Trabalho Ferramenta Weka:   
Classificação, Associação e Clusterização

Marcelo A. R. d’Almeida

Instituto de Computação – Universidade Federal Fluminense (UFF)  
Av. Gal. Milton Tavares de Souza, s/nº, 24210-346 – Niterói – RJ – Brazil

md@id.uff.br

**Abstract.** This work aims to apply the classification, association and clustering tasks, by using the Weka tool in selected datasets. Also seeks to review the results obtained by manipulating the parameters of the algorithms used.

**Resumo.** Este trabalho tem por objetivo aplicar as tarefas de Classificação, Associação e Clusterização, através da utilização da ferramenta Weka, em datasets escolhidos. Também visa também a avaliação dos resultados obtidos com a manipulação dos parâmetros dos algoritmos utilizados.

# 1. Ferramenta Weka

# A ferramenta Weka é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Algoritmos que podem ser aplicados diretamente aos detasets ou serem chamados através de código Java. [1]

# Possui ferramentas de pré-processamento, classificação, regressão, clusterização, regras de associação e visualização. [1]

# 2. Base de Dados

Duas bases foram escolhidas: Tennis Major Tournaments Match Statistics Dataset para as tarefas de Classificação e Clusterização; e Contact Lens Dataset para a tarefa de Associação.

# 2.1. Tennis Major Tournaments Match Statistics Dataset

A base consiste em 8 arquivos contendo estatísticas de partidas masculinas e femininas nos 4 maiores torneios de tênis, no ano de 2013. Os torneios são: Australian Open; French Open - Roland Garros; Wimbledon; e US Open. Cada arquivo contém 42 colunas e mínimo de 76 linhas. O dataset foi obtido no UCI Machine Learning Repository. [2]

**2.1.1. Atributos**

|  |  |  |
| --- | --- | --- |
| **Variável** | **Significado** | **Tipo** |
| Player (1 ou 2) | Nome do jogador | Nominal |
| Result | ‘1’ se Jogador *1 venceu*,  ‘0’ se Jogador 2 venceu | (0/1) |
| FSP (.1 ou .2) | **Primeiro Serviço (**porcentagem) | Real Number |
| FSW (.1 ou .2) | **Primeiro Serviço (ganho)** | Real Number |
| SSP (.1 ou .2) | **Primeiro Serviço** (porcentagem) | Real Number |
| SSW (.1 ou .2) | **Segundo Serviço** (ganho) | Real Number |
| ACE (.1 ou .2) | ***Aces*** ganhos | Numeric-Integer |
| DBF (.1 ou .2) | **Faltas duplas** cometidas | Numeric-Integer |
| WNR (.1 ou .2) | ***Winners*** ganhos | Numeric |
| UFE (.1 ou .2) | **Erros** **‘não forçados’** cometidos | Numeric |
| BPC (.1 ou .2) | ***Break Points*** criados | Numeric |
| BPW (.1 ou .2) | ***Break Points*** ganhos | Numeric |
| NPA (.1 ou .2) | ***Net Points*** tentados | Numeric |
| NPW (.1 ou .2) | ***Net Points*** ganhos | Numeric |
| TPW (.1 ou .2) | Total de **pontos ganhos** | Numeric |
| ST1 .. ST5 (.1 ou .2) | Resultado do ***SET*** | Numeric-Integer |
| FNL (.1 ou .2) | Números de ***Games* vencidos** | Numeric-Integer |
| Round | **Rodada** do torneio | Numeric-Integer |

# Tabela 1 – Atributos do Tennis Major Tournaments Match Statistics Dataset

**2.2. Contact Lens Dataset**

Este dataset também se encontra disponível no UCI Machine Learning Repository, porém, por conveniência de já se encontrar no formato requerido pelo Weka, o ‘.arff’, foi retirado do Sample Weka Data Sets. [3]

