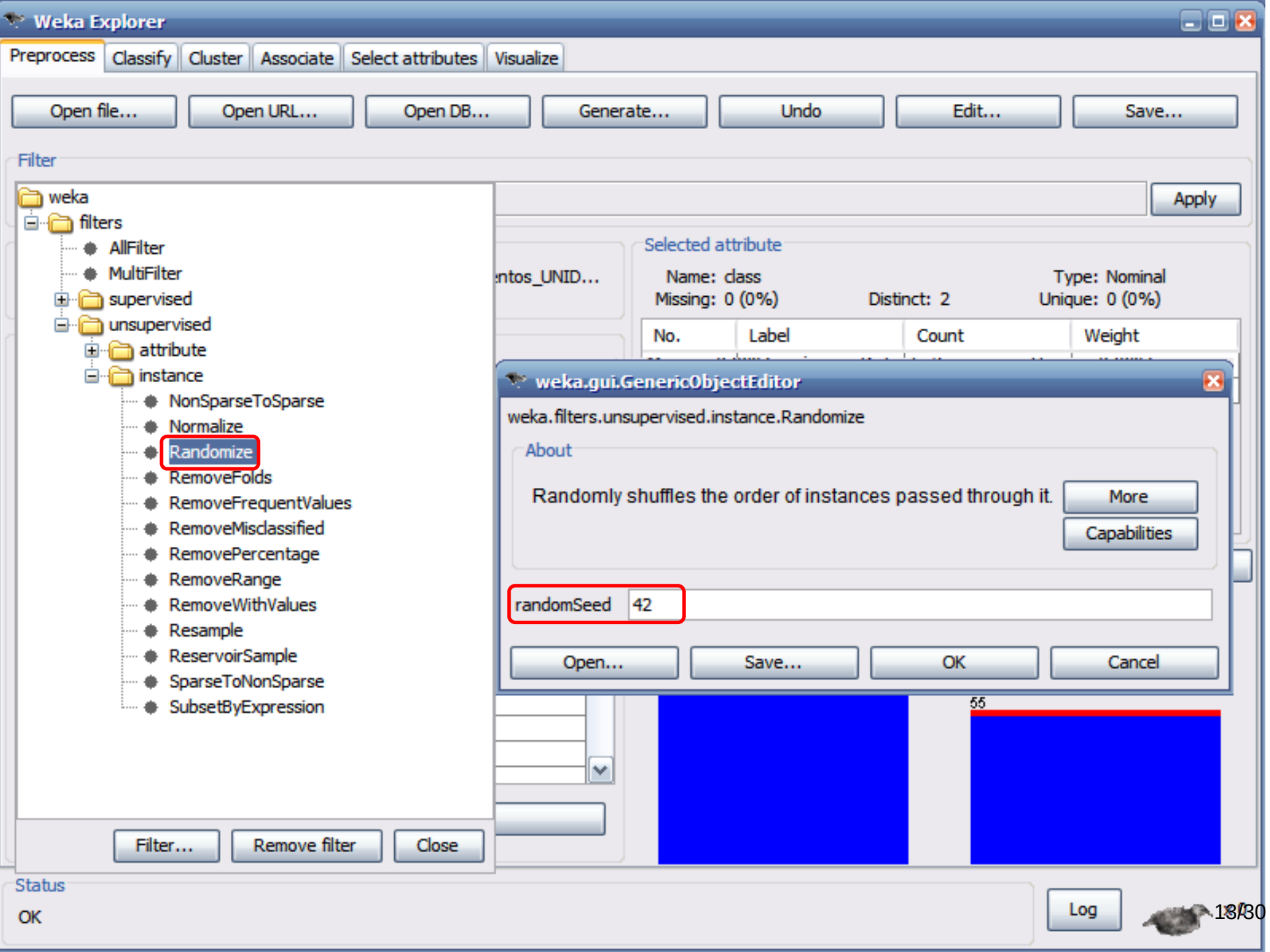
Cross-Validation	Mover (Stratified CV)	Weka (CV)
3 Folds	64,24%	67,40%
5 Folds	68,15%	72,60%
10 Folds	70,77%	74,07% 72,59% (com **stopword**) 71,11% (com stemmer **IteratedLovins**) 72,59% (com stemmer **Lovins**) 70,37% (com stemmer **Snowball**) 73,33% (stopword e, Lovins ou Snowball)

