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

Experimentos Mover x Weka

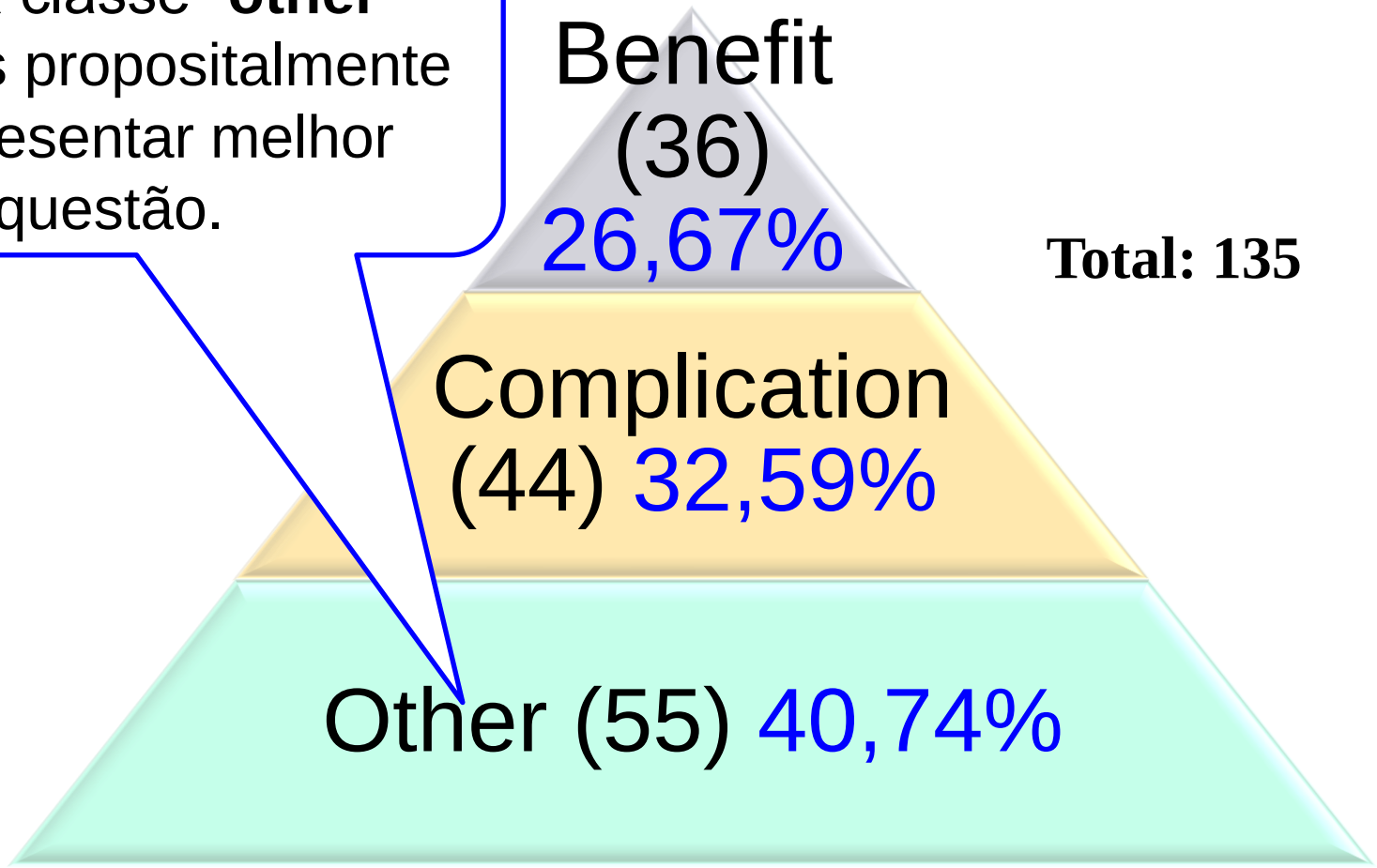
Aluno: Pablo Freire Matos
Orientador: Dr. Ricardo Rodrigues Ciferri
Coorientador: Dr. Thiago Alexandre S. Pardo
Área: Banco de Dados

Experimentos Mover x Weka

- Três Classes
  - Benefit
  - Complication
  - Other

Quantidade de sentença por classe com 4 artigos **(Seleção Tendenciosa)**

Exemplos da classe "**other** " selecionados propositalmente a fim de representar melhor a classe em questão.