O dataset reúne informações que determinam se o paciente deveria (ou não) usar lentes de contato e se elas deveriam ser rígidas ou maleáveis. Contém 5 atributos e 24 instâncias

**2.2.1. Atributos**

|  |  |  |
| --- | --- | --- |
| **Variável** | **Significado** | **Valores possíveis** |
| Age | Idade | Young, Pre-presbyopic, Presbyopic |
| Spectacle prescription | Se a pessoa possui **Miopia** ou **Hipermetropia** | Myope, Hypermetrope |
| Astigmatic | **Astigmatismo** | Yes, No |
| Tear production rate | **Taxa de produção de lágrima** | Reduced, Normal |
| Contact lenses | **Tipo de lente de contato** | Soft, Hard, None |

# Tabela 2 – Atributos do Contact Lens Dataset

# 3. Classificação

Dentre as bases disponíveis no dataset, foi escolhido um torneio e uma modalidade ao caso para a realização dos experimentos. Os algoritmos de classificação foram escolhidos de modo à comparar diferentes abordagens. As seguintes abordagens foram escolhidas: Árvore de decisão; k-neareast-neighbor (k vizinhos mais próximos); e Probabilístico.

- Base utilizada: Tennis Major Tournaments Match Statistics Dataset:

-Torneio: AustraliaOpen

- Modalidade: Masculino (men)

Algoritmos usados:

- NaiveBayes

- IBk (k-NN)

- J48 (Árvore de decisão)

Dado que se trata de uma base de dados com poucos registros, o Leave-one-out foi escolhido como técnica de avaliação.

**3.1. Pré-processamento**

Os testes foram realizados com duas configurações da base: 1 – Todos os atributos mantidos como no original; e 2 – Atributos ‘trivialmente óbvios’ removidos.

Remover os ‘trivialmente óbvios significa remover atributos que possam dar uma informação direta do resultado da partida (atributo classe).

Atributos ‘trivialmente óbvios’ removidos

- Número total de Games vencidos

- Resultados dos Sets

- Número total de pontos

- Números de Winners obtidos

Em ambas as configurações, o atributo ‘Result’ foi convertido para binário (antes representado como valores inteiros 0 ou 1) e usado como atributo classe.

Em cada seção descrevendo um algoritmo estará descrito os parâmetros testados.

**3.2. NaiveBayes**

Parâmetros testados:

- Default

- useKernelEstimator = True

- useSupervisedDiscretization = True

|  |  |  |  |
| --- | --- | --- | --- |
| **Todos os atributos** | **Default** | **Use Kernel Estimator** | **Use Supervised Discretization** |
| **Classificadas corretamente** | 98.4127 % | **99.2063 %** | 97.619 % |

# Tabela 3 – Resultado - todos os atributos

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **a** | **b** |  | **a** | **B** |  | **a** | **b** |  |
| **Matriz de confusão** | 57 | 2 | **a** | 58 | 1 | **a** | 58 | 1 | **a** |
| 0 | 67 | **b** | 0 | 67 | **b** | 2 | 65 | **b** |

# Tabela 4 – Matriz de Confusão referente ao resultado da acima

|  |  |  |  |
| --- | --- | --- | --- |
| **‘Trivialmente óbvios’ removidos** | **Default** | **Use Kernel Estimator** | **Use Supervised Discretization** |
| **Classificadas corretamente** | 85.7143 % | **88.8889 %** | 77.7778 % |

# Tabela 5 – Resultado – ‘trivialmente óbvios’ removidos

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **a** | **b** |  | **a** | **B** |  | **a** | **b** |  |
| **Matriz de confusão** | 55 | 4 | **a** | 56 | 3 | **a** | 48 | 11 | **a** |
| 14 | 53 | **b** | 11 | 56 | **b** | 17 | 50 | **b** |

# Tabela 6 – Matriz de Confusão referente ao resultado da acima

**3.3. IBk**

Parâmetros testados:

