Faculdade de Engenharia da Universidade do Porto

Deteção e caracterização de perturbações de preços de mercado com base em sistemas imunológicos artificiais

Eduardo Fernando Nogueira Rodrigues da Rocha

VERSÃO FINAL

Dissertação realizada no âmbito do
Mestrado Integrado em Engenharia Electrotécnica e de Computadores

Orientador: José Nuno Moura Marques Fidalgo
Co-orientador: João Tomé Saraiva

27 Janeiro de 2017

Resumo

A análise dos preços horários do mercado de eletricidade tem revelado situações em que a evolução se apresenta relativamente estável (ou, pelo menos, interpretável) e outras situações caracterizadas por grande irregularidade (e grande dificuldade de interpretar a evolução dos preços). No primeiro caso, os processos de previsão de preços funcionam satisfatoriamente, enquanto no último, a instabilidade dos preços penaliza fortemente o desempenho da previsão.

O objetivo desta dissertação é analisar e caracterizar as circunstâncias que condicionam a evolução dos preços, de modo a permitir a identificação de estados que potenciam a instabilidade dos preços. As conclusões deste estudo servirão posteriormente para apoiar sistemas de previsão dos preços de mercado.

Assim sendo, será analisado um ficheiro correspondente à evolução do preço horário do mercado de eletricidade em Portugal durante todo o ano de 2013. Uma primeira abordagem será feita utilizando um algoritmo básico de formação de classes (*clustering*), pretendendo-se atribuir classes ao maior número de casos. Numa segunda abordagem, será implementado um outro algoritmo de *clustering* baseado em Sistemas Imunológicos Artificiais, cujos protótipos (detetores) irão formar *clusters* diferentes. No fim, ambos os protótipos das diferentes abordagens irão ser testados para verificar se conseguiram melhorar a qualidade da previsão usando redes neuronais. O objetivo principal é confirmar que a separação em classes permite aumentar o desempenho da previsão.

Palavras-chave: Mercado de eletricidade, Sistemas imunológicos artificiais, *clustering*, redes neuronais, previsão de preços.

Abstract

The analysis of the hourly prices of the electricity market has revealed situations in which the evolution is relatively stable (or at least interpretable) and other situations characterized by great irregularity (and great difficulty in interpreting the evolution of prices). In the first case, the price forecasting processes work satisfactorily, while in the latter price instability strongly penalizes the forecast performance.

The objective of this dissertation is to analyze and characterize the circumstances that condition the evolution of prices, in order to allow the identification of states that promote price instability. The conclusions of this study will later serve to support market price forecasting systems.

Therefore, a file will be analyzed corresponding to the evolution of the hourly price of the electricity market in Portugal throughout the year 2013. A first approach will be made using a basic algorithm of class formation (clustering), aiming to assign classes to the largest number of cases. In a second approach, another clustering algorithm based on Artificial Immune Systems will be implemented, whose prototypes (detectors) will form different clusters. In the end, both prototypes of the different approaches will be tested to check if they have been able to improve the quality of the forecast using neural networks. The main goal is to confirm that class separation increases forecast performance.

***Key-words**: electricity market, artificial immune systems, clustering, neural networks, price forecast.*

Agradecimentos

Com esta dissertação termina o meu percurso académico, que sem a ajuda de algumas pessoas não teria sido possível. Serve este espaço para homenagear as que contribuíram para o meu sucesso na realização deste documento:

Em primeiro lugar, quero desejar um muito obrigado ao meu orientador, Prof. Dr. José Nuno Fidalgo, por todo o acompanhamento que me deu estes últimos meses, pela grande disponibilidade demonstrada, por todos os conselhos e críticas sempre com o objetivo de melhorar o meu trabalho.

Seguidamente, à minha família, especialmente aos meus Pais, que sempre se preocuparam e me apoiaram, garantindo todas as condições necessárias à minha concretização académica.

Por ultimo, a todos os meus amigos que não estão aqui mencionados, mas que sempre me deram motivação para acabar esta fase tão importante da minha vida.

Muito obrigado a todos!

"Everything in this world is magic, except to the magician"

Dr. Robert Ford (*Anthony Hopkins*) em *Westworld* (2016)

Índice

Lista de figuras

Lista de tabelas

Abreviaturas e Símbolos

Lista de abreviaturas (ordenadas por ordem alfabética)

FEUP	Faculdade de Engenharia da Universidade do Porto
MIEEC	Mestrado Integrado em Engenharia Eletrotécnica e de Computadores
MIBEL	Mercado Ibérico de Eletricidade
UE	União europeia
PE	Preço eletricidade
MAE	*Mean absolute error*
MAPE	*Mean absolute percentage error*
RMSE	*Root mean square error*
AIS	*Artificial Immune System*
RN	Rede neuronal
ER	Energias renováveis
ENR	Energias não renováveis
CB	*Clustering* básico

Lista de símbolos

≈	Varia de acordo com
!≈	Varia simetricamente com

Capítulo 1

Introdução

A presente dissertação foi realizada no âmbito do Mestrado Integrado em Engenharia Eletrotécnica e de Computadores (MIEEC), da Faculdade de Engenharia da Universidade do Porto (FEUP).

Neste capítulo será apresentada a temática da dissertação, de forma a enquadrar a sua importância no paradigma atual, assim como a motivação e os objetivos associados à sua elaboração.

No final do capítulo apresenta-se uma descrição breve da estrutura adotada para a dissertação. Os dados utilizados ao longo de todo o processo foram fornecidos pela EDP Distribuição e os programas utilizados para tratamento e análise de dados foram o Microsoft Excel e o MATLAB MathWorks.

1.1 - Enquadramento

O MIBEL (Mercado Ibérico de Eletricidade) resulta de um processo de cooperação entre os Governos de Portugal e Espanha com o intuito de criar um mercado de eletricidade comum. Esta iniciativa contribuiu significativamente não só para a realização do mercado de energia elétrica a nível ibérico, mas também à escala europeia, como um importante passo para a construção do mercado interno de energia da UE. Após um acordo entre os dois países, este mercado teve finalmente início no dia 1 de Julho de 2007 [1], oferecendo benefícios para os consumidores de ambos os países, promovendo o acesso a todos os interessados em condições de igualdade, transparência e objetividade.

O MIBEL é um mercado liberalizado onde os preços adquirem características voláteis e incertas, visto que os mesmos são obtidos através de propostas de oferta e de compra de energia, tal como nos outros sistemas bolsistas. Os agentes do mercado visam maximizar os seus lucros, objetivo esse que está extremamente dependente das suas estratégias de licitação. Sendo assim, é extremamente importante para estes que tenham à sua disposição a máxima quantidade de informação de qualidade sobre as condições futuras do mercado, de forma a poderem analisar os riscos de diferentes estratégias de licitação e definirem a que melhor serve

os seus interesses. Assim, num ambiente competitivo como este, torna-se fulcral prever o preço futuro da energia, não só para a definição de uma estratégia de venda e aumento do lucro por parte dos produtores de energia, mas também para beneficio financeiro de quem a compra e, duma forma geral, de todos os agentes que a transacionam.

1.2 - Motivação

O mercado de energia, por ser um ambiente muito competitivo, faz com que a previsão de preços de eletricidade assuma uma elevada importância. As aplicações da previsão de preços de eletricidade variam consoante o horizonte temporal desejado, podendo ser efetuadas previsões a curto, médio e longo prazo. A capacidade de saber como os preços irão tender no futuro é extremamente vantajosa para todos os participantes e agentes do mercado. As suas estratégias e lucros dependem de uma aposta correta de que o mercado irá tender para uma direção ou outra.

O estudo dos preços horários do mercado de eletricidade tem revelado casos em que a evolução se apresenta relativamente previsível e outras situações caracterizadas por grande imprevisibilidade. Para o primeiro caso, os processos de previsão do PE são capazes de obter resultados satisfatórios, enquanto que no último a instabilidade dos preços prejudica bastante a *performance* da previsão. É então de grande interesse e utilidade desenvolver uma técnica que tenha a capacidade de ultrapassar este último obstáculo, sendo capaz de efetuar a previsão do PE de forma mais eficiente.

1.3 - Objetivos

Pretende-se nesta dissertação analisar e caracterizar as circunstâncias que condicionam a evolução dos preços, de modo a permitir a identificação de estados que potenciam a instabilidade dos preços. O objetivo final é verificar se a identificação de classes e a compartimentação dos dados permite melhorar o desempenho da previsão.

Será analisado um ficheiro com dados de produção, consumo e preço do mercado de energia, referentes a Portugal em cada hora do ano de 2013.

Sendo assim, os objetivos estão divididos em 3 fases:

- Caracterização e classificação básica dos casos horários contidos no ficheiro de dados, em que cada hora do ano está caracterizada por várias variáveis: consumo de eletricidade, produção e consumo de certas energias renováveis e não renováveis, importação, exportação, bombagem e preço;
- Criação de um programa baseado em sistemas imunológicos artificiais (algoritmos de seleção negativa) que irá criar "casos-tipo" (detetores) com um dado preço associado, que possam ser identificados como muito semelhantes aos casos fornecidos;
- Utilização das classificações de casos identificadas nos pontos anteriores para melhorar a previsão de preços usando redes neuronais (RN). Isto será feito através da consideração de previsores especializados para cada tipo de caso.

Em primeiro lugar, irá ser utilizado o Microsoft Excel para classificar casos mais fáceis de explicar, passando-se depois para o desenvolvimento do programa baseado em sistemas imunológicos artificiais usando o software MATLAB MathWorks. Para verificar se a previsão está a melhorar, será utilizada a ferramenta de RN existente no MATLAB.

1.4 - Estrutura da dissertação

Para além deste capítulo introdutório, esta dissertação contém outros 4 capítulos: Revisão bibliográfica, Metodologia, Resultados e Conclusão.

O capítulo 2 é inteiramente dedicado ao estado da arte, onde são apresentadas algumas formas de efetuar a previsão do preço da eletricidade, assim como as teorias em que se baseiam os algoritmos postos em prática nesta dissertação.

No capítulo 3 é feita a caracterização e descrição da metodologia utilizada na resolução do problema proposto nesta dissertação, onde se descrevem as ferramentas de implementação dos algoritmos utilizados.

O capítulo 4 contém uma análise dos resultados obtidos para cada conjunto de dados e no capítulo 5 é feita uma consolidação das conclusões obtidas no decurso da dissertação.

Capítulo 2

Revisão bibliográfica

Existem muitas formas de efetuar a previsão do preço da eletricidade (PE). Porém, neste capítulo irão apenas ser apresentadas de forma sucinta algumas alternativas mais comuns, assim como serão apresentados e descritos os sistemas e técnicas que servirão de base ao modelo proposto para esta dissertação.