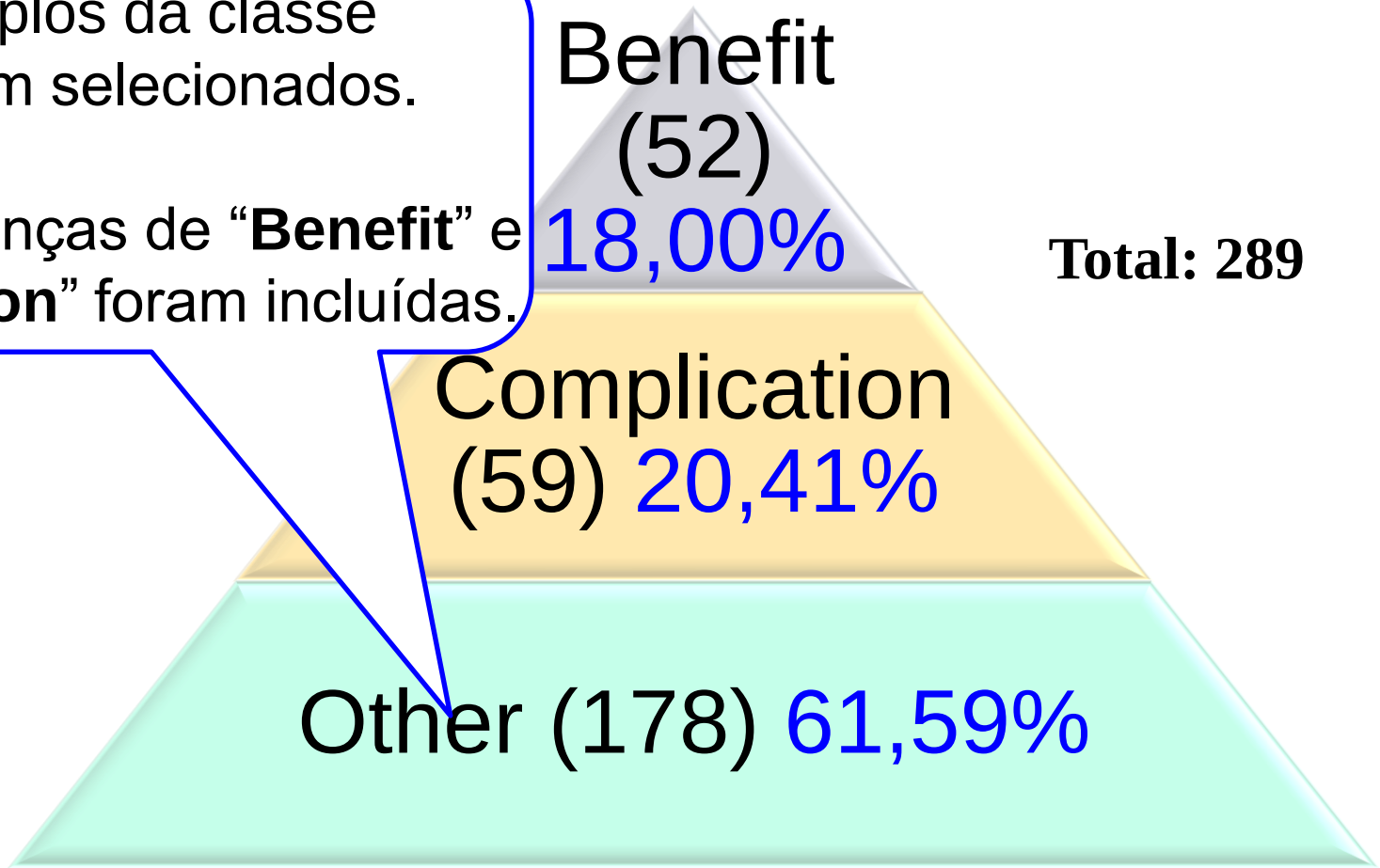


Total: 135

Quantidade de sentença por classe com 4 artigos **(Seleção Mais Sentenças)**

Todos exemplos da classe "**other** " foram selecionados.

Outras sentenças de "**Benefit**" e "**Complication**" foram incluídas.



Total: 289

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. Discretização
    3. RemoveMisclassified (Ruído)
    4. Randomize
    5. Resample (over sampling)
    6. Seleção de atributo
3. Selecionar Classificador

Resultado Weka

- Matriz atributo-valor
  - Frequência mínima = 2
  - 1 a 3 gramas
- Classificação
  - Ganho de Informação (com ou sem)
  - 10-Fold Cross-Validation

Exemplos	**Mover (Stratified CV) NB**	**Weka (CV) NB**	**Weka (CV) SVM**
135 Algumas sentenças de “outros”	70,77%	78,52% (Ganho de **7,75%**) 75,55% (com **stopword**) 74,81% (com stemmer **teratedLovins**) 76,29% (com stemmer **Lovins**) 75,55% (com stemmer **Snowball**) 75,55% (stopword e Snowball)	71,85%
289 Todas sentenças de “outros”	68,21%	70,59% (Ganho de **2,38%**)	**72,66%**