- Variação do k de 1 – 20

|  |  |  |  |
| --- | --- | --- | --- |
| **Todos os atributos** | **Classificados corretamente** |  | **Classificados corretamente** |
| **k = 1** | 92.8571 % | **k = 11** | 96.8254 % |
| **k = 2** | 93.6508 % | **k = 12** | 96.8254 % |
| **k = 3** | 95.2381 % | **k = 13** | 97.619 % |
| **k = 4** | 95.2381 % | **k = 14** | 97.619 % |
| **k = 5** | 94.4444 % | **k = 15** | 97.619 % |
| **k = 6** | 95.2381 % | **k = 16** | 97.619 % |
| **k = 7** | 96.8254 % | **k = 17** | 97.619 % |
| **k = 8** | 97.619 % | **k = 18** | **98.4127 %** |
| **k = 9** | 96.0317 % | **k = 19** | 97.619 % |
| **k = 10** | 97.619 % | **k = 20** | 97.619 % |

# Tabela 7 – Resultado - todos os atributos

|  |  |  |  |
| --- | --- | --- | --- |
| **Todos os atributos** | **k = 1** | **k = 2** | **k = 18** |
| **Classificadas corretamente** | 92.8571 % | 93.6508 % | **98.4127 %** |

# Tabela 8 – Resultados selecionados - todos os atributos

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **a** | **b** |  | **a** | **B** |  | **a** | **b** |  |
| **Matriz de confusão** | 53 | 6 | **a** | 58 | 1 | **a** | 58 | 1 | **a** |
| 3 | 64 | **b** | 7 | 60 | **b** | 1 | 66 | **b** |

# Tabela 9 – Matriz de Confusão referente ao resultado da acima

|  |  |  |  |
| --- | --- | --- | --- |
| **‘Trivialmente óbvios’ removidos** | **Classificados corretamente** |  | **Classificados corretamente** |
| **k = 1** | 72.2222 % | **k = 11** | 84.127 % |
| **k = 2** | 78.5714 % | **k = 12** | 86.5079 % |
| **k = 3** | 76.1905 % | **k = 13** | 85.7143 % |
| **k = 4** | 81.746 % | **k = 14** | 85.7143 % |
| **k = 5** | 80.9524 % | **k = 15** | 87.3016 % |
| **k = 6** | 79.3651 % | **k = 16** | 86.5079 % |
| **k = 7** | 82.5397 % | **k = 17** | 86.5079 % |
| **k = 8** | 84.127 % | **k = 18** | 86.5079 % |
| **k = 9** | 81.746 % | **k = 19** | **88.0952 %** |
| **k = 10** | 83.3333 % | **k = 20** | **88.0952 %** |

# Tabela 10 – Resultado – ‘trivialmente óbvios’ removidos

|  |  |  |  |
| --- | --- | --- | --- |
| **‘Trivialmente óbvios’ removidos** | **k = 1** | **k = 2** | **k = 20** |
| **Classificadas corretamente** | 72.2222 % | 78.5714 % | **88.0952 %** |

# Tabela 11 – Resultados selecionados – ‘trivialmente óbvios’ removidos

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **a** | **b** |  | **a** | **B** |  | **a** | **b** |  |
| **Matriz de confusão** | 38 | 21 | **a** | 51 | 8 | **a** | 50 | 9 | **a** |
| 14 | 53 | **b** | 19 | 48 | **b** | 6 | 61 | **b** |

# Tabela 12 – Matriz de Confusão referente ao resultado da acima

**3.4. J48**

Parâmetros testados:

- Default

- ReducedErrorPruning = True

|  |  |  |
| --- | --- | --- |
| **Todos os atributos** | **Default** | **Reduced Error Pruning** |
| **Classificadas corretamente** | 97.619 % | **98.4127 %** |

# Tabela 13 – Resultados selecionados - todos os atributos

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **a** | **b** |  | **a** | **b** |  |
| **Matriz de confusão** | 57 | 2 | **a** | 57 | 2 | **a** |
| 1 | 66 | **b** | 0 | 67 | **b** |