Resultado Mover x Weka

Ganho de 7,74%

- Naïve Bayes
- Ganho de Informação (com ou sem)
- Matriz atributo-valor
  - Frequência mínima = 2
  - 1 a 3 gramas

Cross-Validation	Mover (Stratified CV)	Weka (CV)
3 Folds	64,24%	68,14%
5 Folds	68,15%	76,30%
10 Folds	70,77%	78,51% 75,55% (com **stopword**) 74,81% (com stemmer **IteratedLovins**) 76,29% (com stemmer **Lovins**) 75,55% (com stemmer **Snowball**) 75,55% (stopword e Snowball)


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
weka
filters
AllFilter
MultiFilter
supervised
unsupervised
attribute
instance
NonSparseToSparse
Normalize
Randomize
RemoveFolds
RemoveFrequentValues
RemoveMisclassified
RemovePercentage
RemoveRange
RemoveWithValues
Resample
ReservoirSample
SparseToNonSparse
SubsetByExpression
Apply
Selected attribute
Name: class
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)
No.
Label
Count
Weight
weka.gui.GenericObjectEditor
weka.filters.unsupervised.instance.Randomize
About
Randomly shuffles the order of instances passed through it.
More
Capabilities
randomSeed
42
Open...
Save...
OK
Cancel
55
Filter...
Remove filter
Close
Status
OK
Log
13/30

Resultado Mover x Weka

Sem Ganho

- Naïve Bayes
- Ganho de Informação (com ou sem)
- Matriz atributo-valor
  - Frequência mínima = 2
  - 1 a 3 gramas
- Randomize

Cross-Validation	Mover (Stratified CV)	Weka (CV)	Weka (CV) (Randomize)
3 Folds	64,24%	68,14%	69,62%
5 Folds	68,15%	76,30%	74,81%
10 Folds	70,77%	**78,51%**	75,55%

Naïve Bayes no WEKA

- Mesmos resultados
  - Naïve Bayes
  - Naïve BayesUpdateable
  - Naïve BayesSimple
- Naïve BayesMultinominal
- Naïve BayesMultinominalUpdateable

Resultado no Weka com 135 exemplos

- Ganho de Informação
- Matriz atributo-valor
  - Frequência mínima = 2
  - 1 a 3 gramas

Valores dos parâmetros padrão

Classificador	10-Fold Cross-Validation	Tempo (segundos)	10-Fold Cross-Validation (Randomize)	Tempo (segundos) (Randomize)
Naïve Bayes	**78,51%**	0,22	75,55%	0,11
OneR				0,11
Prism				1,8
J48				0,94
ID3	54,07%	0,67	56,29	0,7
SVM	71,85%	0,55	71,85%	0,66

Randomize influencia o NB
Randomize NÃO influencia o SVM

Weka Explorer

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

Classifier

Choose **NaiveBayes**

Test options

- (•) Use training set
- ( ) Supplied test set — Set...
- ( ) Cross-validation — Folds 10
- ( ) Percentage split — % 66

More options...

(Nom) class

Start | Stop

Result list (right-click for options)