Configuração dos Experimentos

- Pré-processamento
  - Matriz atributo-valor
    - Frequência mínima = 2
    - 1 a 3 gramas
    - Sem *stopword* e *stemmer*
- Seleção de Atributo
  - Ganho de Informação
- Seis classificadores escolhidos
  - Valores dos parâmetros padrão
  - 10-Fold Cross-Validation

Parâmetros definidos nos experimentos testados

Filtro: Randomize

	135 exemplos		**289** exemplos	
Classificador	**Sem Filtro**	**Com Filtro**	**Sem Filtro**	**Com Filtro**
Naïve Bayes	**78,51%**	75,55%	70,59%	71,63%
OneR	41,48%	40,74%	63,67%	62,63%
Prism	45,92%	40,00%	48,79%	46,71%
J48	50,37%	55,55%	64,71%	65,05%
ID3	54,07%	56,29%	62,28%	59,86%
SVM	71,85%	71,85%	**72,66%**	67,82%

Eliminar Ruído

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

Filtro: RemoveMisclassified do NB

	135 exemplos		289 exemplos	
Classificador	**Sem Filtro**	**Com Filtro**	**Sem Filtro**	**Com Filtro**
Naïve Bayes	**78,51%**	76,14%	70,59%	----------
OneR	41,48%	52,29%	63,67%	----------
Prism	45,92%	51,37%	48,79%	----------
J48	50,37%	66,05%	64,71%	----------
ID3	54,07%	75,22%	62,28%	----------
SVM	71,85%	**78,89%**	**72,66%**	----------
Distribuição das classes	135 exemplos C[0] C[1] C[2] 44 36 55	109 exemplos C[0] C[1] C[2] 38 16 55	289 exemplos C[1] C[2] C[3] 59 52 178	176 exemplos C[0] C[1] C[2] 0 0 176

Filtro: RemoveMisclassified (RM) de cada classificador

	135 exemplos	
Classificador	**Sem Filtro**	**Com Filtro**
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.)