# Tabela 14 – Matriz de Confusão referente ao resultado da acima

|  |  |  |
| --- | --- | --- |
| **‘Trivialmente óbvios’ removidos** | **Default** | **Reduced Error Pruning** |
| **Classificadas corretamente** | 84.9206 % | **86.5079 %** |

# Tabela 15 – Resultados selecionados - ‘trivialmente óbvios’ removidos

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **a** | **b** |  | **a** | **b** |  |
| **Matriz de confusão** | 48 | 11 | **a** | 53 | 6 | **a** |
| 8 | 59 | **b** | 11 | 56 | **b** |

# Tabela 16 – Matriz de Confusão referente ao resultado da acima

**3.5. Comparação**

Três métricas foram utilizadas na comparação: A porcentagem de instâncias classificadas corretamente; F-Measure (média ponderada); e ROC área (média ponderada). NaiveBayes foi o algoritmo que se saiu melhor, como pode ser conferido nos gráficos a seguir:

**3.5.1. Instâncias Classificadas Corretamente**

Gráfico 1 - Instâncias classificadas corretamente

**3.5.2. F-Measure**

Gráfico 2 – F-Measure

**3.5.3. ROC Area**

Gráfico 3 – ROC area

# 4. Associação

Ao realizar experimentos relacionados a tarefa de Associação com o dataset das partidas de tênis, percebeu-se que a quantidade de regras óbvias unido com uma necessidade de discretização mais apurada, idealmente realizada por um especialista no domínio, favoreceram pela escolha da não utilização desta base na seção de Associação. Invés disso, para apresentar um resultado mais interessante e proveitoso neste estudo introdutório à ferramenta, foi escolhido em seu lugar, o dataset das lentes de contato, uma base mais comportada e com atributos categóricos, eliminando também o pré-processamento necessário para aplicação do algoritmo Apriori (discretização de valores contínuos).

Base utilizada: Contact Lens Dataset

Algoritmo usado:

- Apriori

Parâmetros testados

- Suporte mínimo: 0.1, 0.2 e 0.3

- Confiança mínima: 0.75, 0.8 e 1

- Lift: 1.5, 2.0, 4.0 e 6.0

- Car: True; False

**4.1. Apriori**

As tabelas a seguir representam o total de regras mineiradas.

|  |  |  |  |
| --- | --- | --- | --- |
| **Conf \ MinSup (‘car’=True)** | **0.1** | **0.2** | **0.3** |
| **0.75** | **108 (23)** | **18 (8)** | **2 (-)** |
| **0.8** | **89 (16)** | **15 (7)** | **2 (-)** |
| **1.00** | **83 (14)** | **10 (5)** | **1 (-)** |

# Tabela 17 – Total de regras mineiradas (Conf – MinSup)

|  |  |  |  |
| --- | --- | --- | --- |
| **Lift \ MinSup** | **0.1** | **0.2** | **0.3** |
| **1.5** | **320** | **28** | **2** |
| **2.0** | **166** | **10** | **–** |
| **4.0** | **26** | **2** | **-** |
| **6.0** | **4** | **-** | **-** |

# Tabela 18 – Total de regras mineiradas (Lift – MinSup)

**4.1.2. Regras interessantes**

Dentre as regras mineiradas, algumas das mais importantes, se não as mais importantes, foram selecionadas:

# contact-lenses=soft ==> astigmatism=no tear-prod- rate=normal minsup:(0.2) conf:(1) <lift:(4)> astigmatism=no tear-prod-rate=normal ==> contact- lenses=soft minsup:(0.2) conf:(0.83) <lift:(4)> tear-prod-rate=reduced ==> contact-lenses=none minsup:(0.3) conf:(1) <lift:(1.6)> spectacle-prescrip=myope astigmatism=yes tear-prod-rate=normal ==> contact-lenses=hard minsup:(0.1) conf:(1) <lift:(6)>