21:26:59 - bayes.NaiveBayes

Classifier output

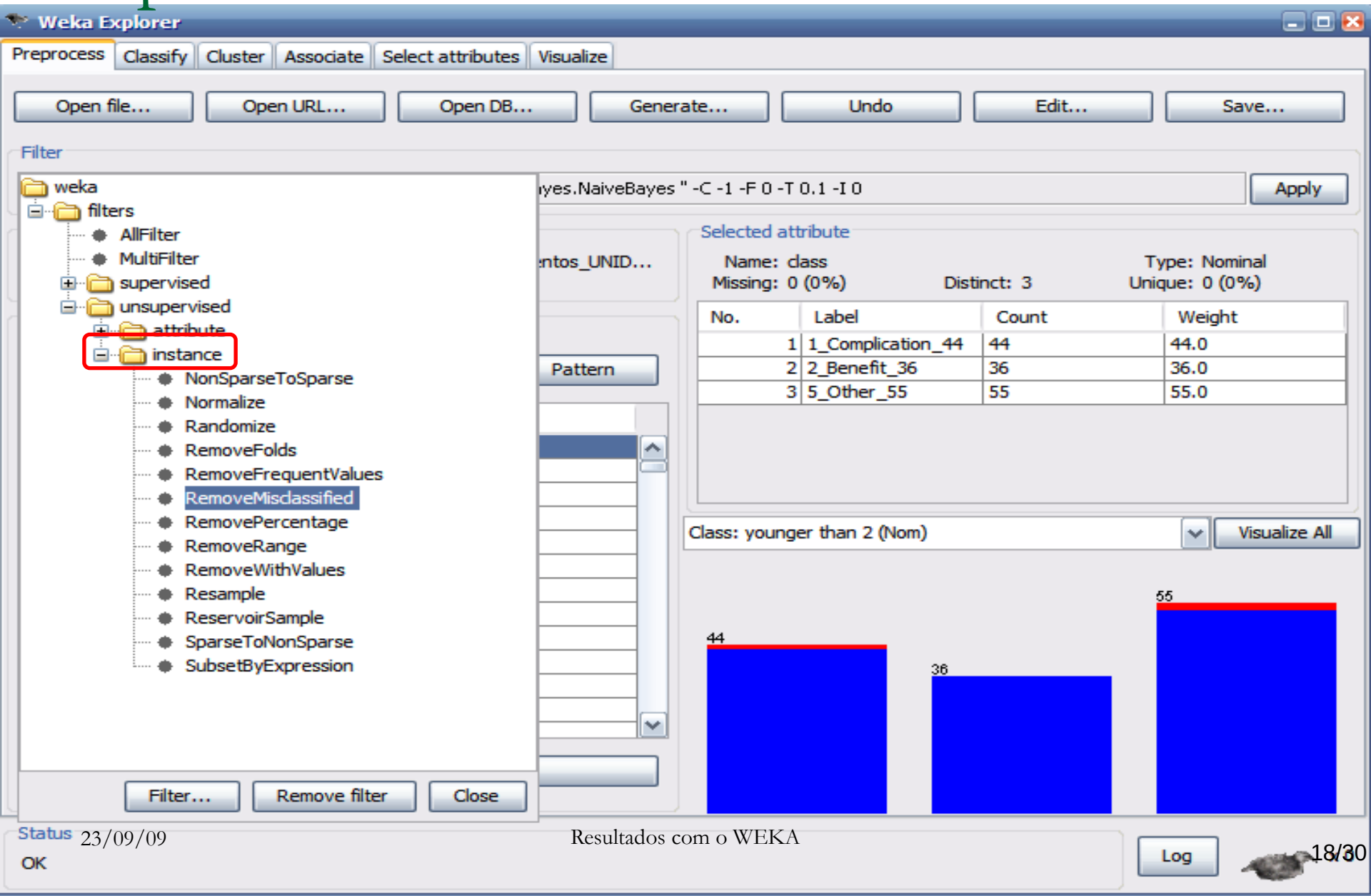
```
Correctly Classified Instances         127               94.0741 %
Incorrectly Classified Instances         8                5.9259 %
Kappa statistic                          0.9089
Mean absolute error                      0.0467
Root mean squared error                  0.1895
Relative absolute error                 10.666  %
Root relative squared error             40.5048 %
Total Number of Instances              135

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Clas
                 0.886     0          1         0.886     0.94       0.998    1_C
                 0.917     0          1         0.917     0.957      0.999    2_B
                 1         0.1        0.873     1         0.932      1        5_C
Weighted Avg.    0.941     0.041      0.948     0.941     0.941      0.999

=== Confusion Matrix ===

  a  b  c   <-- classified as
 39  0  5 |  a = 1_Complication_44
  0 33  3 |  b = 2_Benefit_36
  0  0 55 |  c = 5_Other_55
```

