

**Universidade do Minho**

Departamento de Informática

Mestrado em Engenharia Informática  
Mestrado Integrado em Engenharia Informática

Perfil de Sistemas Inteligentes

Unidade Curricular de Aprendizagem e Extração de Conhecimento

2015/2016

Aprendizagem e Extração de Conhecimento

André Geraldes a67673

Bruno Barbosa a67646

Francisco Dourado pg30458

**Abstract**

This report describes all the knowledge extraction procedures used in three distinct datasets. Every decision made throughout the project is presented as well as the relevant implementation details explored.

**Resumo**

Neste relatório é descrito todos os procedimentos de extração de conhecimento usados em três conjuntos de dados distintos. São apresentadas todas as decisões tomadas e explorados durante todo o processo assim como os detalhes de implementação mais relevantes.

Índice

[1 Introdução 4](#_Toc437790327)

[1.1 Enquadramento e objetivos 4](#_Toc437790328)

[1.2 Planeamento 5](#_Toc437790329)

[1.3 Estrutura do documento 5](#_Toc437790330)

[2 Desenvolvimento 6](#_Toc437790331)

[2.1 Dataset de Avaliação de Automóveis 6](#_Toc437790332)

[2.2 Dataset do Consumo de Energia Elétrica 6](#_Toc437790333)

[2.3 Dataset da Central de Energia Elétrica 7](#_Toc437790334)

[2.3.1 Preparação do Dataset 7](#_Toc437790335)

[2.3.2 Associação 7](#_Toc437790336)

[2.3.3 Classificação 8](#_Toc437790337)

[2.3.4 Segmentação 9](#_Toc437790338)

[2.4 Dataset da Idade de Moluscos Marinhos 10](#_Toc437790339)

[2.4.1 Discretização do Dataset 10](#_Toc437790340)

[2.4.2 Regras de Associação 10](#_Toc437790341)

# Introdução

Os dados recolhidos a partir do mundo real não podem ser imediatamente utilizados para a extração de conhecimento visto que possuem imensos problemas. Problemas esses que podem estar relacionados com inconsistências, informação incompleta ou até mesmo lixo.

Surge, deste modo, a necessidade de tratar esses dados com a devida antecedência para que se adequem a uma determinada ferramenta de extração de conhecimento. A preparação dos dados para a extração de conhecimento segue um conjunto de passos bem definidos: seleção, pré-processamento, transformação, mineração dos dados (*Data Mining*) e interpretação.

Discretização, limpeza, integração, transformação e redução, são exemplos de tarefas indispensáveis desde que selecionamos os dados até à extração de conhecimento a partir dos mesmos.

## Enquadramento e objetivos

Com a liberdade de escolha dos conjuntos de dados a estudar, o nosso grupo focou-se em três *datasets* distintos. Um primeiro relacionado com a avaliação de automóveis, um outro associado ao consumo de energia elétrica e um terceiro relativo à energia gerada numa central elétrica. A escolha destes *datasets* teve em conta algumas considerações que achamos importante referir. O número de atributos deverá estar compreendido entre 3 a 6 unidades e o número de instâncias não deverá ser superior a 10000 nem inferior a 1500 unidades. A razão desta escolha deve-se ao facto de querermos trabalhar com *datasets* minimamente robustos mas que simultaneamente não sejam demasiado grandes.

Os objetivos deste trabalho consistem assim, na análise, interpretação e transformação dos dados e na extração de conhecimento através dos diversos relacionamentos entre os atributos.

## Planeamento

Cada elemento do grupo ficou responsável por tratar de um *dataset*. Neste caso, a distribuição foi feita da seguinte forma:

* André Geraldes – Avaliação de automóveis
* Bruno Barbosa – Consumo de energia elétrica
* Francisco Dourado – Energia gerada por uma central elétrica

## Estrutura do documento

O presente documento está dividido em dois capítulos e um anexo. No primeiro capítulo é feita uma breve exposição do problema descrevendo as considerações iniciais decididas pelo grupo.