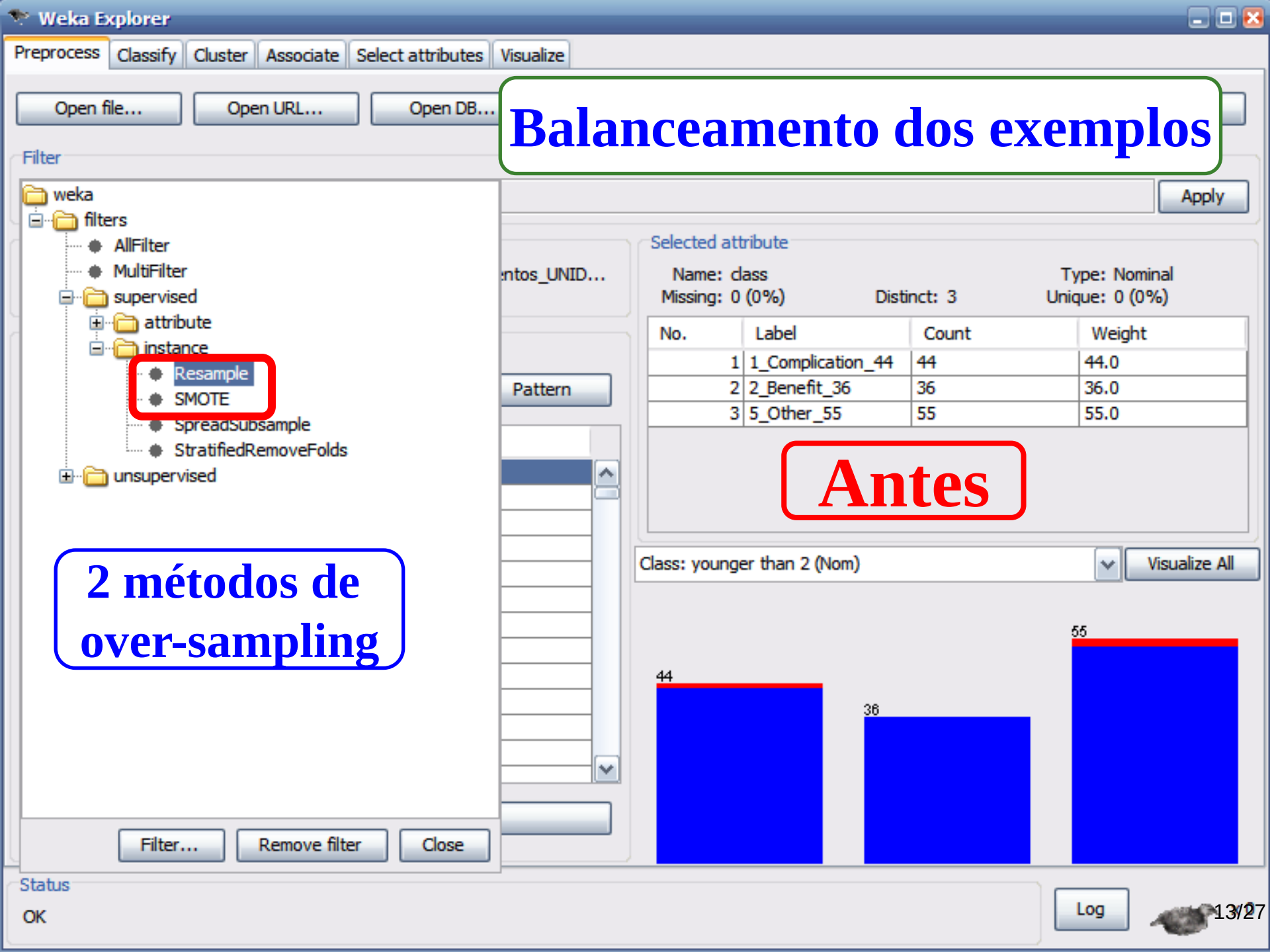
289 exemplos

Sem filtro				**RM do NB**				**RM do OneR**				**RM do J48**			
Dist	C[1]	C[2]	C[3]	Dist	C[1]	C[2]	C[3]	Dist	C[1]	C[2]	C[3]	Dist	C[1]	C[2]	C[3]
3	59	52	178	3	0	0	176	3	13	0	174	3	44	31	176

Filtro: RemoveMisclassified do J48

Mover (Stratified CV) 135 exemplos	**Mover (Stratified CV)** 289 exemplos
70,77% NB	68,21% NB

	135 exemplos - ganho de **5,26%**		289 exemplos - Ganho de **11,07%**	
Classificador	**Sem Filtro**	**Com Filtro**	**Sem Filtro**	**Com Filtro**
Naïve Bayes	**78,51%**	**76,03%**	**70,59%**	74,90%
OneR	41,48%	50,41%	63,67%	70,92%
Prism	45,92%	45,45%	48,79%	62,95%
J48	50,37%	65,29%	64,71%	**79,28%**
ID3	54,07%	63,64%	62,28%	68,92%
SVM	71,85%	72,73%	**72,66%**	78,88%
Distribuição das Classes	135 exemplos C[0] C[1] C[2] 44 36 55	109 exemplos C[0] C[1] C[2] 38 16 55	289 exemplos C[0] C[1] C[2] 59 52 178	251 exemplos C[0] C[1] C[2] 44 31 176


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Balanceamento dos exemplos
Filter
weka
filters
AllFilter
MultiFilter
supervised
attribute
instance
Resample
SMOTE
SpreadSubsample
StratifiedRemoveFolds
unsupervised
Apply
Selected attribute
Name: class
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
No.
Label
Count
Weight
1
1_Complication_44
44
44.0
2
2_Benefit_36
36
36.0
3
5_Other_55
55
55.0
Antes
Pattern
2 métodos de over-sampling
Class: younger than 2 (Nom)
Visualize All
44
36
55
Filter...
Remove filter
Close
Status
OK
Log
13/27

Weka Explorer

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

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

Filter

Choose | **Resample** -B 0.0 -S 1 -Z 100.0 | Apply

Depois

Current relation

Relation: C__Documents and Settings_Pablo_Mesa_Documentos_UNID...
Instances: 135 Sum of weights: 135

Selected attribute

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

No.	Label	Count	Weight
1	1_Complication_44	42	42.0
2	2_Benefit_36	46	46.0
3	5_Other_55	47	47.0