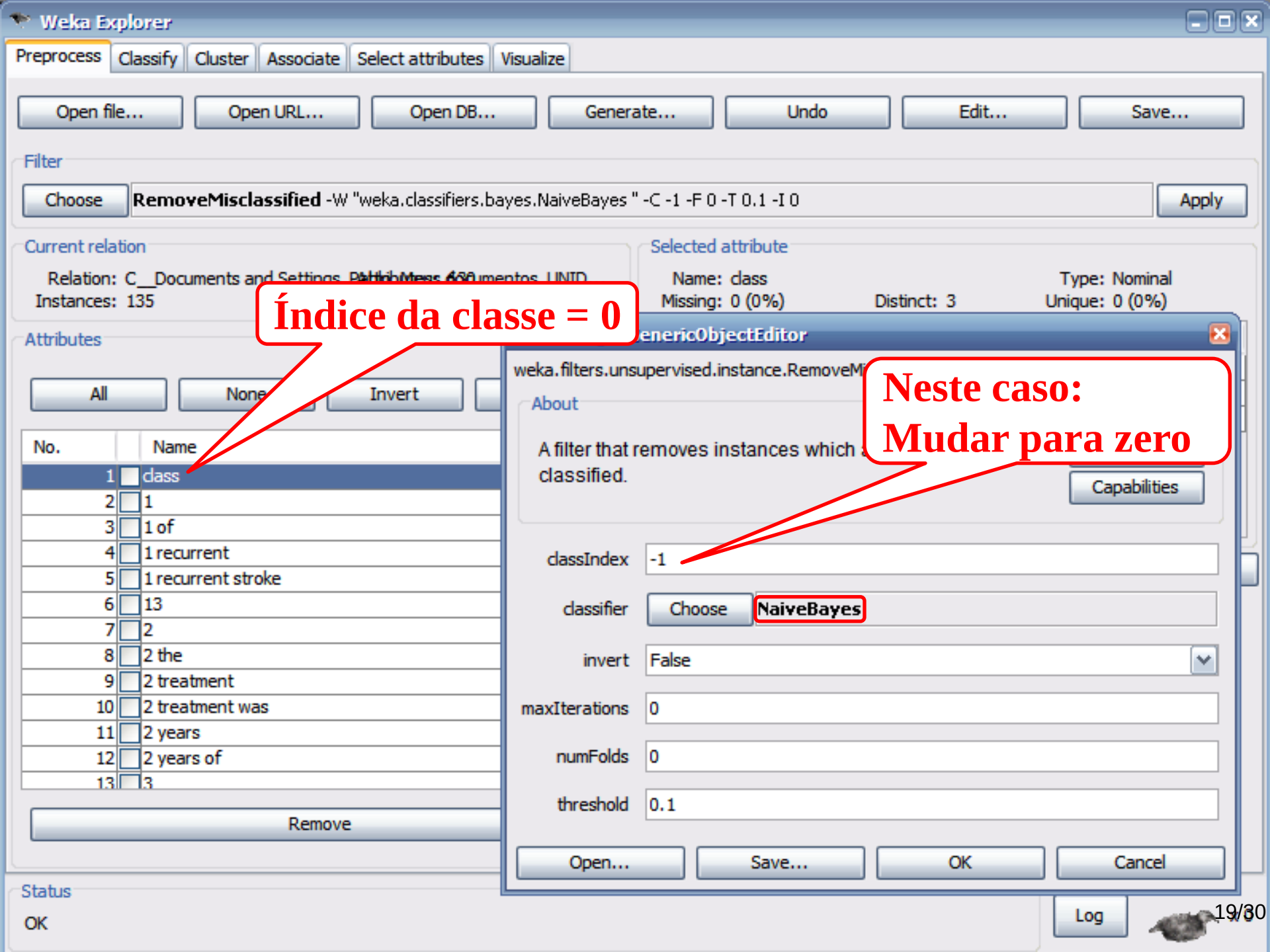
Status

OK

Log

Aplicar Filtro: Remove Misclassified




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
RemoveMisclassified -W "weka.classifiers.bayes.NaiveBayes " -C -1 -F 0 -T 0.1 -I 0
Apply
Current relation
Relation: C__Documents and Settings
Instances: 135
Selected attribute
Name: class
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
Índice da classe = 0
Attributes
All
None
Invert
No.
Name
1 class
2 1
3 1 of
4 1 recurrent
5 1 recurrent stroke
6 13
7 2
8 2 the
9 2 treatment
10 2 treatment was
11 2 years
12 2 years of
13 3
Remove
GenericObjectEditor
weka.filters.unsupervised.instance.RemoveMi
About
A filter that removes instances which
classified.
Capabilities
Neste caso:
Mudar para zero
classIndex
-1
classifier
Choose
NaiveBayes
invert
False
maxIterations
0
numFolds
0
threshold
0.1
Open...
Save...
OK
Cancel
Status
OK
Log
19/30