# age=young astigmatism=yes tear-prod-rate=normal ==> contact-lenses=hard minsup:(0.1) conf:(1) <lift:(6)> spectacle-prescrip=hypermetrope astigmatism=yes ==> contact-lenses=none minsup:(0.2) conf:(0.83)

age=presbyopic ==> contact-lenses=none   
minsup:(0.2) conf:(0.75)

spectacle-prescrip=hypermetrope astigmatism=yes ==> contact-lenses=none minsup:(0.1) conf:(0.83)

age=presbyopic ==> contact-lenses=none   
minsup:(0.1) conf:(0.75)

**‘car’ = True ==> No Lift measured**

# 5. Clusterização

# Mesma base utilizada na classificação foi utilizada para clusterização.

# Base utilizada: Tennis Major Tournaments Match Statistics Dataset

# - Torneio: AustraliaOpen

# - Modalidade: Masculino (men)

# Foi retirado o atributo ‘Result’, atributo tido como classe, para que não interfira no resultado da clusterização.

# Algoritmos usados:

# - SimpleKMeans (k-means)

# - DBSCAN

# 5.1. *SimpleKMeans*

# Parâmetros testados:

# - Número de clusters: 2, 3, 4 e 5

|  |  |
| --- | --- |
| **k - cluster \ Distância** | **Euclidiana** |
| **2 – 0**  **- 1** | 49 (39%)  77 (61%) |
| **3 – 0**  **- 1**  **- 2** | 21 (17%)  33 (26%)  72 (57%) |
| **4 – 0**  **- 1**  **- 2**  **- 3** | 23 (18%)  33 (26%)  32 (25%)  38 (30%) |
| **5 – 0**  **- 1**  **- 2**  **- 3**  **- 4** | 17 (13%)  25 (20%)  21 (17%)  17 (13%)  46 (37%) |