Em geral, as previsões do PE podem ser obtidas através de métodos deterministas ou probabilísticos, havendo várias técnicas distintas para cada caso [2]. Contudo, nesta dissertação serão abordados outros processos com vista a prever os PE, recorrendo a dois algoritmos: um algoritmo de *clustering* básico (CB) e um algoritmo imunológico artificial (AIS – *Artificial immune system*). Para validar os resultados do objetivo desta tese irá ser usada a ferramenta de redes neuronais para testes de previsão de preços.

2.1 - Previsão Determinista do PE

2.1.1 – Horizonte temporal

O período temporal compreendido entre o momento em que a previsão é efetuada e o instante para o qual se quer determinar uma dada variável é designado por horizonte temporal.

Dependendo do horizonte temporal em questão, pode classificar-se a previsão como sendo de curto, médio ou longo prazo [2].

O que separa as previsões destes três tipos de horizontes temporais são essencialmente o tipo de técnicas e variáveis de influência que são utilizadas no seu desenvolvimento, as suas aplicações e avaliações de desempenho (um determinado erro da previsão pode ser aceitável para uma previsão de longo prazo e não ser aceitável para uma de curto prazo). De acordo com [3], estes tipos de previsões definem-se como:

- Previsões de PE de longo prazo: dizem respeito a horizontes de previsão mensais, trimestrais ou anuais e são usadas para análises de aproveitamento de investimentos, como é o caso da determinação de novas localizações ou fontes de combustível para as centrais [4].

- Previsões de PE de médio prazo: inserem-se em previsão do PE no intervalo temporal que compreende alguns dias até alguns meses. São utilizadas em cálculos de balanços, em planeamentos de expansão da geração e de manutenção, na realocação de recursos e na realização de contratos bilaterais [5] [6].
- Previsões de PE de curto prazo, as mais comuns, compreendem de alguns minutos até alguns dias. São de extrema importância nas operações diárias dos mercados, por exemplo, na preparação de propostas de compra e oferta, na implementação de outras funções de operação de sistemas de potência e em funções de operação de sistemas energéticos [7] [6]. Uma central que seja capaz de prever o preço a curto prazo da eletricidade pode ajustar o planeamento da sua produção de forma a maximizar o seu lucro [4] [8].

O exemplo de teste estudado nesta dissertação aborda apenas a previsão a curto-prazo.

2.1.2 – Seleção de variáveis

Para que a previsão do PE seja a mais fiável possível é necessária uma boa seleção das variáveis que serão utilizadas como entradas no modelo de previsão [2] [9].

Uma vez que estas seleções são frequentemente baseadas em pressupostos por parte de quem está a construir o modelo de previsão, para tentar evitar lapsos [10], é essencial o pré-processamento, a seleção e a forma como serão organizadas as variáveis que vão ser utilizadas [11].

Nesta dissertação serão utilizados diferentes tipos de variáveis, sendo estas classificadas como: variáveis cronológicas, variáveis de preço e variáveis de produção/consumo.

As variáveis cronológicas ajudam a caracterizar o comportamento do preço de mercado em termos de sazonalidade diária, semanal e mesmo em termos anuais. As condições atmosféricas apresentam uma grande relação com o preço de mercado que se deve essencialmente à variabilidade da carga e das condições de geração, nomeadamente de energias renováveis [2].

Dada a sua relação com o consumo, também é usual a consideração de variáveis referentes à hora, dia, mês, ano, feriados e estação do ano, sendo estas variáveis de livre acesso a todos os agentes do mercado.

As variáveis de preço representam valores históricos de preços de mercado (que podem ir de dias até anos atrás) e/ou presentes, e existem dada a relação entre os valores futuros do preço com os valores conhecidos previamente desta variável. Estas são normalmente facultadas pela maioria dos operadores do mercado, porém a obtenção de valores correspondentes a preços de períodos muito longínquos no passado pode revelar-se bastante difícil.

Por ultimo, as variáveis de produção elétrica têm o objetivo de representar alguns dos fatores externos que tem influência no mercado dos preços da eletricidade, podendo ser variáveis históricas ou presentes. Exemplos de variáveis que pertencem a esta classe são a geração de energia solar, térmica, eólica, hídrica e o consumo de energia, todas relevantes no modelo usado nesta dissertação. Para as previsões que irão ser feitas ao longo desta dissertação, em relação ao preço da eletricidade no MIBEL, os valores destas variáveis são diariamente fornecidas pela EDP Distribuição.

2.1.3 - Avaliação de Desempenho

Quanto às medidas de caracterização da qualidade da previsão não há uma consensualidade no que diz respeito à sua literatura, sendo necessário um certo cuidado na sua avaliação [2].

Irão ser abordadas as medidas de precisão utilizadas na presente dissertação, sendo predominantes os erros absolutos e os erros quadráticos.

Nas equações seguintes a variável P_h representa o preço real e $P_h{}'$ o preço previsto para o período de carga h. Obtido o cálculo destas variáveis é possível calcular o MAE *(mean absolute error), o MAPE (mean absolute percentage error) e o RMSE (root mean square error).*

$$\boldsymbol{MAE_T} = \frac{1}{T}\sum_{h=1}^{T}\left[\left|P_h - P_h^{'}\right|\right]$$

$$\boldsymbol{MAPE_T} = \frac{1}{T}\sum_{h=1}^{T}\left[\frac{\left|P_h - P_h^{'}\right|}{P_h}\right]$$

$$\boldsymbol{MAPE'_T} = \frac{\boldsymbol{MAE_T}}{< P_h >}$$

$$\boldsymbol{RMSE_T} = \sqrt{\frac{1}{T}\sum_{h=1}^{T}\left[\left(P_h - P_h^{'}\right)^2\right]}$$

- Tal como o nome sugere, o MAE é uma média dos erros absolutos. Esta medida usa a mesma escala dos dados que estão a ser medidos e é conhecida como uma medida de precisão dependente da escala, sendo que por isso não pode ser usada para fazer comparações entre séries de escalas diferentes [12]. O MAE é uma medida comum do erro de previsão na análise de séries temporais [13].
- O MAPE é uma medida da precisão de métodos de previsão estatísticos e expressa o resultado em percentagem. Apesar do conceito do MAPE parecer convincente e simples, tem algumas desvantagens [14], como não poder ser usado em dados cujo valor é nulo, uma vez que iria dar origem a uma divisão por 0. Como existem valores nulos no conjunto de dados a tratar, neste caso usa-se o MAPE' que consiste em dividir o erro absoluto médio (MAE) pelo valor médio da grandeza ($<P_h>$).
- O RMSE representa o desvio padrão da amostra das diferenças entre os valores previstos e os valores observados. Essas diferenças individuais são chamadas resíduos quando os cálculos são executados sobre a amostra de dados que foi usada para estimativa e são chamados erros de previsão quando computados fora da amostra. Esta é uma medida muito usada na caracterização da qualidade da previsão, até porque se relaciona diretamente com o critério de minimização do erro mais utilizado (mínimos quadrados). No entanto, não será o mais fácil de interpretar, ruma vez que é dependente da escala da grandeza a prever [13].

2.2 – Previsão Probabilística do PE

Enquanto as previsões deterministas consistem na determinação de um valor determinístico para o valor futuro da variável que se pretende prever, as previsões probabilísticas assumem um conjunto de resultados possíveis para esse futuro. Para certas aplicações, tais como análise de risco ou na licitação com margens de segurança, esta classe de previsões é mais adequada do que as deterministas, já que conseguem quantificar a incerteza dos valores previstos [15] [16]. Contudo, embora a sua utilização esteja a ser cada vez mais usual em diversas áreas, nas previsões do PE a sua aplicação ainda não foi muito explorada [2]. Existem vários métodos de representação das incertezas, nomeadamente as funções densidade de probabilidade, funções de distribuição cumulativa e intervalos de previsão. No entanto, não iremos abordar nenhuma delas nesta tese de dissertação, uma vez que não se inserem na metodologia utilizada para a resolução do problema.

2.3 – Agrupamento (*Clustering)*

Clustering é uma técnica de *Data Mining* usada para construir agrupamentos automáticos de dados segundo o seu grau de semelhança. O critério de semelhança faz parte da definição do problema a resolver e depende do algoritmo a implementar.

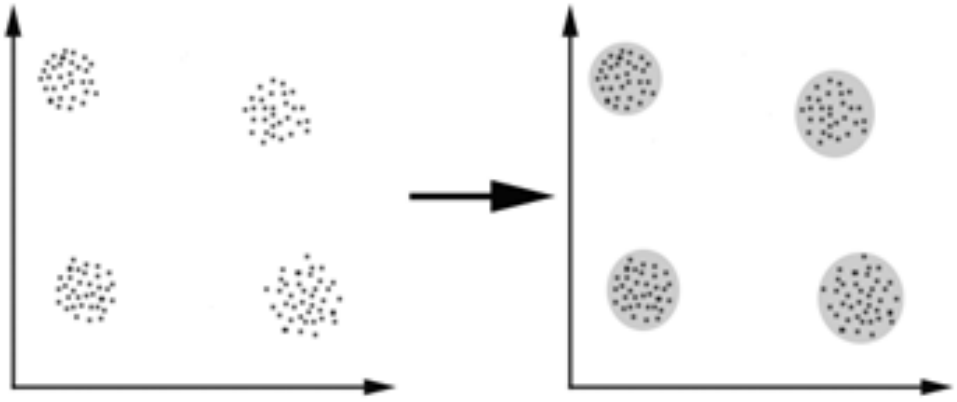


Figura 1 – *Clustering* baseado em distância [17].

Neste caso, pode-se ver claramente os 4 *clusters* nos quais os dados podem ser divididos. O critério de similaridade mais comum é uma medida direta ou indireta de distância: dois ou mais objetos pertencem ao mesmo *cluster* se estiverem "próximos" de acordo com uma distância dada (neste caso distância euclidiana). Cada objeto é caracterizado por um conjunto de parâmetros de modo que o algoritmo de *clustering* vê cada objeto como um vetor com N

variáveis de entrada. Algumas técnicas de *clustering* transformam este espaço N-dimensional num espaço bidimensional no qual são calculadas as tais distâncias.

Quando se tem um conjunto de objetos é natural olhar para eles tentando perceber o quão semelhantes ou diferentes eles são uns dos outros. Uma abordagem comum consiste em definir uma função de distância entre os objetos, com a interpretação de que objetos com uma distância menor são mais semelhantes. O *clustering* (agrupamento) surge quando tentamos classificar ou organizar objetos em grupos coerentes. Dada uma função de distância, o algoritmo divide os objetos em grupos para que, intuitivamente, objetos dentro do mesmo grupo estejam próximos e objetos de diferentes grupos estejam distantes [18].

Portanto, o objetivo do *clustering* é determinar o agrupamento intrínseco num conjunto de dados não marcados (aprendizagem não supervisionada). No entanto, não há nenhum critério absoluto "melhor" independente do objetivo final do agrupamento. Consequentemente, é o utilizador que deve definir este critério, de tal forma que o resultado do agrupamento satisfará as suas necessidades [19].