Weka Explorer

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

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

Filter

Choose | **RemoveMisclassified** -W "weka.classifiers.bayes.NaiveBayes " -C 0 -F 0 -T 0.1 -I 0 | Apply

Current relation

Relation: C__Documents and Settings_Pabbo_Meus Documentos_UNID...
Instances: 109 Sum of weights: 109

Instâncias inicial = 135
Redução de 19,25%

Attributes

All | Nor

No.	Name
1	class
2	1
3	1 of
4	1 recurrent
5	1 recurrent stroke
6	13
7	2
8	2 the
9	2 treatment
10	2 treatment was
11	2 years
12	2 years of
13	3

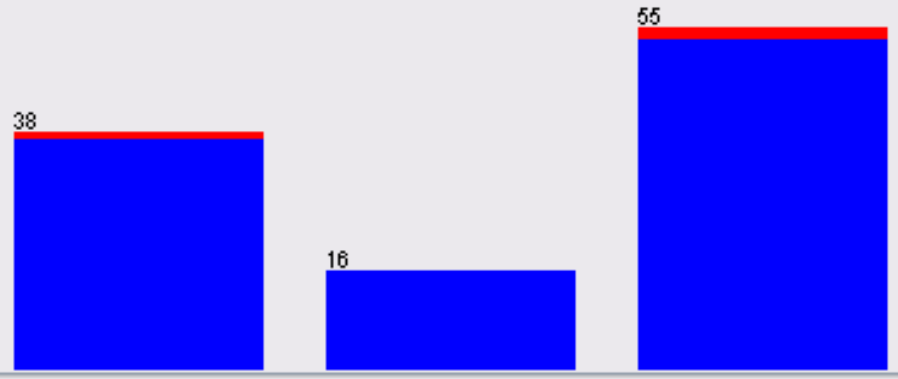
Remove

Selected attribute

Name: class Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

	Label	Count	Weight
1	1_Complication_44	38	38.0
2	2_Benefit_36	16	16.0
3	5_Other_55	55	55.0

Class: younger than 2 (Nom) | Visualize All



Status

OK | Log

Weka Explorer

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

Classifier

Choose **AttributeSelectedClassifier** -E "weka.attributeSelection.InfoGainAttributeEval " -S "weka.attributeSelection.Ranker -T -1.7976931348623157E308 -f

Test options

- Use training set
- Supplied test set — Set...
- Cross-validation — Folds 10
- Percentage split — % 66

More options...

(Nom) class

Start | Stop

Result list (right-click for options)

21:26:59 - bayes.NaiveBayes
21:32:47 - meta.AttributeSelectedClassifier
21:39:26 - meta.AttributeSelectedClassifier
21:40:17 - meta.AttributeSelectedClassifier

Classifier output

```
Correctly Classified Instances         109               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0181
Root mean squared error                  0.0752
Relative absolute error                  4.4957 %
Root relative squared error             16.7837 %
Total Number of Instances              109

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Cla
                 1         0          1         1         1          1        1_(
                 1         0          1         1         1          1        2_I
                 1         0          1         1         1          1        5_(
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===

  a  b  c   <-- classified as
 38  0  0 |  a = 1_Complication_44
  0 16  0 |  b = 2_Benefit_36
  0  0 55 |  c = 5_Other_55
```

Status

OK

Log

Resultado no Weka (109 exemplos-NB)

- Pré-processamento
  - Matriz atributo-valor
    - Frequência mínima = 2
    - 1 a 3 gramas
  - RemoveMisclassifiedTest do Naïve Bayes (NB)