Attributes

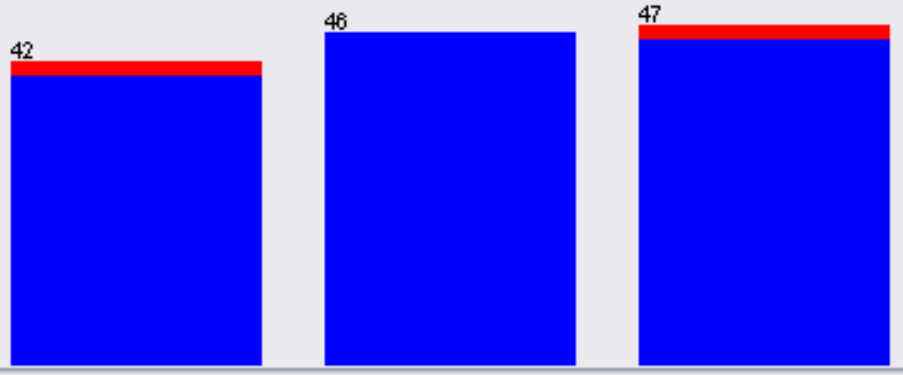
All | ...rn

Mantém a quant. original de exemplos

No.	
1	da
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

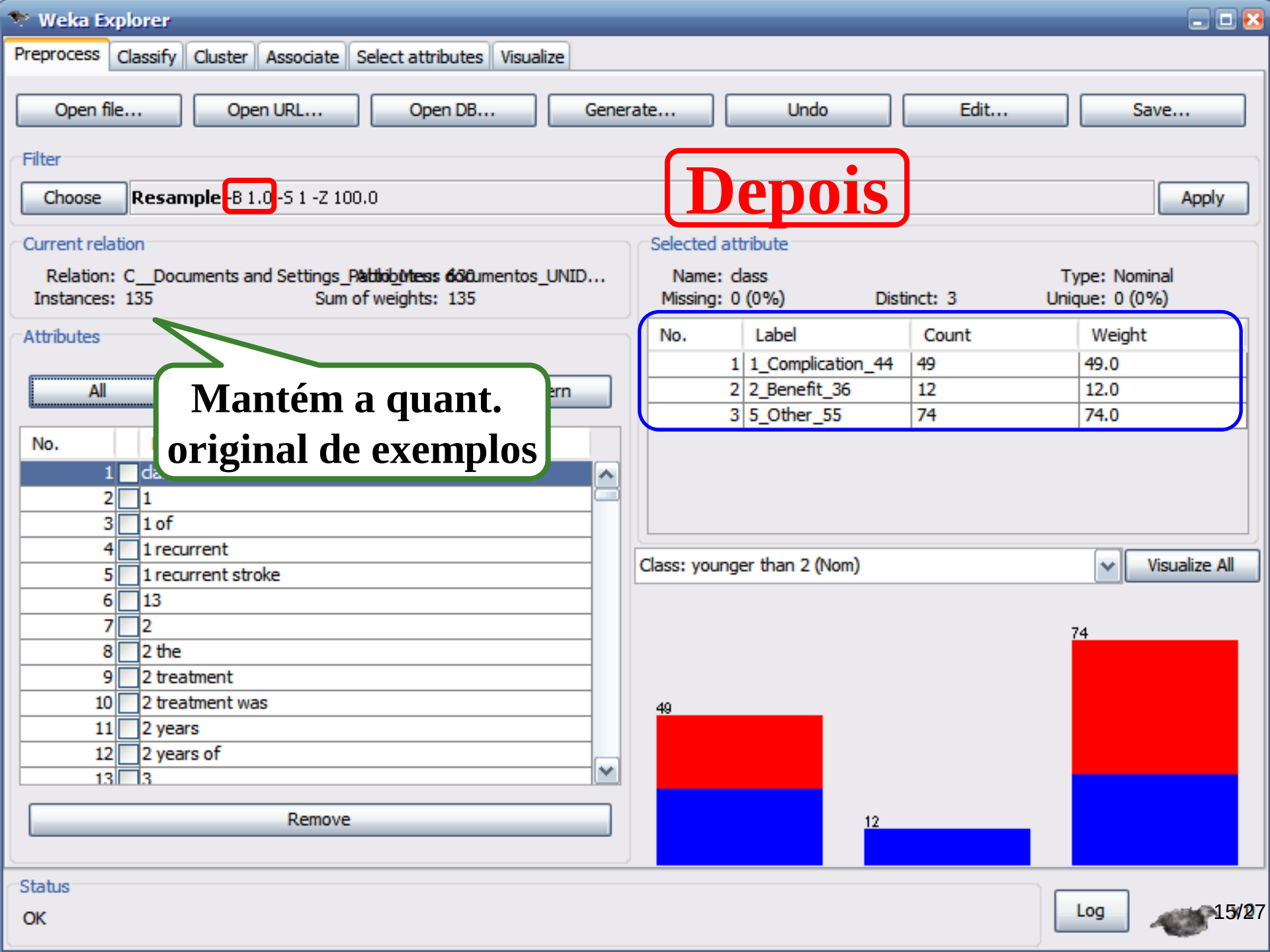
Remove

Class: younger than 2 (Nom) | Visualize All



Status

OK | Log


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
Resample -B 1.0 -S 1 -Z 100.0
Apply
Depois
Current relation
Relation: C__Documents and Settings_...umentos_UNID...
Instances: 135
Sum of weights: 135
Selected attribute
Name: class
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
No. Label Count Weight
1 1_Complication_44 49 49.0
2 2_Benefit_36 12 12.0
3 5_Other_55 74 74.0
Attributes
All
Mantém a quant. original de exemplos
No.
1
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
Class: younger than 2 (Nom)
Visualize All
74
49
12
Status
OK
Log
15/27