Algoritmos de *clustering* podem ser aplicados em muitas áreas, por exemplo: *marketing*, biologia, seguros, urbanismo, classificação de documentos, entre outros [17] [18].

Os principais requisitos que um algoritmo de agrupamento deve satisfazer são [17]:

- Escalabilidade;
- Conseguir lidar com diferentes tipos de atributos;
- Descobrir *clusters* de forma arbitrária;
- Capacidade de lidar com o ruído e *outliers*;
- Insensibilidade à ordem dos registos de entrada;
- Alta dimensionalidade;
- Interpretação e usabilidade.

Alguns exemplos de algoritmos de *clustering* podem ser classificados como [17]:

- *Clustering* exclusivo - os dados são agrupados de forma exclusiva, de modo a que se um certo conjunto de dados pertence a um *cluster*, então ele não pode ser incluído noutro *cluster*;
- *Clustering* de sobreposição - utiliza conjuntos *fuzzy* para agrupar dados, de modo que cada ponto pode pertencer a dois ou mais *clusters* com diferentes graus de adesão. Nesse caso, os dados serão associados a um valor de associação apropriado;
- *Clustering* hierárquico - baseia-se na ideia de recursividade. A condição inicial é obtida definindo todos os pontos de referência, obtendo-se um conjunto primário de *clusters*. Em seguida, aplica-se o mesmo algoritmo a cada *cluster* separadamente. Depois de algumas iterações, os *clusters* finais desejados são alcançados.
- *Clustering* probabilístico - usa uma abordagem probabilística, sendo os resultados apresentados em termos de probabilidade de pertença a um dado *cluster*.

2.3.1 Medida de similaridade

A similaridade entre dois pares de objetos pode ser medida de diversas maneiras, sendo que a mais utilizada para medir dados é a distância euclidiana [20]. Uma vertente importante de um algoritmo de *clustering* é a sua distância, medida entre os pontos de dados. Se os componentes dos vetores de dados estão todos nas mesmas unidades físicas, então é possível que a medida de distância euclidiana simples seja suficiente para agrupar com sucesso os grupos de dados semelhantes. No entanto, mesmo neste caso ela pode por vezes ser enganadora. A figura abaixo ilustra isso com um exemplo das medidas de largura e altura de um objeto. Apesar de ambas as medições serem consideradas nas mesmas unidades físicas, uma decisão informada tem que ser feita quanto à escala relativa. Como mostra a figura, diferentes escalonamentos podem levar a diferentes agrupamentos [17].

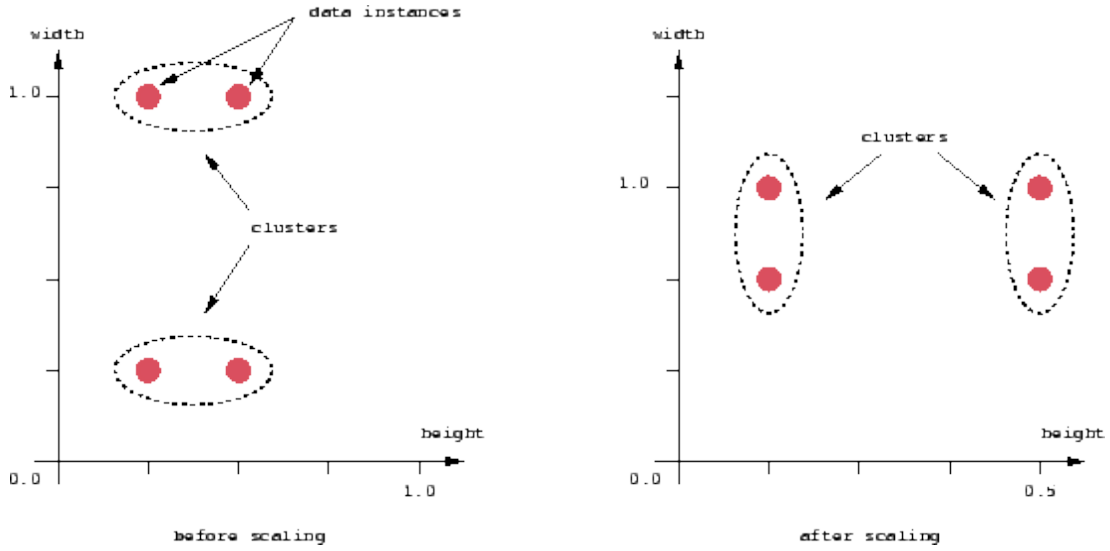


Figura 2 – Diferenças nos resultados na aplicação de diferentes algoritmos para o mesmo problema [17].

Tal como é observado, não existe um algoritmo único que seja objetivamente o melhor. O algoritmo de *clustering* mais apropriado para um problema específico precisa de ser escolhido experimentalmente, a não ser que haja uma razão matemática para preferir um modelo de *clustering* sobre outro. Um algoritmo que é projetado para um tipo de modelo não é útil num conjunto de dados que contém um tipo de modelo radicalmente diferente. O conhecimento do domínio deve ser usado para orientar a formulação de uma medida de distância adequada para cada aplicação particular [19].

Podem-se ver aplicações de *clustering* para previsão do PE em [21] e [22]. No caso desta dissertação, será utilizada a distância euclidiana para agrupar casos mais semelhantes e estes apenas poderão pertencer a um *cluster* (detetor). Por isso, pode-se afirmar que a classificação irá ser baseada num algoritmo de *clustering* exclusivo baseado em distância.

2.4 – Sistemas imunológicos artificiais

Sistemas Imunológicos Artificiais ou Artificial Immune Systems (AIS, em Inglês) são sistemas da área da Inteligência Computacional, baseados em regras inspiradas nos princípios e processos do sistema imunológico do ser humano. A principal ideia destes sistemas é transpor a função do sistema imunológico biológico para sistemas computacionais e investigar a sua aplicação para resolver problemas computacionais de matemática, engenharia e tecnologia da informação [23].

A função primária do sistema imunológico é proteger os corpos humanos de agentes infeciosos (como vírus, bactérias e outros parasitas) vulgarmente conhecidos como patógenos. A resposta imune é estimulada pelo reconhecimento de uma molécula associada chamada antígeno. O sistema imunológico geralmente funciona de acordo com dois mecanismos: imunidade inata e adaptativa. A imunidade inata é dirigida contra patógenos gerais que entram no corpo enquanto a imunidade adaptativa permite lançar um ataque contra qualquer invasor que o sistema inato não consegue remover [24]. O algoritmo implementado neste estudo é inspirado neste conceito, embora não seja comparável em todos os aspetos.

Existem 4 tipos distintos de algoritmo destes sistemas [25]:

- *Clonal Selection Algorithm;*
- *Negative Selection Algorithm;*
- *Immune Network Algorithms;*
- *Dendritic Cell Algorithms.*

No entanto, como na presente dissertação vai ser usado um *Negative Selection Algorithm* (algoritmo de seleção negativa), apenas este será descrito.

A seleção negativa refere-se à identificação e supressão de células auto-reativas, isto é, células T, que podem selecionar e atacar tecidos próprios. Esta classe de algoritmos é tipicamente usada para classificação e reconhecimento de padrões onde o espaço do problema é modelado no complemento do conhecimento disponível. Por exemplo, no caso de um domínio de deteção de anomalias, o algoritmo prepara um conjunto de detetores de padrões exemplares, treinados em padrões normais, que modelam e detetam padrões não vistos ou anómalos [26].

Assim sendo, o ponto de partida deste algoritmo é identificar um conjunto de casos, S, que definem o estado normal do sistema. A tarefa seguinte é gerar um conjunto de detetores, T, que apenas ligam / reconhecem o complemento de S. Estes detetores podem então ser aplicados a novos dados para classificá-los como sendo parte do conjunto normal ou não [27]. Pode-se ver uma aplicação deste tipo de algoritmo em [28].

```
input   : S_seen = set of seen known self elements
output  : D = set of generated detectors

begin

   repeat
      Randomly generate potential detectors and place them in a set P
      Determine the affinity of each member of P with each member of the self set S_seen
      If at least one element in S recognises a detector in P according to a recognition threshold,
         then the detector is rejected, otherwise it is added to the set of avaliable detectors D
   until  Stopping criteria has been met

end
```

Figura 3 – Pseudo-código para gerar e aplicar detetores [27].

No caso desta dissertação, estes detetores T serão padrões horários fictícios criados pelo algoritmo implementado (MATLAB), que irão ser comparados com todos os outros casos contidos no ficheiro a ser analisado, usando como medida a distância euclidiana. O Capítulo 3 descreve este procedimento em detalhe.

2.5 – *Redes neuronais (RN)*

De acordo com [29], as redes neuronais artificiais (RN) são inspiradas nos modelos do sistema nervoso central humano, mas de forma mais simplificada. Esta estrutura composta por unidades computacionais conectadas entre si são chamadas de neurónios e têm capacidade de aprendizagem.

Esta capacidade de aprendizagem é dada pelas interações com um ambiente, cujas informações obtidas são armazenadas em conexões (sinapses) entre neurónios. Estes neurónios estão organizados num grupo pequeno de camadas que se situam entre as entradas (inputs) e as saídas (outputs) da rede [2].

Às ligações entre neurónios (sinapses) está associado um determinado peso, que é ajustado durante o processo de aprendizagem até que se atinja um certo objetivo, normalmente relacionado com a minimização dum erro de aproximação [29].

Sendo assim, a operação de qualquer RN pressupõe um processo de treino, onde são dadas como *inputs* as variáveis de influência, e como *targets* os *outputs* ou variáveis que se querem prever. Este procedimento marca o inicio da aprendizagem onde a rede neuronal vai determinar quais as melhores relações entre as entradas e as saídas, de forma que o erro médio quadrático entre os outputs produzidos pela rede e os *outputs* reais (*target*) seja mínimo [30].

Dando por terminada a fase de treino inicia-se a fase de teste da rede neuronal, onde são aplicados os pesos a um conjunto de *inputs* que não foram previamente fornecidos à rede, mas que possuam as mesmas características (estrutura de dados) utilizados na fase de treino. É então calculado um conjunto de *outputs*, de forma a obter-se os valores da variável que se pretendia prever inicialmente.

Finalmente, procede-se à avaliação de desempenho da rede neuronal através das medidas de avaliação apresentadas em 2.1.3.

As RN são classificadas consoante a sua arquitetura e algoritmo de aprendizagem, sendo que a primeira diz respeito à maneira como se organizam as ligações entre neurónios e a

segunda diz respeito à maneira como a rede neuronal ajusta os pesos das suas conexões [2]. Quanto à arquitetura, existem dois grupos de RN: as RN *feedforward* e as RN recorrentes. Neste trabalho são apenas utilizadas redes *feedforward* e estas já são bastante conhecidas no meio académico, pelo que não se vai apresentar uma descrição dos seus princípios base neste trabalho. Quanto aos algoritmos de treino, a literatura mostra que existem várias alternativas, sendo o algoritmo de *Levemberg-Marquatt* um dos mais eficientes [31].