Valores dos parâmetros padrão

Classificador	10-Fold Cross-Validation (Resultado Anterior)	Tempo (segundos) (Resultado Anterior)	10-Fold Cross-Validation	Tempo (segundos)
Naïve Bayes	78,51%	0,22	76,14%	0,11
OneR	41,48%	0,09	52,29%	0,09
Prism	45,92%	1,47	51,37%	0,91
J48	50,37%	0,91	66,05	0,52
ID3	54,07%	0,67	75,22%	0,47
SVM	71,85%	0,55	**78,89%**	0,33

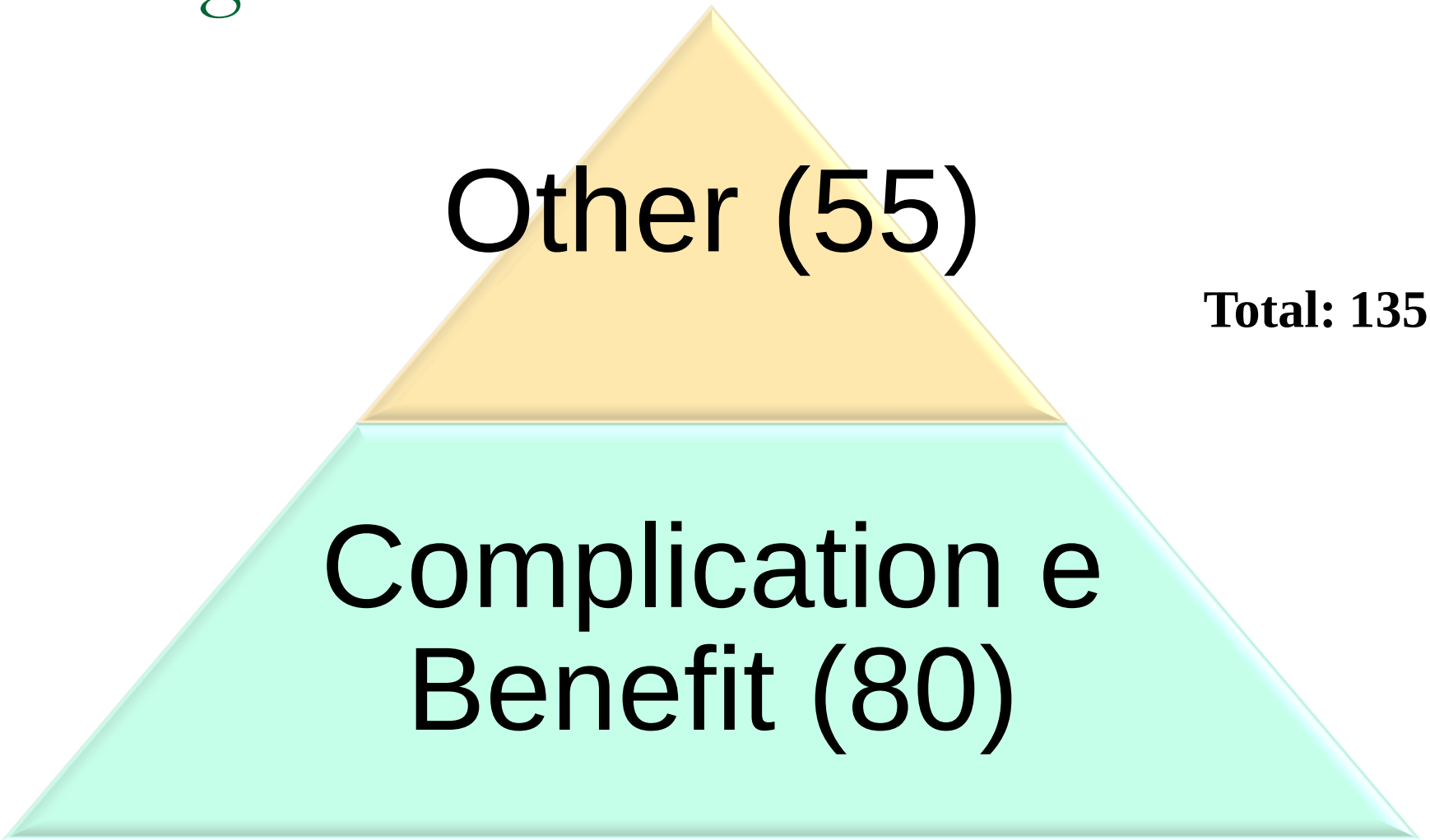
Resultado no Weka

- Pré-processamento
  - Matriz atributo-valor
    - Frequência mínima = 2
    - 1 a 3 gramas
  - RemoveMisclassifiedTest de cada classificador