No segundo capítulo, encontra-se toda a parte referente ao tratamento dos dados, que inclui uma exposição da cada *dataset* e todos os procedimentos aplicados sobre eles.

Nos anexos estão expostas algumas imagens que ajudam a complementar o trabalho desenvolvido.

# Desenvolvimento

Neste capítulo são apresentados todos os procedimentos utilizados durante a análise e o tratamento dos vários *datasets*, bem como as questões mais relevantes a eles associadas.

## Dataset de Avaliação de Automóveis

## Dataset do Consumo de Energia Elétrica

## Dataset da Central de Energia Elétrica

Este Dataset denota a produção de energia de uma central elétrica de ciclo combinado, entre 206 e 2011. O Dataset tem 9568 entradas, com os campos com a seguinte nomenclatura:

|  |  |  |
| --- | --- | --- |
| AT | Atmospheric Temperature | Temperatura Atmosférica |
| V | Exhaust Vacuum | Vácuo gerado pelo Sistema de Exaustão |
| RH | Relative Humidity | Humidade Relativa |
| AP | Ambient Pressure | Pressão Ambiente |
| PE | Eletrical Energy Production | Produção de Energia Elétrica |

Uma central elétrica de ciclo combinado é composta por turbinas de gás, turbinas a vapor e geradores de vapor via recuperadores de calor. Dos campos acima identificados, o campo referente ao Vácuo influencia o funcionamento das turbinas a vapor. Os campos referentes à Temperatura Atmosférica, à Humidade Relativa e à Pressão Ambiente influenciam a turbina a gás.

### Preparação do Dataset

Utilizando a ferramenta Weka, realizou-se uma discretização dos valores de todos os campos do Dataset em 5 intervalos de valores com largura idêntica. Os valores discretos foram classificados da seguinte forma:

|  |  |  |
| --- | --- | --- |
| VL | Very Low | Muito Baixo |
| L | Low | Baixo |
| A | Average | Médio |
| H | High | Alto |
| VH | Very High | Muito Alto |

### Associação

Como este Dataset contém um conjunto de dados acerca das condições ambientais e atmosféricas e a quantidade de energia produzida nessas condições, o algoritmo **Apriori** foi aplicado tendo em conta uma associação à classe PE. Todos os parâmetros definidos foram:

|  |  |  |
| --- | --- | --- |
| car | True | Encontrar regras associadas à classe. |
| classIndex | -1 | Usar PE (última classe) como a classe para encontrar regras |
| delta | 0.005 | Decrementar o facto de suporte por este valor |
| minMetric | 0.7 | Mínimo de 0.7 de confiança nas associações |
| numRules | 100 | Encontrar, no máximo, 100 regras |

Os resultados obtidos foram os seguintes:

1. AT=H V=H 1575 ==> PE=L 1198 conf:(0.76)
2. V=H AP=A 1636 ==> PE=L 1240 conf:(0.76)
3. AT=H AP=A 1728 ==> PE=L 1289 conf:(0.75)
4. AT=L V=L 2322 ==> PE=H 1681 conf:(0.72)
5. AT=L 2565 ==> PE=H 1797 conf:(0.7)

Aplicou-se também o algoritmo **Apriori** sem a restrição de associação a uma dada classe. Todos os parâmetros definidos foram:

|  |  |  |
| --- | --- | --- |
| delta | 0.005 | Decrementar o facto de suporte por este valor |
| minMetric | 0.7 | Mínimo de 0.7 de confiança nas associações |
| numRules | 100 | Encontrar, no máximo, 100 regras |

Os resultados obtidos, para além dos descritos acima, foram os seguintes:

