

**Universidade do Minho**

Departamento de 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](#_Toc437555921)

[1.1 Enquadramento e objetivos 4](#_Toc437555922)

[1.2 Planeamento 5](#_Toc437555923)

[1.3 Estrutura do documento 5](#_Toc437555924)

[2 Desenvolvimento 6](#_Toc437555925)

[2.1 Dataset de Avaliação de Automóveis 6](#_Toc437555926)

[2.2 Dataset do Consumo de Energia Elétrica 6](#_Toc437555927)

[2.3 Dataset da Central de Energia Elétrica 6](#_Toc437555928)

[3 Anexos 7](#_Toc437555929)

[3.1 Exemplo 7](#_Toc437555930)

# 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 Qualidade de Vinhos

Este dataset foi criado através de várias análises a vários vinhos verdes Portugueses neste caso de variante tinto. O dataset contém 1599 instâncias e 12 atributos, e contém maioritariamente vinhos médios (Quality 5 e 6) e poucos abaixo de 5 e acima de 7. Os atributos contidos no dataset são:

|  |  |
| --- | --- |
| Atributo | Gama de valores |
| Fixed acidity | [4.6 : 15.9] |
| Volatile acidity | [0.12 : 1.58] |
| Citric acid | [0 : 1] |
| Residual sugar | [0.9 : 15.5] |
| Chlorides | [0.012 : 0.611] |
| Free sulfur dioxide | [1 : 72] |
| Total sulfur dioxide | [6 : 289] |
| Density | [0.99 : 1.004] |
| pH | [2.74 : 4.01] |
| Sulphates | [0.33 : 2] |
| Alcohol | [8.4 : 14.9] |
| Quality | [3 : 8] |

Estes atributos não são balanceados, os 11 primeiros atributos foram obtidos pelas várias análises e o atributo Quality foi obtido pelos dados de um sensor.

Após uma análise ao dataset concluímos que a ideia seria relacionar os 11 atributos com o último e tentar descobrir quais as suas influências na qualidade do vinho.

Inicialmente todos os atributos do dataset estavam na forma Numeric, foi necessário colocar o atributo qualidade na forma {0,1,2,3,4,5,6,7,8,9,10} e discretizar o resto dos atributos para se poder procurar regras de associação. Após uma análise cuidada percebemos que os atributos não podem ser todos discretizados da mesma forma, foi preciso perceber de que forma cada um deles devia ser discretizado, quantidade de intervalos e frequências.

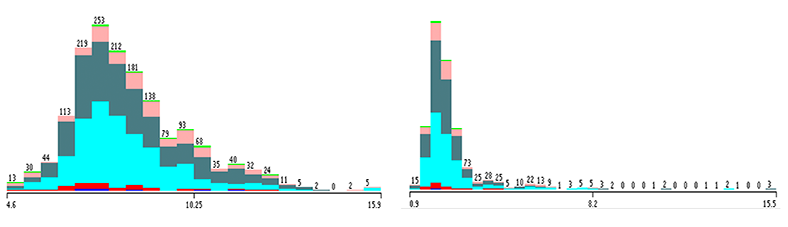


Ilustração . Fixed acidity à esquerda e Residual sugar à direita

Como podemos ver na figura anterior em que temos valores continuos é necessário adaptar a discretização ao atributo. Assim decidimos colocar números de intervalos adaptados a cada atributo, desde 8 intervalos para o pH até 35 atributos para o Chlorides, foi necessário números de intervalos altos porque estamos a trabalhar com medidas muitos precisas e cada valor mínimo pode fazer a diferença no output.

Para a associação utilizamos o algoritmo Apriori, alterando algumas definições, colocamos o car a True para relacionar com o atributo Quality e fizemos testes com diferentes valores nos campos lowerBoundMinSupport e minMetric.

Os testes iniciais, com grau de confiança mínimo de 0.6 e um suporte mínimo de 10% resultou em algumas regras, por exemplo:

sulphates='(-inf-0.53875]' alcohol='(9.05-9.7]' 135 ==> quality=5 113 conf:(0.84)

volatile acidity='(0.558-0.631]' alcohol='(9.05-9.7]' 115 ==> quality=5 88 conf:(0.77)

free sulfur dioxide='(15.2-22.3]' alcohol='(9.05-9.7]' 109 ==> quality=5 83 conf:(0.76)

fixed acidity='(7.425-7.99]' alcohol='(9.05-9.7]' 111 ==> quality=5 84 conf:(0.76)

chlorides='(0.0719-0.091867]' sulphates='(-inf-0.53875]' 165 ==> quality=5 122 conf:(0.74)

Após vários teste verificamos que só encontrávamos regras que implicavam uma qualidade de 5 no vinho, então decidimos baixar o suporte, pois como dizia as informações no dataset possui poucos abaixo de qualidade 5 ou acima de qualidade 7, então baixamos o valor do suporte para 0.05 e o grau de confiança para 0.5 para procurar regras que implicassem outras qualidades e encontramos:

sulphates='(0.719667-0.775333]' alcohol='(10.675-11]' 20 ==> quality=6 16 conf:(0.8)

total sulfur dioxide='(34.3-48.45]' sulphates='(0.775333-0.831]' 23 ==> quality=6 18 conf:(0.78)

volatile acidity='(0.412-0.509333]' free sulfur dioxide='(4.55-8.1]' pH='(3.248-3.375]' 27 ==> quality=6 21 conf:(0.78)

fixed acidity='(6.86-7.613333]' alcohol='(11.325-11.65]' 22 ==> quality=6 17 conf:(0.77)

residual sugar='(1.734286-2.151429]' chlorides='(0.063343-0.080457]' sulphates='(0.775333-0.831]' 21 ==> quality=6 16 conf:(0.76)

alcohol='(10.675-11]' 178 ==> quality=6 90 conf:(0.51)

E algumas regras com qualidade 7.

citric acid='(0.333333-0.4]' density='(0.994156-0.995518]' 28 ==> quality=7 17 conf:(0.61)

volatile acidity='(0.314667-0.412]' chlorides='(0.063343-0.080457]' density='(0.994156-0.995518]' 28 ==> quality=7 16 conf:(0.57)

volatile acidity='(0.217333-0.314667]' residual sugar='(2.151429-2.568571]' 36 ==> quality=7 19 conf:(0.53)

sulphates='(0.7475-0.95625]' 287 ==> quality=7 86 conf:(0.5)

Disto podemos retirar várias associações, por exemplo, a quantidade de álcool no vinho é um fator que influencia na qualidade de um vinho, um vinho com uma quantidade de álcool entre 9.05 e 9.7 será um vinho de qualidade 5 enquanto que uma quantidade entre 10.675 e 11.65 será um vinho de qualidade 6. Um vinho de qualidade 5 terá uma quantidade de sulfatos inferior a 0.53875 enquanto que um vinho de qualidade 6 terá entre 0.719667 a 0.775333 e um vinho de qualidade 7 terá entre 0.7475 e 0.95625. Um vinho com qualidade 5 ou 6 terá uma acidez volátil entre 0.412 e 0.509333, e um vinho com qualidade 7 terá entre 0.217333 e 0.412.

## Dataset do Consumo de Energia Elétrica

## Dataset da Central de Energia Elétrica

# Anexos

## Exemplo