Valores dos parâmetros padrão

Classificador	10-Fold Cross-Validation (Resultado Anterior)	10-Fold Cross-Validation
Naïve Bayes	78,51%	76,14% (109 exemplos)
OneR	41,48%	100% (64 ex)
Prism	45,92%	45,92% (135 ex)
J48	50,37%	65,28 (121 ex)
ID3	54,07%	54,07% (135 ex)
SVM	71,85%	71,85% (135 ex)

Quantidade de sentença por classe com 4 artigos



Total: 135

Resultado Mover x Weka

Ganho de 5,33%

- Naïve Bayes
- Ganho de Informação (com ou sem)
- Pré-processamento
  - Matriz atributo-valor
    - Frequência mínima = 2
    - 1 a 3 gramas
  - RemoveMisclassifiedTest e Randomize

Cross-Validation	Weka (CV) – 2 classes com 135 exemplos	Weka (CV) – 2 classes com 130 exemplos (RemoveMisclassifiedTest)	Weka (CV) – 2 classes com 135 exemplos (Randomize)	Weka (CV) – 2 classes com 130 exemplos (Remove e Randomize)
10 Folds	78,51	82,30%	82,96%	83,84%

Resultado Mover x Weka

Ganho de 5,33%

- Naïve Bayes
- Ganho de Informação (com ou sem)
- Pré-processamento
  - Matriz atributo-valor
    - Frequência mínima = 2
    - 1 a 3 gramas
  - RemoveMisclassifiedTest e Randomize

Cross-Validation	Mover (Stratified CV)	Weka (CV) – 3 classes com 135 exemplos	Weka (CV) – 3 classes com 109 exemplos (Remove)	Weka (CV) – 2 classes com 135 exemplos	Weka (CV) – 2 classes com 130 exemplos (Remove e Randomize)
10 Folds	70,77%	78,51%	76,14%	78,51	83,84%

Considerações Finais

- Pré-processamento
  1. Gerar matriz atributo-valor
     1. Frequência, n-gramas, stopwords, stemmer ...
     2. IDF e TF transform, lematização
  2. Converter numérico para nominal
  3. Randomize
  4. RemoveMisclassifiedTest
  5. Igualar os exemplos (over e under sampling)
  6. Seleção de atributo

Considerações Finais

- Pré-processamento
  - Matriz atributo-valor
    - Frequência mínima = 2
    - 1 a 3 gramas
    - sem stopword
    - sem stemmer
    - testar lematização
  - Randomize e RemoveMisclassifiedTest
    - com 2 classes (Other, Complication e Benefit)
  - RemoveMisclassifiedTest
    - Com 3 classes (Complication, Benefit e Other)

Referências

- ANTHONY, L.; LASHKIA, G. V. Mover: a machine learning tool to assist in the reading and writing of technical papers. **IEEE Transactions on Professional Communication**, v. 46, n. 3, p. 185-193, 2003.

- BURSTEIN, J.; MARCU, D.; KNIGHT, K. Finding the WRITE stuff: automatic identification of discourse structure in student essays. **Intelligent Systems**, IEEE, v. 18, n. 1, p. 32-39, 2003.

- BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. **SIGKDD Explorations Newsletter**, v. 6, n. 1, p. 20-29, 2004.

- HEY, D. F.; FELTRIM, V. D. Uma investigação sobre a aplicação de algoritmos de aprendizado à classificação de papéis retóricos. In: VIII Fórum de Informática e Tecnologia de Maringá, XI Mostra de Trabalhos de Informática, 2008, Maringá. **Anais...** Universidade Estadual de Maringá, 2008. p. 94-104.

Referências (Cont.)

- 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

Resultados com o WEKA

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