1. AT=L PE=H 1797 ==> V=L 1681 conf:(0.94)
2. PE=H 2116 ==> V=L 1975 conf:(0.93)
3. AT=L RH=H 1112 ==> V=L 1019 conf:(0.92)
4. AT=L AP=A 1073 ==> V=L 975 conf:(0.91)
5. AT=L 2565 ==> V=L 2322 conf:(0.91)
6. V=L PE=H 1975 ==> AT=L 1681 conf:(0.85)
7. PE=H 2116 ==> AT=L 1797 conf:(0.85)
8. PE=H 2116 ==> AT=L V=L 1681 conf:(0.79)
9. AT=H V=H 1575 ==> PE=L 1198 conf:(0.76)
10. V=H AP=A 1636 ==> PE=L 1240 conf:(0.76)
11. AT=H AP=A 1728 ==> PE=L 1289 conf:(0.75)
12. AT=L V=L 2322 ==> PE=H 1681 conf:(0.72)
13. AT=L 2565 ==> PE=H 1797 conf:(0.7)

Determinados resultados obtidos são de carácter irrelevante no contexto, como por exemplo:

1. PE=H 2116 ==> V=L 1975 conf:(0.93)

É claro que a Produção de Energia não tem qualquer influência no estado do Vácuo, apesar de a associação ter sido encontrada.

### Classificação

Para a Classificação, aplicou-se ao Dataset uma Arvore de Decisão J48. Este algoritmo foi aplicado tendo em atenção que o atributo **PE** fosse usado como a classe. Desta forma, obtemos uma árvore de decisão onde podemos ver as condições necessárias para obtermos uma melhor produção de energia.

A Árvore de Decisão gerada é a seguinte:

**AT = VL: VH (825.0/211.0)**  
**AT = L  
| V = VL: H (71.0/6.0)  
| V = L: H (2322.0/641.0)**  
| V = A: A (159.0/51.0)  
| V = H: A (1.0)  
| V = VH: A (12.0/2.0)  
AT = A  
| V = VL: A (2.0)  
| V = L: A (788.0/175.0)  
| V = A  
| | AP = VL: A (0.0)  
| | AP = L: L (160.0/32.0)  
| | AP = A: A (453.0/172.0)  
| | AP = H: A (198.0/40.0)  
| | AP = VH: A (6.0/1.0)  
| V = H: L (766.0/246.0)  
| V = VH: L (63.0/9.0)  
AT = H  
| V = VL: L (0.0)  
| V = L  
| | RH = VL: L (3.0/1.0)  
| | RH = L: L (38.0/7.0)  
| | RH = A: L (41.0/12.0)  
| | RH = H: A (9.0/3.0)  
| | RH = VH: VL (3.0/1.0)  
| V = A: L (526.0/70.0)  
| V = H  
| | AP = VL: VL (15.0/3.0)  
| | AP = L  
| | | RH = VL: VL (4.0/1.0)  
| | | RH = L: VL (42.0/13.0)  
| | | RH = A: L (171.0/77.0)  
| | | RH = H: L (144.0/57.0)  
| | | RH = VH: L (24.0/9.0)  
| | AP = A: L (1007.0/166.0)  
| | AP = H: L (168.0/24.0)  
| | AP = VH: L (0.0)  
| V = VH  
| | AP = VL: VL (21.0/2.0)  
| | AP = L: VL (400.0/141.0)  
| | AP = A  
| | | RH = VL: VL (1.0)  
| | | RH = L: L (34.0/16.0)  
| | | RH = A: L (98.0/48.0)  
| | | RH = H: VL (159.0/66.0)  
| | | RH = VH: L (80.0/31.0)  
| | AP = H: L (36.0/7.0)| | AP = VH: VL (0.0)  
AT = VH: VL (718.0/212.0)

Podemos ver, marcados a negrito, os caminhos que denotam as condições necessárias para obter uma produção alta ou muito alta de energia elétrica.