As RN são uma das técnicas mais utilizadas no ramo da previsão do PE de curto-prazo. Nos artigos [32-39] é possível observar algumas destas abordagens para diferentes cenários.

Capítulo 3

Metodologia

A metodologia adotada neste estudo pressupões dois grandes passos: identificação de *clusters* e interpretação das principais classes identificadas. A fase final consiste em testes de previsão de preços usando, por um lado, todo o conjunto de dados e, por outro lado, os dados separados pelos diversos grupos (*clusters*). Em princípio, se a previsão no segundo caso (por *cluster*) resultar melhor que a previsão global com todos os dados, então poderemos concluir que os clusters identificaram comportamentos próprios em termos de preços de mercado.

Em termos de agrupamento de dados irão ser realizadas duas abordagens distintas: uma primeira por separação básica dos dados, baseada exclusivamente na comparação de amplitudes entre pares de variáveis do vetor de entrada; uma segunda técnica foi baseada na aplicação de AIS:

- Na separação básica, todas as variáveis consideradas relevantes foram divididas em 5 níveis (de "baixo" a "alto") e efetuou-se uma comparação dos níveis das variáveis "consumo", "ENR", "ER", "importação" e "exportação" com o "preço", para cada respetiva hora do ano. De cada teste resulta um número que representa uma classificação distinta (1 para o 1º teste, 5 para o último teste). Depois de atribuídas as classes, é possível visualizar o protótipo das mesmas;
- Na aplicação baseada em AIS, serão criados detetores aleatórios (vetores com o mesmo tamanho dos dados fornecidos), que possam identificar estados muito semelhantes a eles. Quando um detetor atinge um número mínimo de emparelhamentos, é criada uma classe;
- No fim de cada algoritmo, serão criadas RN para todas as classes, tendo em conta também a variável "preço 24h antes", sendo que as primeiras 24h do ano não são *inputs* da RN por não possuírem esta variável. As RNs têm como entradas as variáveis todas exceto o "preço", sendo que este é o seu *target* desejado. Os resultados da previsão da RN com classes são então comparados aos resultados da RN inicial (sem classes).

A atribuição de classes no Excel e posteriormente em MATLAB foi feita de acordo com o seguinte fluxograma:

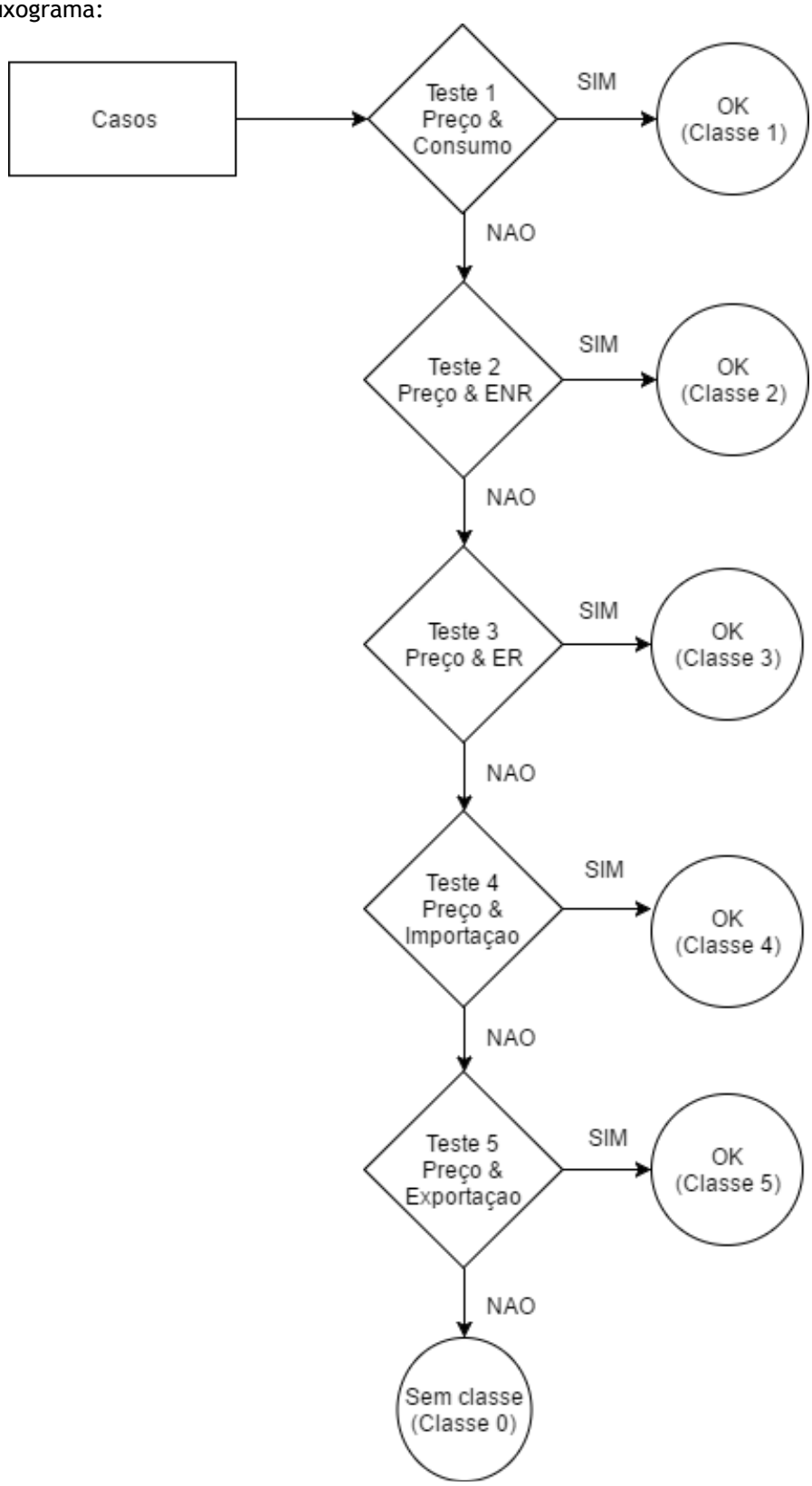


Figura 4 – Fluxograma do algoritmo desenvolvido em Excel.

3.1 Análise dos dados disponibilizados

Como foi referido anteriormente, irá ser analisado um ficheiro correspondente ao preço e consumo horários da eletricidade em Portugal, em todo o ano de 2013 (tirando os fins-de-semana). As variáveis contidas neste ficheiro que serão utilizadas na dissertação são as seguintes:

- Data (dia, mês, ano);
- Dia da semana (dds);
- Hora do dia;
- Consumo no dia;
- Carvão;
- Fuel;
- Gás natural;
- Albufeiras;
- Fios de água;
- Importação;
- Exportação;
- PRE hidráulica;
- PRE térmica;
- PRE fotovoltaica;
- PRE ondas;
- Bombagem;
- Preço de mercado.

Data	Mês	dia	dds	Hora	Consur	Preço	Carvã	Fuel	Gás Natu	Albufeir
01/01/2013	1	1	3	00:00	4905,7	32,2	920,9	0	0	646,85
01/01/2013	1	1	3	01:00	4710,3	32,1	920,4	0	0	373,55
01/01/2013	1	1	3	02:00	4414,9	31,4	920,2	0	0	125,4

Tabela 1 – Extrato do ficheiro a ser analisado.

3.2 – Caracterização e classificação de casos (CB)

3.2.1 – 1ª Abordagem (valores nominais)

Tal como já foi referido anteriormente, em primeiro lugar irá ser desenvolvido em Excel um algoritmo de classificação de classes hierárquico para verificar quantos casos é possível agrupar e quantos não são classificados. Cada variável irá ser dividida de forma homogénea em 5 níveis: "BAIXO", "MÉDIO BAIXO", "MÉDIO", "MÉDIO ALTO" e "ALTO". Por exemplo, no nível BAIXO encontram-se todos os casos entre 0 e 20%, relativamente ao valor máximo da variável; o nível MÈDIO BAIXO é dos 20% aos 40%, e assim sucessivamente.

O 1º teste engloba apenas as variáveis "consumo" e "preço". Segundo a lógica conhecida, se o Consumo é baixo, o Preço também deverá ser baixo; se o Consumo é elevado, então o Preço também deverá ser elevado. Sendo assim, após a realização deste teste teremos a classe em que a categoria do "consumo" será a mesma do "preço" e será caracterizada como sendo a classe "1". Os casos que não cumprirem esta condição serão numerados com "0", indicando assim que não estão inseridos em nenhuma classe. Sendo assim:

- 1 – ambos com o mesmo nível nas categorias "consumo" e "preço";
- 0 – nenhuma das condições foi satisfeita, logo o caso continua por explicar.

Data	Hora	Consumo	Preço	Categoria Consumo	Categoria Preço	Classe 1
1/01/2013	00:00:00	4905,725	48	MÉDIO BAIXO	MÉDIO	0
1/01/2013	01:00:00	4710,3	45	MÉDIO BAIXO	MÉDIO	0
1/01/2013	02:00:00	4414,875	31,27	MÉDIO BAIXO	MÉDIO BAIXO	1
1/01/2013	03:00:00	4036,9	21	BAIXO	BAIXO	1

Tabela 2 – Excerto da folha Excel para tratamento de dados (Classe 1).

Este 1º teste foi capaz de explicar 2441 de 6264 casos, ou seja, cerca de 38,97% de todos os casos. A Figura 5 mostra a divisão de casos: a coluna à esquerda inclui os casos que pertencem à Classe 1 e a da direita os que não pertencem.

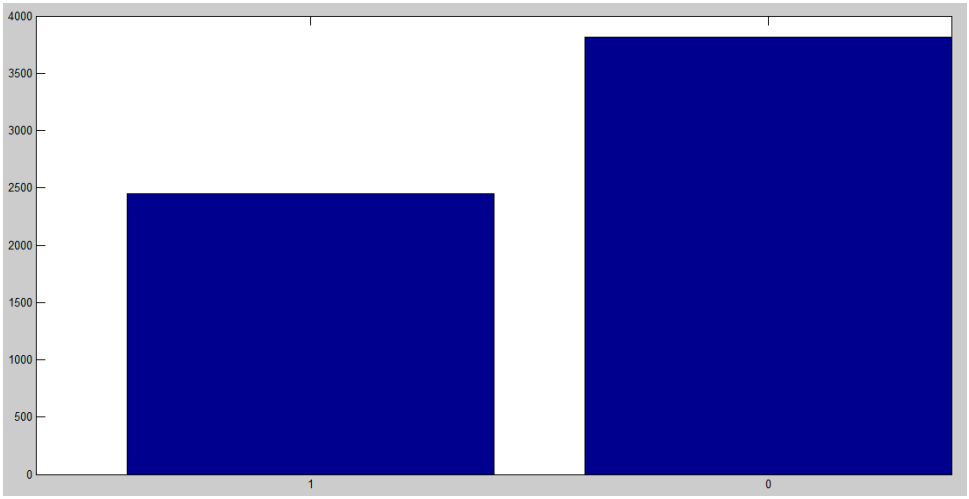


Figura 5 – Resultados do 1º teste.