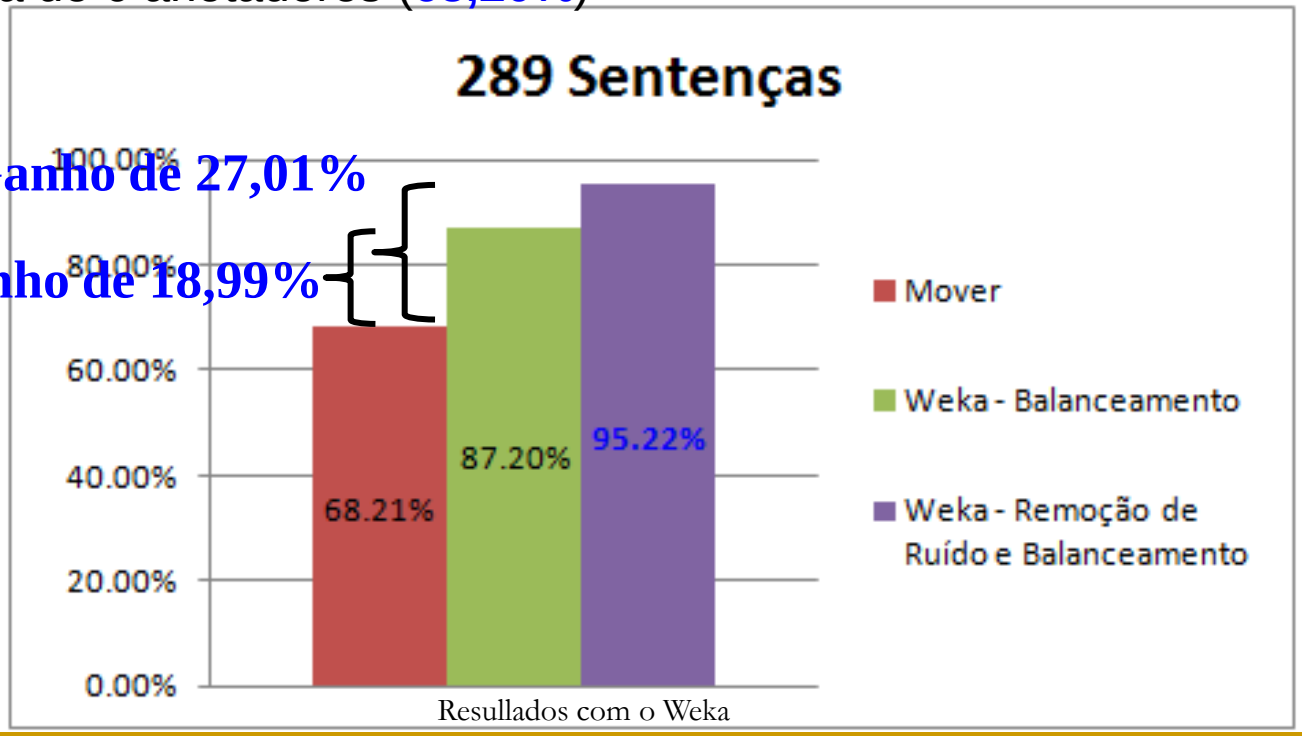
Resumo dos resultados: 3 classes

Seleção de Atributo: Meta

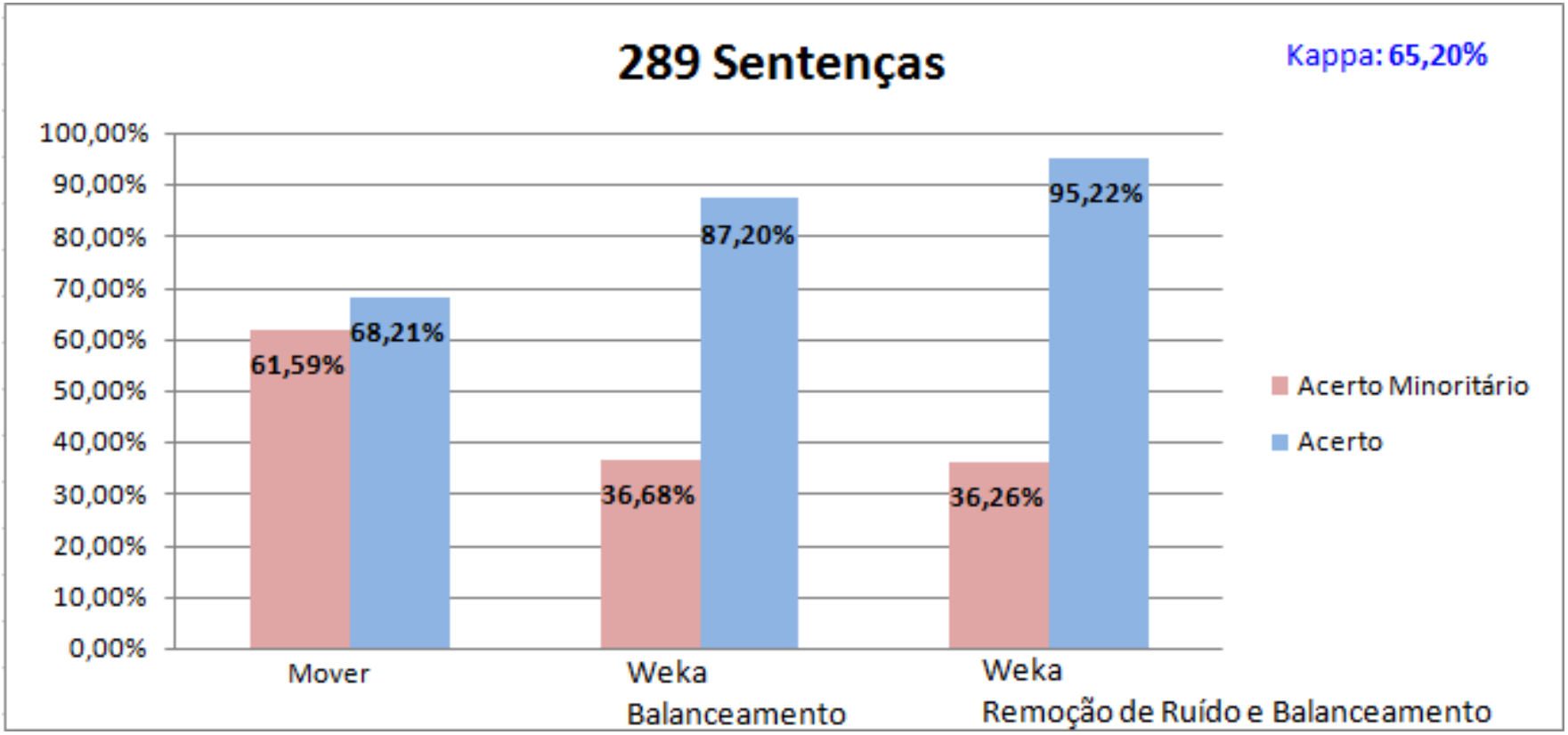
Exemplos	Mover	Sem Filtro	Resample bias 0	RemoveMisclassified (J48) e Resample bias 0
135	70,77% NB	54,07% ID3 71,85% SVM **78,52%** NB 135 exemplos: C[0] C[1] C[2] 44 36 55	83,70% NB 84,44% SVM 85,19% ID3 135 exemplos: C[0] C[1] C[2] 45 47 43 Ganho de **6,67%**	74,38% ID3 80,17% NB **85,12%** SVM 121 exemplos: C[0] C[1] C[2] 41 40 40 Ganho de **6,6%**