### Segmentação

Após a identificação dos caminhos encontrados para uma produção de energia elétrica mais elevada, correu-se um Algoritmo de Segmentação **SimpleKMeans** com **Distancia Euclidiana** para encontrar o padrão de valores precisos mais indicado para uma produção mais eficiente. Como um maior número de segmentos significa um conjunto de parâmetros mais específicos, optou-se por procurar 10 segmentos. Os segmentos seguintes foram encontrados:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| AT | 12.0329 | 23.7021 | 22.7534 | 13.6189 | 8.2729 | 24.5452 | 19.4603 | 28.8567 | 29.2475 | 15.37 |
| V | 41.0251 | 47.5184 | 61.4147 | 43.2812 | 40.2086 | 69.5728 | 53.2346 | 70.4313 | 66.9633 | 42.8143 |
| AP | 1009.3984 | 1008.8114 | 1015.3291 | 1019.6974 | 1020.309 | 1008.8922 | 1011.8665 | 1007.8135 | 1011.6557 | 1016.6665 |
| RH | 86.3302 | 59.2717 | 69.3167 | 83.593 | 81.683 | 83.6243 | 85.7504 | 64.7948 | 48.2171 | 61.406 |
| PE | 472.0184 | 447.9896 | 447.3762 | 466.4503 | 481.74 | 437.6776 | 451.2381 | 434.1948 | 437.4324 | 466.3271 |
|  | 975 (10%) | 529 (6%) | 1169 (12%) | 964 (10%) | 1114 (12%) | 1067 (11%) | 921 (10%) | 1063 (11%) | 928 (10%) | 838 (9%) |

Como podemos verificar, o segmento nº 4 oferece as condições necessárias para uma produção de energia elétrica mais eficiente.

## Dataset da Idade de Moluscos Marinhos

A idade destes moluscos é calculada via o número de anéis existentes na carapaça, depois de a cortar pelo cone e aplicar um corante. A idade é calculada pela soma do número de anéis com 1,5 anos. Utilizando um conjunto de medições físicas, é possível prever a idade do molusco medido. O Dataset utilizado tem 4177 entradas e 9 atributos. Os atributos são:

|  |  |
| --- | --- |
| Sex | {M, F, I(Infantil)} |
| Length | Medição mais longa da carapaça |
| Diameter | Diâmetro, medido perpendicularmente a Length |
| Height | Medição da carapaça com tecido vivo |
| Whole Weight | Peso do Molusco |
| Shucked Weight | Peso do tecido vivo |
| Viscera Weight | Peso após remoção do sangue |
| Shell Weight | Peso após secagem do molusco |
| Rings | Número de anéis do molusco |

### Preparação do Dataset

Este Dataset foi discretizado da seguinte forma:

* **Sex** com os valores (M,F,I);
* **Length** em 10 intervalos equidistantes;
* **Diameter** em 10 intervalos equidistantes;
* **Height** em 10 intervalos equidistantes;
* **WholeWeight** em 10 intervalos equidistantes;
* **ShuckedWeight** em 10 intervalos equidistantes;
* **VisceraWeight** em 10 intervalos equidistantes;
* **ShellWeight** em 10 intervalos equidistantes;
* **Rings** com inteiros de 1 a 29.

### Associação

Para este Dataset, foi aplicado um Algoritmo **Apriori** para regras de associação na classe Rings, com os seguintes parâmetros:

|  |  |  |
| --- | --- | --- |
| car | True | Encontrar regras associadas à classe. |
| classIndex | -1 | Usar Rings (última classe) como a classe para encontrar regras |
| delta | 0.005 | Decrementar o facto de suporte por este valor |
| minMetric | 0.8 | Mínimo de 0.7 de confiança nas associações |
| numRules | 100 | Encontrar, no máximo, 100 regras |

Os resultados obtidos foram os seguintes:

1. Sex=I WholeWeight='(0.28435-0.5667]' 465 ==> Rings='(5.666667-10.333333]' 418 conf:(0.9)
2. Sex=I VisceraWeight='(0.07645-0.1524]' 488 ==> Rings='(5.666667-10.333333]' 434 conf:(0.89)
3. Sex=I ShellWeight='(0.10185-0.2022]' 490 ==> Rings='(5.666667-10.333333]' 434 conf:(0.89)
4. Sex=I ShuckedWeight='(0.1497-0.2984]' 498 ==> Rings='(5.666667-10.333333]' 437 conf:(0.88)
5. WholeWeight='(0.28435-0.5667]' ShuckedWeight='(0.1497-0.2984]' 621 ==> Rings='(5.666667-10.333333]' 520 conf:(0.84)
6. WholeWeight='(0.28435-0.5667]' ShuckedWeight='(0.1497-0.2984]' ShellWeight='(0.10185-0.2022]' 541 ==> Rings='(5.666667-10.333333]' 450 conf:(0.83)
7. Diameter='(0.293-0.3525]' 513 ==> Rings='(5.666667-10.333333]' 426 conf:(0.83)
8. WholeWeight='(0.28435-0.5667]' 783 ==> Rings='(5.666667-10.333333]' 647 conf:(0.83)
9. WholeWeight='(0.28435-0.5667]' ShuckedWeight='(0.1497-0.2984]' VisceraWeight='(0.07645-0.1524]' 515 ==> Rings='(5.666667-10.333333]' 425 conf:(0.83)
10. WholeWeight='(0.28435-0.5667]' VisceraWeight='(0.07645-0.1524]' 569 ==> Rings='(5.666667-10.333333]' 467 conf:(0.82)
11. VisceraWeight='(0.07645-0.1524]' ShellWeight='(0.10185-0.2022]' 754 ==> Rings='(5.666667-10.333333]' 618 conf:(0.82)
12. WholeWeight='(0.28435-0.5667]' VisceraWeight='(0.07645-0.1524]' ShellWeight='(0.10185-0.2022]' 511 ==> Rings='(5.666667-10.333333]' 418 conf:(0.82)
13. Length='(0.445-0.519]' VisceraWeight='(0.07645-0.1524]' ShellWeight='(0.10185-0.2022]' 525 ==> Rings='(5.666667-10.333333]' 429 conf:(0.82)
14. WholeWeight='(0.28435-0.5667]' ShellWeight='(0.10185-0.2022]' 623 ==> Rings='(5.666667-10.333333]' 507 conf:(0.81)
15. ShuckedWeight='(0.1497-0.2984]' VisceraWeight='(0.07645-0.1524]' ShellWeight='(0.10185-0.2022]' 628 ==> Rings='(5.666667-10.333333]' 508 conf:(0.81)
16. Diameter='(0.3525-0.412]' ShellWeight='(0.10185-0.2022]' 611 ==> Rings='(5.666667-10.333333]' 493 conf:(0.81)
17. ShellWeight='(0.10185-0.2022]' 1023 ==> Rings='(5.666667-10.333333]' 824 conf:(0.81)

Faz sentido aplicar o algoritmo **Apriori** com a classe Rings já que o objectivo é obter um conjunto de valores que permitam estimar a idade dos moluscos.

Aplicando sem regras de associação à classe Rings, são obtidas 595 regras de associação. Vamos listar aqui algumas das associações com confiança superior a 0.99. São elas:

1. WholeWeight='(1.1314-1.41375]' ShellWeight='(0.30255-0.4029]' 428 ==> Height='(0.113-0.226]' 427 conf:(1)
2. Sex=I WholeWeight='(-inf-0.28435]' ShuckedWeight='(-inf-0.1497]' ShellWeight='(-inf-0.10185]' 479 ==> VisceraWeight='(-inf-0.07645]' 477 conf:(1)
3. ShellWeight='(0.30255-0.4029]' 798 ==> Height='(0.113-0.226]' 792 conf:(0.99)
4. WholeWeight='(1.1314-1.41375]' 616 ==> Height='(0.113-0.226]' 611 conf:(0.99)
5. Sex=F Diameter='(0.4715-0.531]' 433 ==> Height='(0.113-0.226]' 429 conf:(0.99)
6. ShellWeight='(0.2022-0.30255]' Rings='(5.666667-10.333333]' 679 ==> Height='(0.113-0.226]' 671 conf:(0.99)

Este conjunto de associações permite obter um conjunto de informações sobre os moluscos se imediata relação com a idade, como por exemplo, a informação baseada na 1ª regra:

**Um molusco com um peso entre 1.1314g e 1.41375g e um peso da carapaça entre 0.30255g e 0.4029g terá sempre um comprimento entre 0.113mm e 0.226mm.**

### Classificação