Depois de um caso ser atribuído a uma classe, já não irá participar nos restantes testes, sendo substituído por um "V" no resultado final do teste realizado.

O 2º e 3º testes foram realizados comparando a categoria "energias não renováveis" (ENR) e "energias renováveis" (ER) com o "preço", respetivamente. Como tal, as ENR englobam o "carvão", "fuel" e "gás natural", enquanto que as ER englobam as "albufeiras", "fios de água" e a "PRE total" de todas as fontes de energia renovável incluídas no ficheiro.

No entanto, sendo que no 2º teste as classes são atribuídas da mesma forma do que no 1º (mesmo nível=classe), no 3º teste são atribuídas de forma diferente, uma vez que, geralmente, quanto mais ER foram produzidas, menor será o "preço": a ER e o Preço variam de forma simétrica – usa-se aqui o símbolo "!≈" para representar este tipo de relação. Então, as classes no 2º e 3º testes foram atribuídas da seguinte maneira:

- 2 – ambos com o mesmo nível de "preço" e "ENR";
- 3 – ER "ALTO" e preço "BAIXO", ER "MÉDIO ALTO" e preço "MÉDIO BAIXO", idem até ER "BAIXO" e preço "ALTO";
- 0 - nenhuma das condições foi satisfeita.

Categoria energias não renovaveis	Categoria energias renovaveis	Classe 2	Classe 3
MÉDIO BAIXO	MÉDIO BAIXO	0	0
MÉDIO BAIXO	MÉDIO BAIXO	0	0
MÉDIO BAIXO	MÉDIO BAIXO	V	V
MÉDIO BAIXO	MÉDIO BAIXO	V	V

Tabela 3 – Excerto de dados e resultados classe 2 e 3.

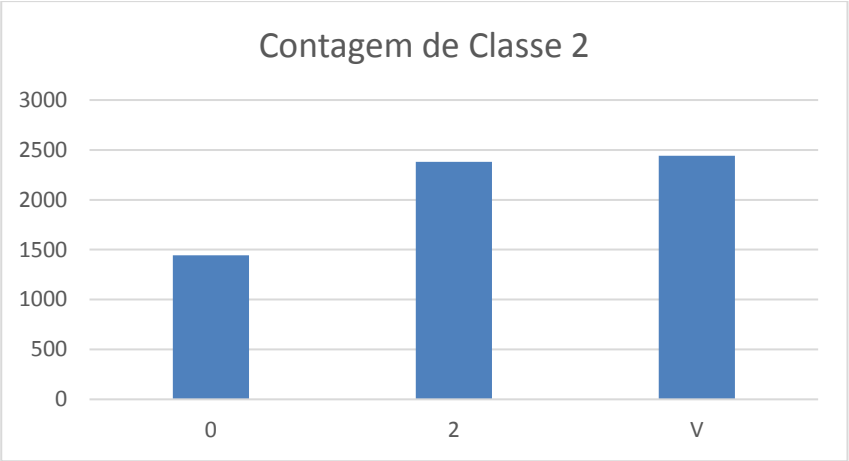


Figura 6 – Gráfico de resultados do teste da classe 2.

Classe 2	Contagem de Classe 2
0	1443
2	2380
V	2441

Tabela 4 – Resultados da classe 2.

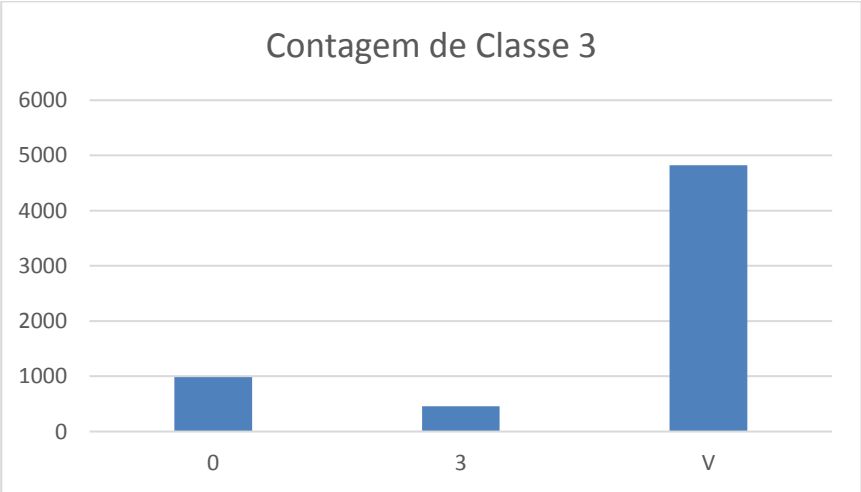


Figura 7 – Gráfico de resultados do teste da classe 3.

Classe 3	Contagem de Classe 3
0	986
3	457
V	4821

Tabela 5 – Resultados da classe 3.

O 4º e 5º testes foram realizados comparando as variáveis “importação” e “exportação” com o “preço”, respetivamente. No 4º teste a lógica de atribuição de classes foi feita como no 1º, enquanto que no 5º foi feita como no 3º.

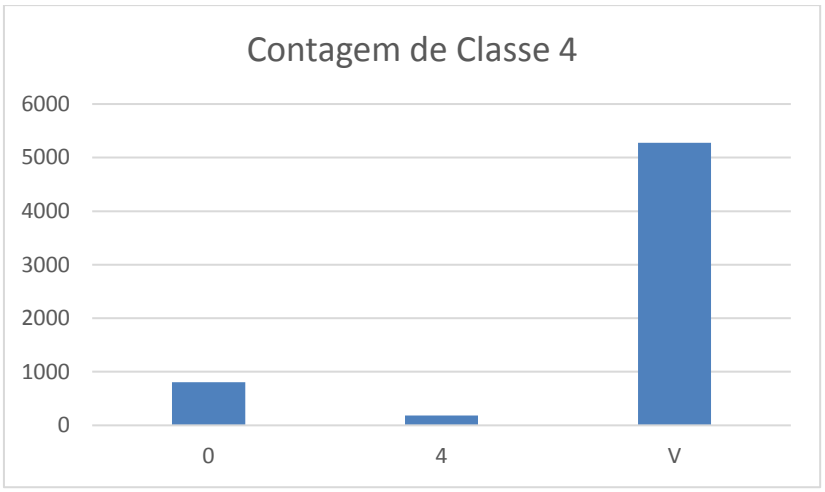


Figura 8 – Gráfico de resultados do teste da classe 4.

Classe 4	Contagem de Classe 4
0	805
4	181
V	5278

Tabela 6 – Resultados da classe 4.

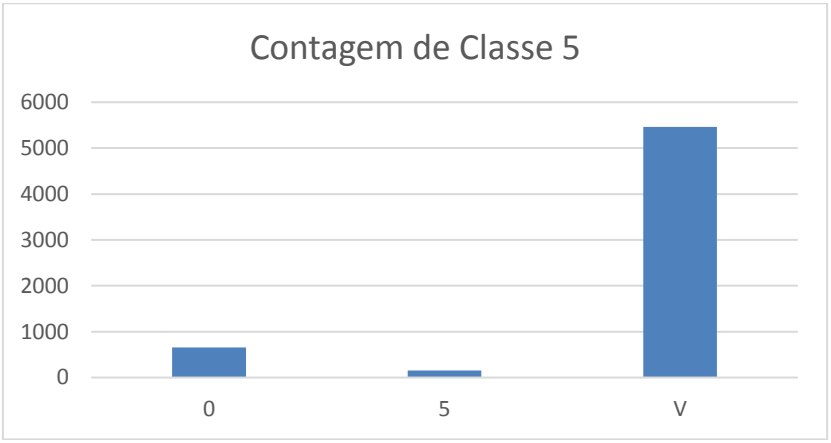


Figura 9 – Gráfico de resultados do teste da classe 5.

Classe 5	Contagem de Classe 5
0	654
5	151
V	5459

Tabela 7 – Resultados da classe 5.

No final, ficou-se apenas com 654 casos sem classe, o que representa apenas 9,96% da totalidade dos casos.

3.2.2 – 2ª Abordagem (normalizações e percentagens)

Em seguida, procedeu-se a uma abordagem diferente, no que toca aos *inputs* dos testes dos quais resulta uma classe: agora, após o 1º teste, os seguintes serão realizados tendo em conta a percentagem das mesmas variáveis em função do respetivo consumo (as entradas foram normalizadas em relação ao consumo da hora).

energias renovaveis/consumo	energias não renovaveis/consumo	categoria energias renovaveis/consumo	categoria energias não renovaveis/consumo	Teste 2	Teste 3
0,720	0,188	MÉDIO BAIXO	MÉDIO BAIXO	0	0
0,655	0,195	MÉDIO BAIXO	MÉDIO BAIXO	0	0
0,612	0,208	MÉDIO BAIXO	MÉDIO BAIXO	V	V
0,650	0,228	MÉDIO BAIXO	MÉDIO BAIXO	V	V

Tabela 8 – Extrato do ficheiro para atribuição de classes 2 e 3 (2ª abordagem).

Sendo assim, nesta nova versão do "Teste 2", a variável a comparar com o "preço" foi "energias não renováveis/consumo" e no "Teste 3" foi "energias renováveis/consumo":

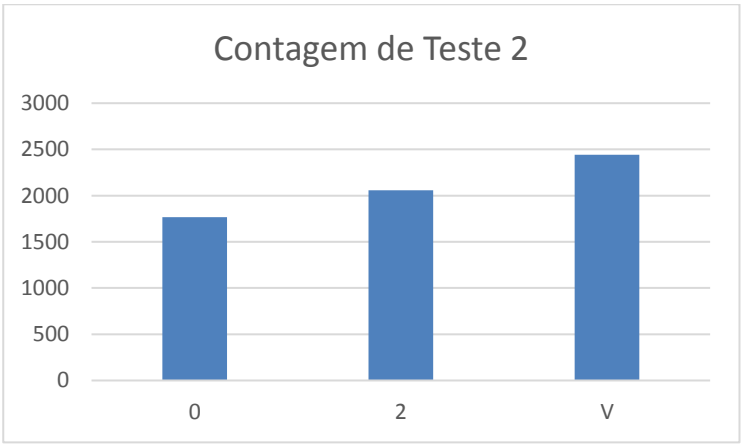


Figura 10 – Gráfico de resultados para o teste 2 (2ª abordagem).

Teste 2	Contagem de Teste 2
0	1766
2	2057
V	2441

Tabela 9 – Resultados teste 2 (2ª abordagem).