# Tabela 19 – Resultado da clusterização

# 5.1.1. Análise de desvio padrão

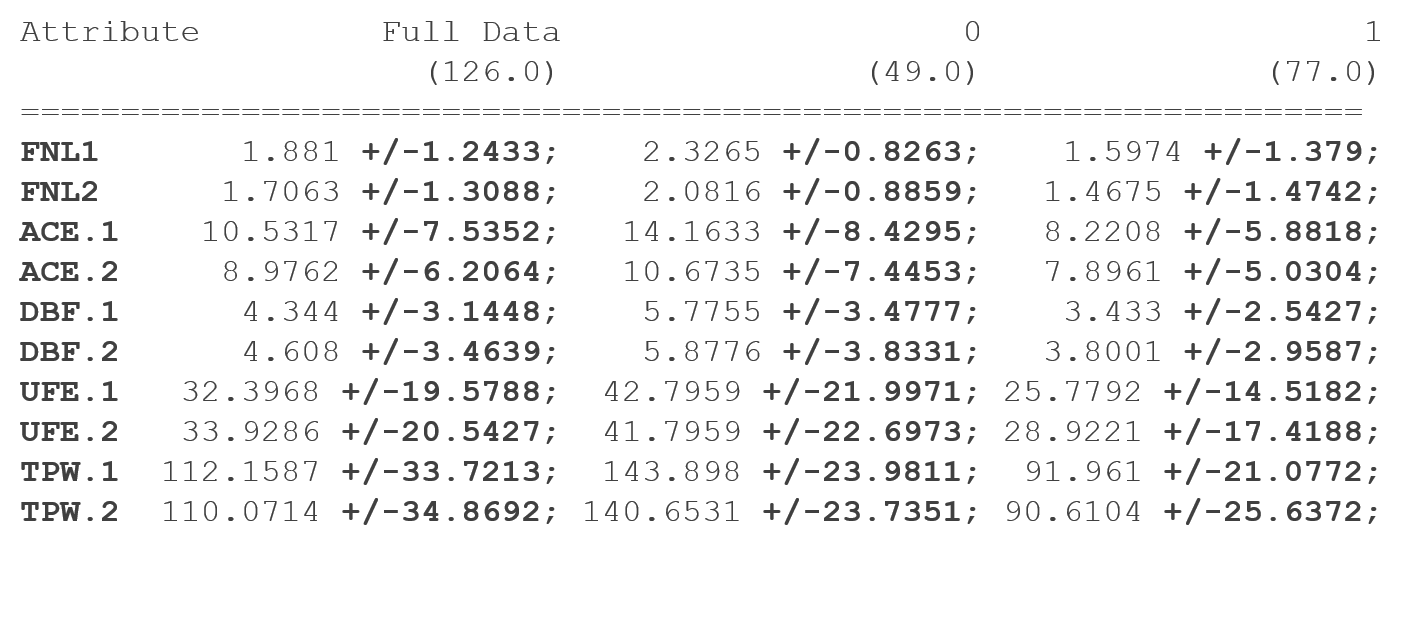


Figura 1 – Análise de desvio padrão para k = 2

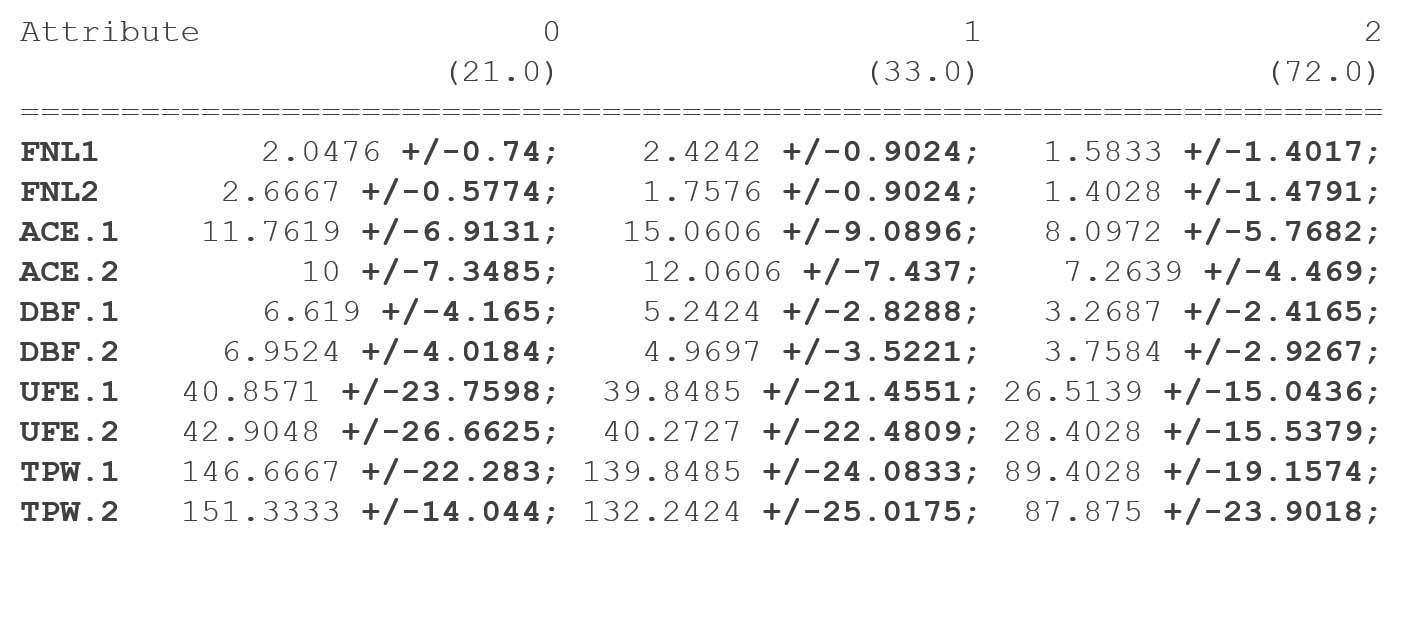


Figura 2 – Análise de desvio padrão para k = 3

# 5.2. *DBSCAN*

# Parâmetros testados:

# - Epsilon (raio): 2, 2.37 e 2.5

# - Mínimo de pontos (densidade) 2, 3, 6 e 10

# - Default = (raio = 0.9, densidade = 6)

# - (raio = 4, densidade = 20)

|  |  |  |
| --- | --- | --- |
| **Configuração (raio, densidade) - cluster** | **Resultado da clusterização** | **Instâncias não clusterizadas** |
| **Default (0.9, 6)** | - | 126 |
| **(2.5, 10) - 0** | 80 (100%) | 46 |
| **(2.5, 6) - 0** | 89 (100%) | 37 |
| **(2.5, 3) - 0**  **- 1**  **- 2** | 97 (92%)  5 (5%)  3 (3%) | 21 |
| **(2.5, 2) - 0**  **- 1**  **- 2**  **- 3**  **- 4** | 97 (89%)  5 (5%)  2 (2%)  3 (3%)  2 (2%) | 17 |
| **(2, 10) - 0** | 22 (100%) | 104 |
| **(2, 6) - 0**  **- 1** | 23 (77%)  7 (23%) | 96 |
| **(2, 3) - 0**  **- 1**  **- 2**  **- 3** | 7 (16%)  24 (53%)  11 (24%)  3 (7%) | 81 |
| **(2, 2) - 0**  **- 1**  **- 2**  **- 3**  **- 4**  **- 5**  **- 6** | 7 (14%)  24 (24%)  11 (22%)  2 (4%)  2 (4%)  2 (4%)  3 (6%) | 75 |
| **(2.37, 10) - 0**  **- 1** | 33 (49%)  35 (51%) | 58 |
| **(2.37, 6) - 0**  **- 1** | 33 (44%)  42 (56%) | 51 |
| **(2.37, 3) - 0**  **- 1** | 33 (39%)  51 (61%) | 42 |