289	68,21% NB	64,71% J48 70,59% NB **72,66%** SVM 289 exemplos: C[0] C[1] C[2] 59 52 178	80,97% ID3 83,04 NB 87,20% SVM 289 exemplos: C[0] C[1] C[2] 93 106 90 Ganho de **14,54%**	88,84% NB 90,04% ID3 **95,22%** SVM 251 exemplos: C[0] C[1] C[2] 82 91 78 Ganho de **22,56%**

Análise dos Resultados - Classificação

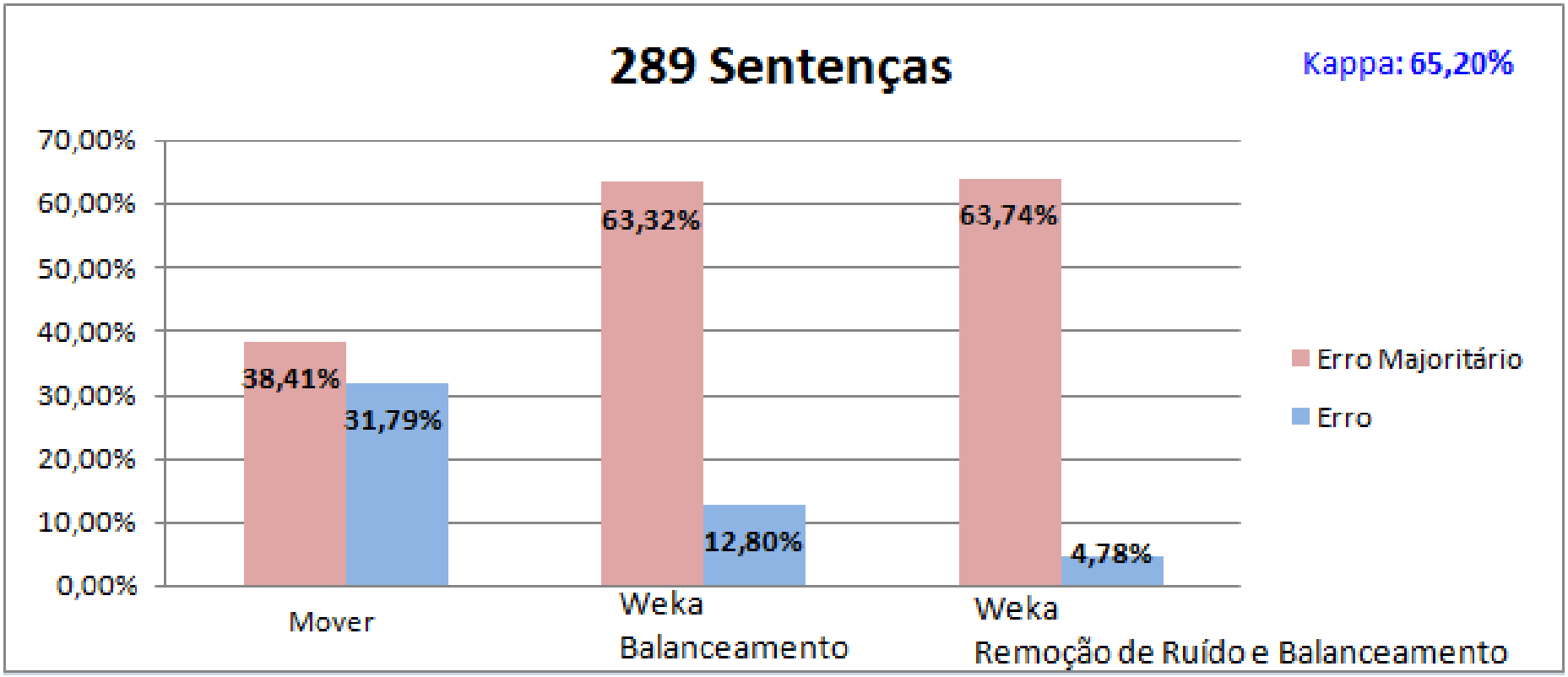
- Taxa de Precisão (P)
  - *Baseline* ≤ *P* ≤ *Topline*
- *Baseline:* Mover
- *Topline*: Taxa de precisão humana
  - 50 sentenças (efeito positivo, efeito negativo e outros)
  - Kappa de 6 anotadores (**65,20%**)