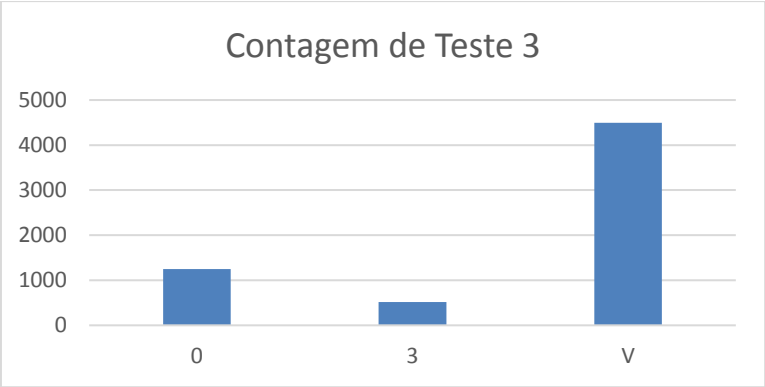


Figura 11 – Gráfico de resultados do teste 3 (2ª abordagem).

Teste 3	Contagem de Teste 3
0	1250
3	516
V	4498

Tabela 10 – Resultados do teste 3 (2ª abordagem).

À semelhança do que muda em relação à 1ª abordagem, no "Teste 4" compara-se com o preço a categoria "importação/consumo" e no "Teste 5" a categoria "exportação/consumo".

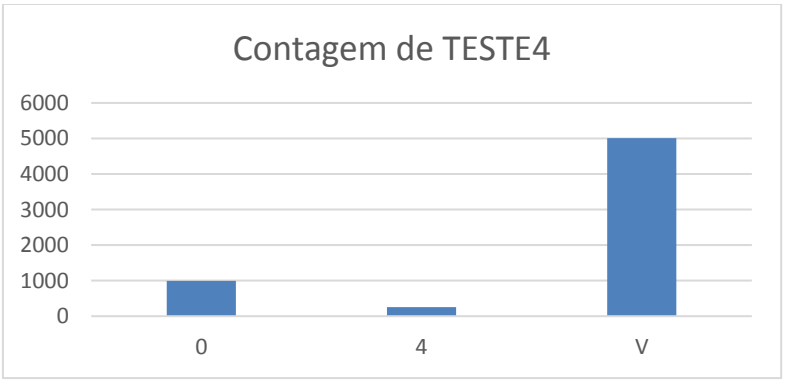


Figura 12 – Gráfico de resultados do teste 4 (2ª abordagem).

TESTE4	Contagem de TESTE4
0	997
4	253
V	5014

Tabela 11 – Resultados do teste 4 (2ª abordagem).

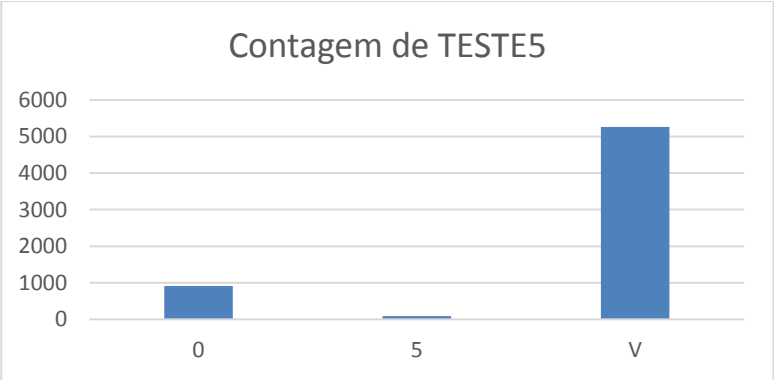


Figura 13 – Gráfico de resultados do teste 5 (2ª abordagem).

TESTE5	Contagem de TESTE5
0	910
5	87
V	5267

Tabela 12 – Resultados do teste 5 (2ª abordagem).

3.3 – Algoritmo imunológico artificial

Esta secção descreve a implementação dum processo de *clustering* baseado em AIS.

Após terminar a tarefa inicial no Excel, procedeu-se à implementação do mesmo algoritmo no MATLAB para verificar a coerência entre os resultados obtidos no Excel e para facilitar a visualização das características dos protótipos das classes.

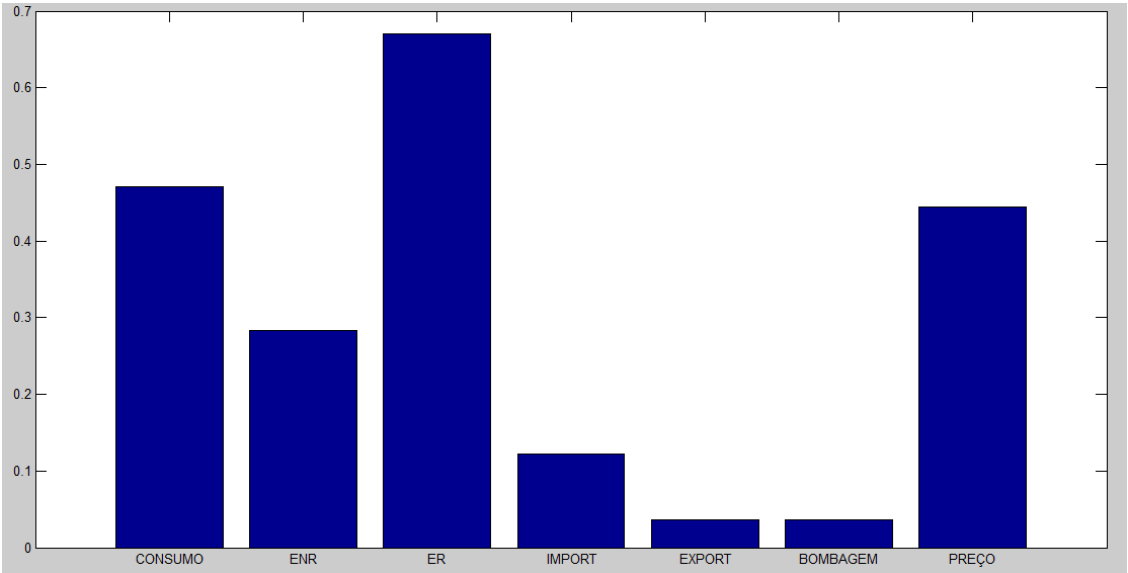


Figura 14 – Protótipo classe 1.

Esta figura demonstra o protótipo resultante da classe 1 criada no algoritmo executado anteriormente e, tal como esperado, demonstra a semelhança entre os níveis de consumo e preço, assim como uma elevada utilização de energias renováveis. Os valores do "consumo" e do "preço" foram normalizados, sendo que as outras variáveis são percentagens do valor do consumo do protótipo.

Em seguida, criou-se o algoritmo de seleção negativa previsto como principal objetivo desta tese de dissertação.

Depois de todos os valores das variáveis estarem preparados conforme descrito para a figura anterior (para os casos do "consumo" e "preço", normalizados; para os outros, divididos pelo consumo da hora, de modo a que representem uma percentagem do consumo), procedeu-se à aplicação do seguinte algoritmo:

1. Classificação apriorística (classificação igual ao 1º teste efetuado no Excel, atribuindo classe quando o grau do consumo é igual ao do preço);
2. Para os restantes casos (fora de 1), gerar N detetores aleatórios;
3. Calcular distância euclidiana de cada caso a cada detetor;
4. Selecionar a distância mínima de cada caso, atribuindo uma classificação provisória consoante o detetor mais próximo;
5. Verificar se essa distância é inferior ao limite pretendido X - neste caso considera-se que há um *match* entre o caso e o detetor;
6. Contar, para cada detetor, quantos *matches* ele tem;
7. Verificar se o número de casos associados ao detetor atingiu o número mínimo de *matches* Y; se confirma, guarda-se o detetor;
8. Identificar quais os casos foram emparelhados a cada detetor;
9. Atribuir classe a cada um desses casos;
10. Contar a quantos casos foi atribuída classe;
11. O ciclo termina quando o resultado de 10 atingir Z ou quando forem efetuados M ciclos, sendo Z o número de casos com classe atribuída.

50x7 double

	1	2	3	4	5	6	7
1	0.2494	0.5113	0.9949	0.0705	0.4394	0.0513	0.2697
2	0.2438	0.0649	0.3881	0.2308	0.3772	0.1962	0.7063
3	0.8093	0.1586	0.5214	0.1586	0.0850	0.0415	0.2143
4	0.3485	0.3111	0.7073	0.1006	0.0549	0.2730	0.4017
5	0.4376	0.4669	0.4300	0.1024	0.2582	0.0489	0.7868
6	0.4377	0.3159	0.5264	0.2322	0.0224	0.0996	0.4010
7	0.8042	0.2013	0.4949	0.5015	0.1901	0.1847	0.4081
8	0.1414	0.4605	0.5956	0.4248	0.0685	0.0287	0.8561
9	0.5566	0.0924	1.0485	0.1662	0.1325	0.0571	0.8884
10	0.1563	0.3992	1.1513	0.2123	0.1897	0.0897	0.9214
11	0.7397	0.1852	1.0863	0.1439	0.1946	0.2807	0.3370
12	0.3264	0.5262	0.9104	0.2639	0.4200	0.1228	0.1805

Tabela 13 – Alguns candidatos a detetores gerados aleatoriamente.

Para o problema desta tese de dissertação foi pré-definido que seriam criados 50 detetores de cada vez, um valor que pode ser alterado pelo utilizador do programa, uma vez que se trata de especificar um parâmetro da aplicação MATLAB.

Nos pontos 4 e 5, calcula-se então a distância euclidiana de cada detetor a cada caso e determina-se a distância mínima de cada caso a um dos detetores (a menor distância indica a classificação). Seguidamente, após alguns testes, definiu-se como 0.4 a distância mínima de semelhança entre ambos, para o caso ser aceite como parte do protótipo do detetor. Após o detetor atingir um número de *matches* considerável, este será então gravado como sendo um protótipo plausível.

6264x1 double

	1
1	0.3065
2	0.2503
3	0.2845
4	0.2824
5	0.2832
6	0.2922
7	0.3064
8	0.3115
9	0.3294
10	0.3126
11	0.2827
12	0.2716

Tabela 14 – Distâncias euclidianas mínimas (a um certo candidato a detetor) das primeiras 12h do ano.