| **(2.37, 2) - 0**  **- 1**  **- 2**  **- 3**  **- 4** | 33 (37%)  51 (57%)  2 (2%)  2 (2%)  2 (2%) | 36 |
| **(4, 20) - 0** | 126 (100%) | - |

# Tabela 20 – Resultado da clusterização

# 5.2.1. Comparação entre parâmetros

|  |  |  |  |
| --- | --- | --- | --- |
| **Raio \ Densidade** | **Relativamente pequena** | **Supostamente ideal** | **Relativamente grande** |
| **Relativamente pequeno** | Não efetua clusterização (muitos ‘não clusterizados’) | | | |  | |  | | Clusterização balanceada, ‘não clusterizados’ aumentam |
| **Supostamente ideal** | –  ––  ––– | Melhor resultado | –  ––  ––– |
| **Relativamente grande** | Clusterização desbalanceada, mais clusters, ‘não clusterizados’ diminuem | |  | |  | | | | Um cluster contendo a maioria dos elementos |

# Tabela 21 – Comparação entre os parâmetros

# 6. Trabalhos Futuros

Como possíveis trabalhos futuros que expandam o estudo exploratório com a ferramenta Weka: Modificar o dataset para avaliar jogadores individualmente invés de partidas; Avaliar as diferenças entre os toneios; Avaliar a diferença entre as partidas femininas e masculinas; e Discretizar adequadamente a base das partidas de tênis para avaliar associações adequadas entre os atributos.

# Referências

[1]http://www.cs.waikato.ac.nz/ml/weka/

[2]https://archive.ics.uci.edu/ml/datasets/Tennis+Major+Tournament+Match+Statistics

[3]http://storm.cis.fordham.edu/~gweiss/data-mining/datasets.html