Análise dos Resultados - Classificação



Análise dos Resultados - Classificação



Considerações Sobre a Classificação

- Pré-processamento: 3 classes (*Complicação, Benefício* e *Outro*)
  - Matriz atributo-valor
    - Frequência mínima = 2
    - 1 a 3 gramas
    - sem stopword e sem stemmer
  - Eliminar Ruído com **J48**
  - Balanceamento (Bias 1 em Java - Mantém a distribuição das classes)
  - Seleção de Atributo (Meta)
    - Ganho de Informação
- Classificador
  - Naive Bayes
  - Support Vector Machine

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. RemoveMisclassified
  5. Resample (over sampling)
  6. Igualar os exemplos (under sampling)
  7. Seleção de atributo

Discussão com 135 sentenças

- Randomize
  - influencia negativamente o NB
  - NÃO influencia o SVM
- Attribute Selection
  - NÃO influencia o NB
  - NÃO influencia o SVM
  - Influencia positivamente todos os outros algoritmos testados

Discussão (cont. 1) - com 135 sentenças

Filtro	Problema	Observações
Seleção de Atributo	•Filter e Meta geram atributos diferentes •O Filter é melhor do que o Meta Resultados diferentes	•O Filter não interfere positivamente nem negativamente no resultado do NB e SVM •Entretanto, o Filter melhora o resultado dos outros algoritmos (J48, ID3, Prism, OneR)

Discussão (cont. 2) - 135 sentenças

Filtro	Problema	Definição Bias
Resample	A implementação do Bias 0 no Explorer é similar a implementação do Bias 1 no Java Resultados diferentes **- Algo está errado!!!**	A value of 0 leaves the class distribution as-is. A value of 1 ensures the class distribution is uniform in the output data.

Distribuição Inicial das sentenças		
C[0]	C[1]	C[2]
44	36	55

Total: 135 sentenças

	Explorer			Java		
Bias 0	C[0]	C[1]	C[2]	C[0]	C[1]	C[2]
	42	46	47	49	26	60
Bias 1	C[0]	C[1]	C[2]	C[0]	C[1]	C[2]
	49	12	74	45	47	43

Discussão (cont. 3) – 135 sentenças

Filtro: Resample

	Explorer	Java
Bias 0	C[0] C[1] C[2] 49 12 74	C[0] C[1] C[2] 49 26 60
Bias 1	C[0] C[1] C[2] 42 46 47	C[0] C[1] C[2] 45 47 43

	Explorer	Java
Bias 0	J48 - 85,18 % - 3º melhor NB - 88,14% - 2º melhor SVM - 92,59%	NB - 78,52% - 2º pior ID3 - 86,67% - 2º melhor SVM - 91,11%
Bias 1	ID3 - 77,03 % - 3º melhor SVM - 82,96% - 2º melhor NB - 83,70%	NB - 83,70% - 3º melhor SVM - 84,44% - 2º melhor ID3 - 85,19%

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

Experimentos Mover x Weka

Obrigado!!!

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