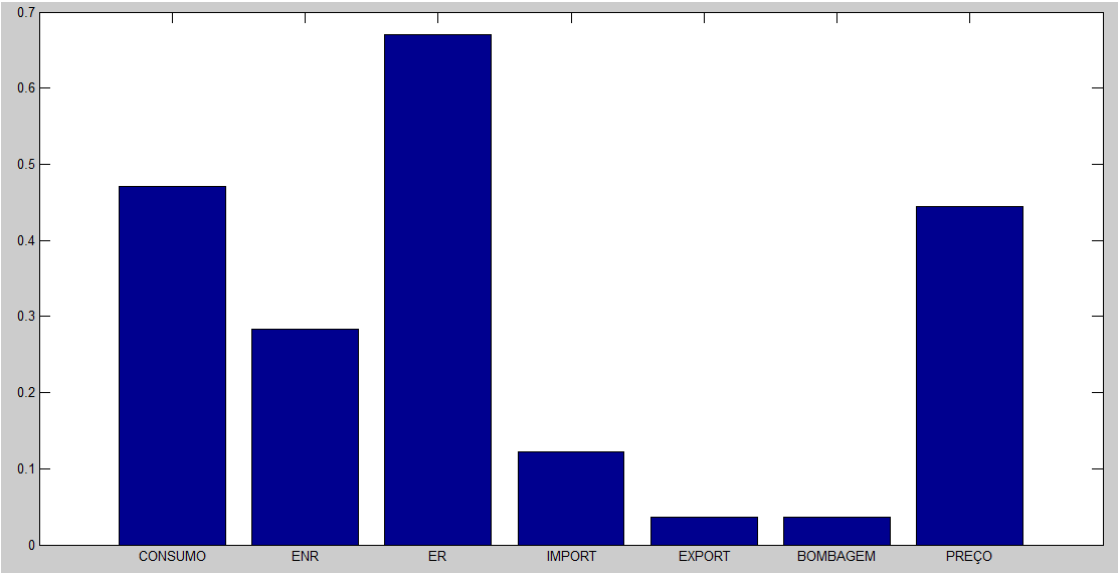
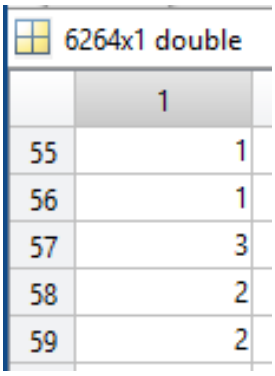


Figura 15 – Protótipo de um detetor.

Para esta dissertação foi decidido que um número aceitável de *matches* seria o triplo do número total de casos a dividir pelo número total de detetores (neste caso, 3*6264/50~=376)[1]. Quando o detetor é gravado como sendo aceite como um "anticorpo" do sistema imunológico, uma classe é atribuída a cada caso semelhante a ele.



6264x1 double

	1
55	1
56	1
57	3
58	2
59	2

Tabela 15 – Classes atribuídas a casos (cada classe representa um detetor, exceto nas classes 0 e 1).

Este processo é iterativo e pode ser alterado consoante a vontade do utilizador. Nesta implementação, o ciclo está programado para parar apenas quando foram emparelhados a um detetor mais de metade dos casos que nos foram fornecidos para análise, ou então quando o ciclo for repetido 100 vezes.

3.4 Previsão de preços

Pretende-se com este estudo verificar se vale ou não a pena separar os dados (casos) por *clusters*. Para analisar os resultados, iremos recorrer à *toolbox nnet* existente no MATLAB. Para tal, é necessário organizar os dados em termos de entradas e saída (valor a prever). Como saída tem-se o "preço", sendo o vetor de entradas constituído por todas as outras variáveis (consumo, energia renováveis, etc.). Para além destas variáveis de entrada, considerou-se também a variável "preço no dia anterior à mesma hora", de forma a incluir mais uma informação potencialmente útil para obter melhores resultados. No entanto, não foi considerado o 1º dia do ano para este estudo, uma vez que não se possui informação acerca do "preço no dia anterior".

Foi então criada uma RN em que as entradas são os dados da matriz de casos e o *target* é a coluna de "preço" da mesma, de forma a posteriormente ser possível comparar os resultados dessa RN com os resultados das RNs de cada classe.

Note-se que estas técnicas de *clustering* se baseiam no pressuposto que está disponível o preço ou uma estimativa do preço. Logo, para estes algoritmos serem aplicados em situação

[1] Este é um ponto a merecer a atenção em futuros desenvolvimentos.

real, teríamos que considerar as estimativas de produção e consumo existentes fornecidas pelas devidas entidades competentes e criar uma RN em que o *target* seria também a previsão de preço a que se tivesse acesso. O esquema seguinte ilustra como se procederia para aplicar estas técnicas na realidade:

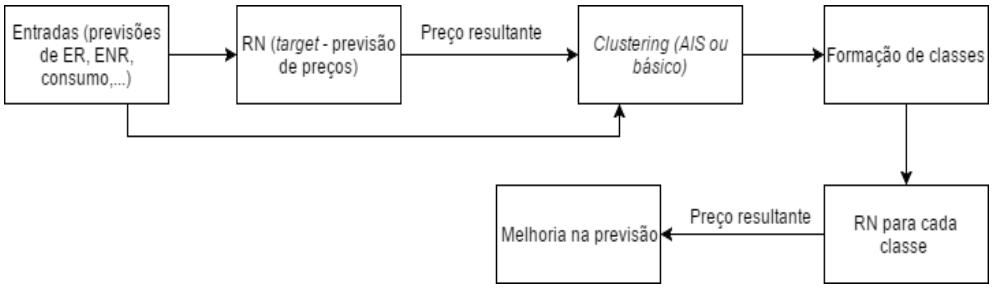


Figura 16 – Esquema ilustrativo de como aplicar os algoritmos em contexto real.

Capítulo 4

Resultados

Este capítulo é dedicado à apresentação dos resultados mais relevantes do estudo desenvolvido. Irá ser efetuada uma análise às diferenças entre os *clusters* criados pelos diferentes algoritmos, assim como uma comparação da eficiência de ambos os métodos (*clustering* básico e do algoritmo AIS) em relação à previsão do PE. Para isso serão comparados os erros MAE, MAPE' e RMSE.

4.1 *Clustering* básico (2ª abordagem)

Tal como já foi demonstrado na secção da Metodologia, o protótipo da classe 1, tal como esperado, apresenta valores relativos de "preço" e "consumo" bastante semelhantes (preço≈consumo). Neste contexto, o símbolo "≈" significa "varia de acordo com", ou seja, o preço é elevado quando o consumo é elevado e será baixo quando o consumo também for baixo. Para além disso, todos os valores das restantes variáveis parecem variar num intervalo relativamente restrito, não apresentando grandes desvios em relação às suas médias. Passemos então à análise dos *clusters* seguintes:

Classe 2

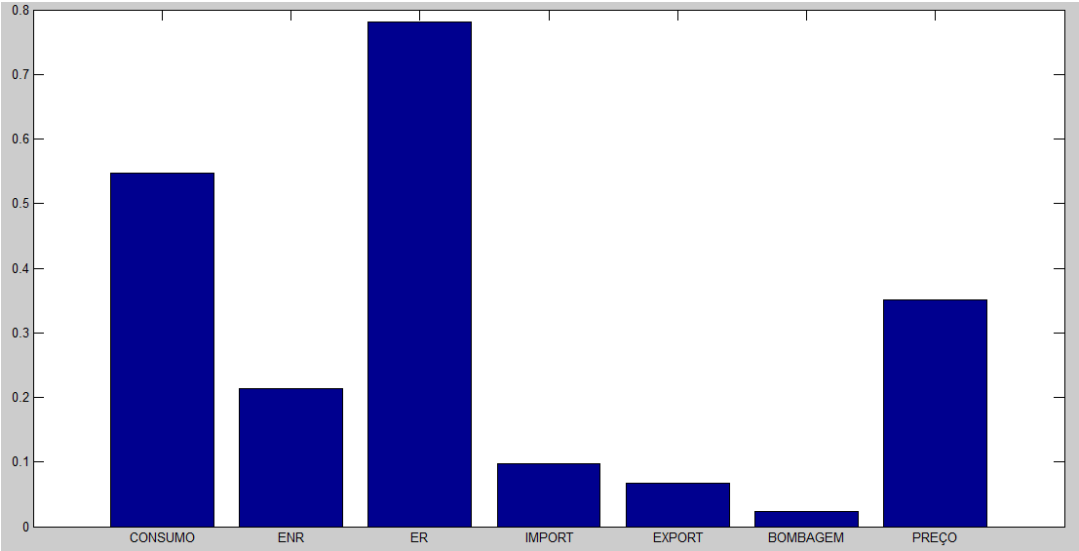


Figura 17 – Protótipo classe 2 (2ª abordagem, ENR≈preço).

Esta classe agrega todos os casos em que ENR≈Preço – a Energia Não Renovável (ENR) varia de acordo com o Preço.

Apesar de não ser evidente nesta figura, os níveis de ENR e Preço estão dentro do esperado, uma vez que a gama de valores do rácio entre ENR e consumo na mesma hora estão compreendidos entre 0.014 e 0.569, o que significa que o valor próximo de 0.2 demonstrado no gráfico pode estar correlacionado com o valor perto de 0.4 do Preço, que se encontra normalizado. Aliás, o nível "médio baixo" está compreendido para o rácio entre 0.125 e 0.236 e para o preço entre 0.2 e 0.4, o que suporta ainda mais o resultado deste gráfico. A barra de ENR parece estar um bocado alta, o que deve contribuir para o afastamento da barra de Preço do limite superior do nível "médio baixo".

Classe 3

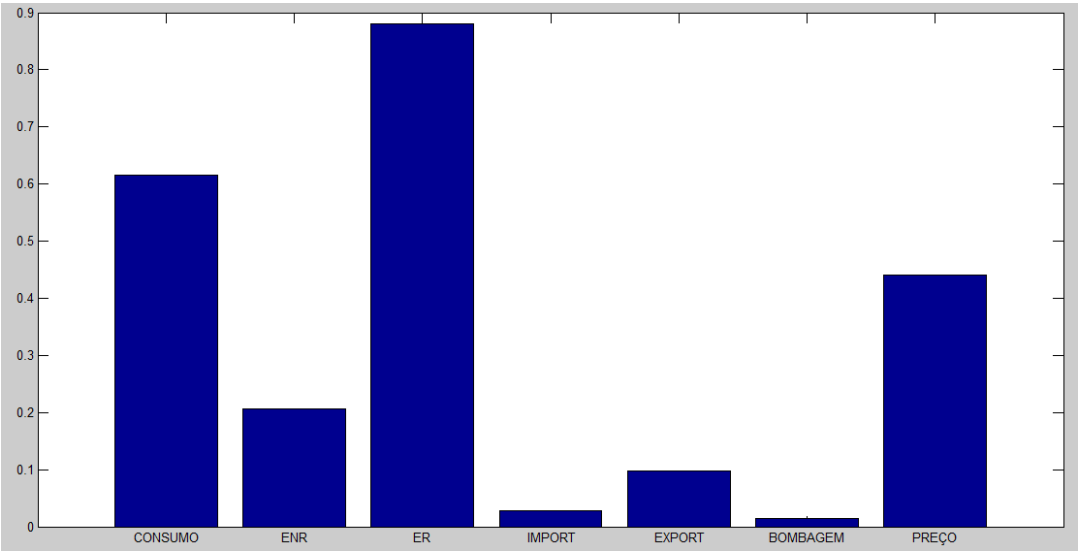


Figura 18 – Protótipo classe 3 (2º método, ER !≈ preço).

Os valores do rácio entre ER e Consumo estão compreendidos entre 0.232 e 1.46. Entre 0.724 e 0.97 o seu nível é "médio", assim como o do Preço entre 0.4 e 0.6. Logo, este gráfico também se encontra dentro do previsto, sendo este o único nível em que faz sentido que coexistam. Note-se que neste caso a ER e o Preço variam de forma simétrica – usa-se aqui o símbolo "!≈" para representar este tipo de relação.

Classe 4

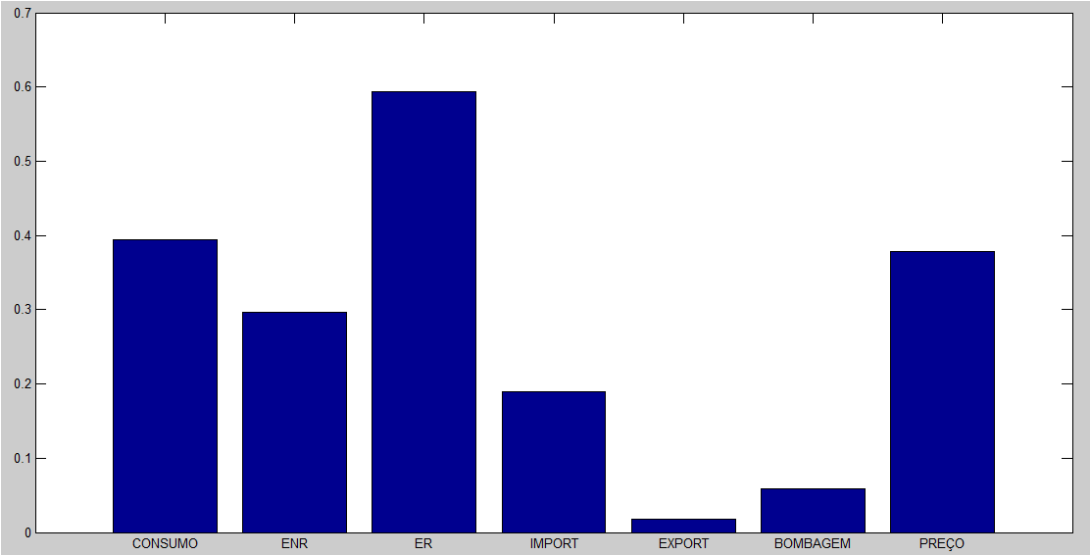


Figura 19 – Protótipo classe 4 (2ª abordagem, importação≈preço).

Para o rácio entre Importação e Consumo, os valores estão compreendidos entre 0 e 0.525. O nível "médio baixo" encontra-se entre 0.105 e 0.21 e, tal como já foi visto anteriormente, a barra de Preço está dentro do mesmo nível. De notar uma maior presença de ENR, que terá possivelmente uma contribuição significativa para o valor do Preço.

Classe 5

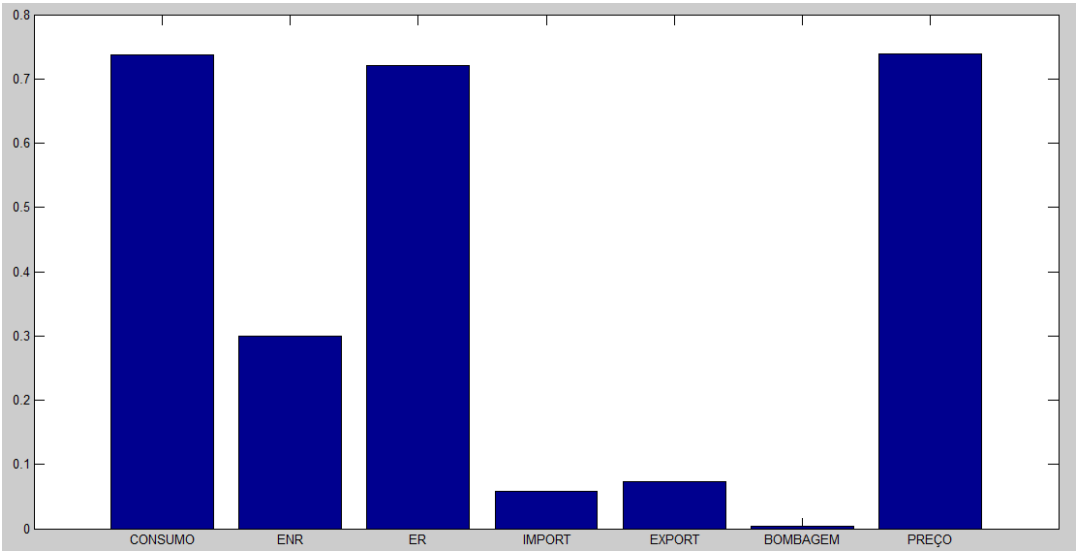


Figura 20 – Protótipo classe 5 (2ª abordagem, exportação !≈ preço).

O rácio entre Exportação e Consumo engloba os valores entre 0 e 0.453 e o valor demonstrado na figura parece estar no limite entre os níveis "baixo" e "médio baixo". Logo, para haver concordância, o nível do Preço deveria estar no limite entre os níveis "médio alto" e "alto", o que não acontece, parecendo estar próximo do limite inferior do nível "médio alto". No entanto, um maior nível de Consumo, assim como de ENR, podem explicar esta elevada barra de Preço e o pequeno número de ocorrências para esta classe (apenas 87) pode também ajudar a explicar as discrepâncias entre os valores obtidos e os esperados. De facto, os períodos de Exportação são relativamente escassos, o que significa que o valor médio será sempre baixo. Estas circunstâncias sugerem que esta variável (tal como a Importação) possam ter um tratamento diferenciado, no qual seriam apenas considerados os casos positivos, isto é, os casos em que existe Exportação.

4.2 Algoritmo AIS

Tal como no *clustering* básico, a Classe 1 é determinada a priori, englobando os casos Consumo≈Preço. Passemos agora à demonstração e análise dos outros agrupamentos que foram criados pelos detetores do algoritmo AIS:

Classe 2

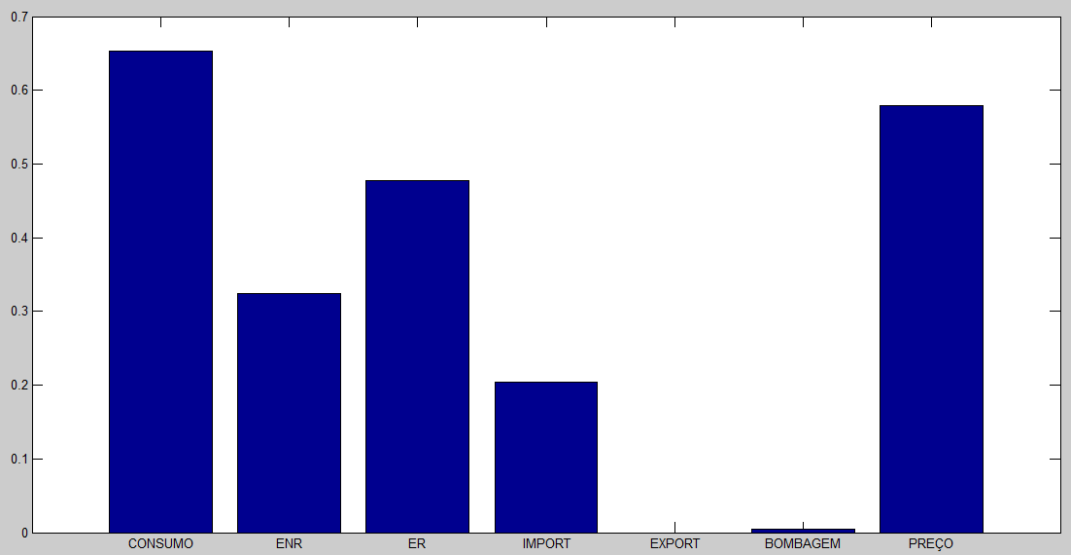


Figura 21 – Protótipo classe 2 (AIS).

Esta classe 2 parece agrupar os casos em que tanto o Consumo como o Preço são elevados, sendo ENR relativamente elevada, ER relativamente baixa e uma quantidade considerável de importação. Esta classe engloba 379 casos.

Classe 3

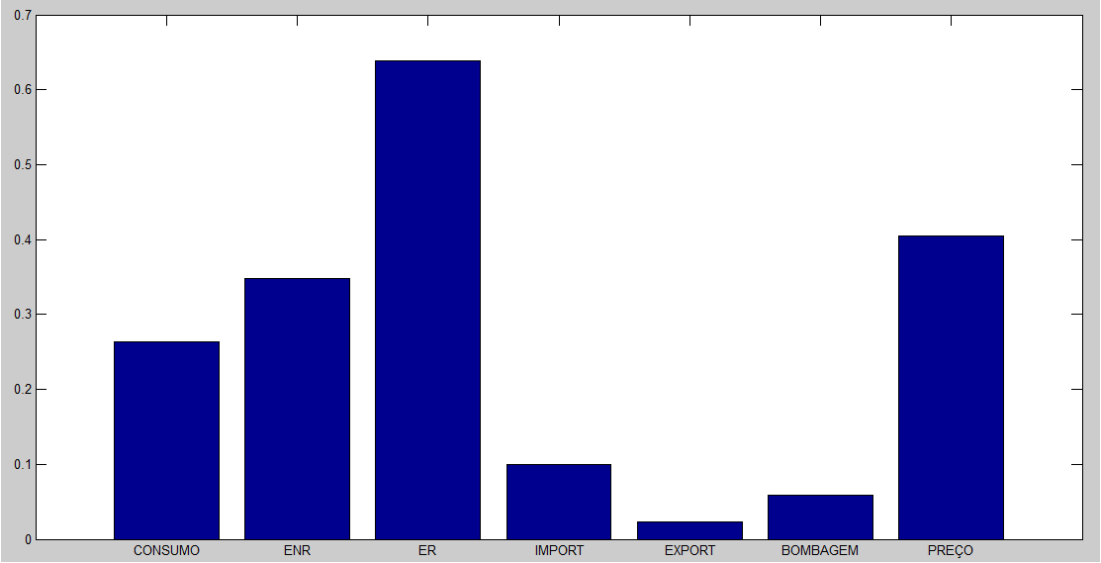


Figura 22 – Protótipo classe 3 (AIS).

Esta classe parece ser caracterizada por níveis baixos de Consumo e um Preço um pouco abaixo da média, mas uma relativamente alta utilização de ENR. Provavelmente, esta presença de ENR será responsável pela excedência do nível de Preço em relação ao Consumo, assim como uma pequena mas não residual quantidade de Importação e Bombagem. Esta classe contém 647 casos associados.

Classe 4

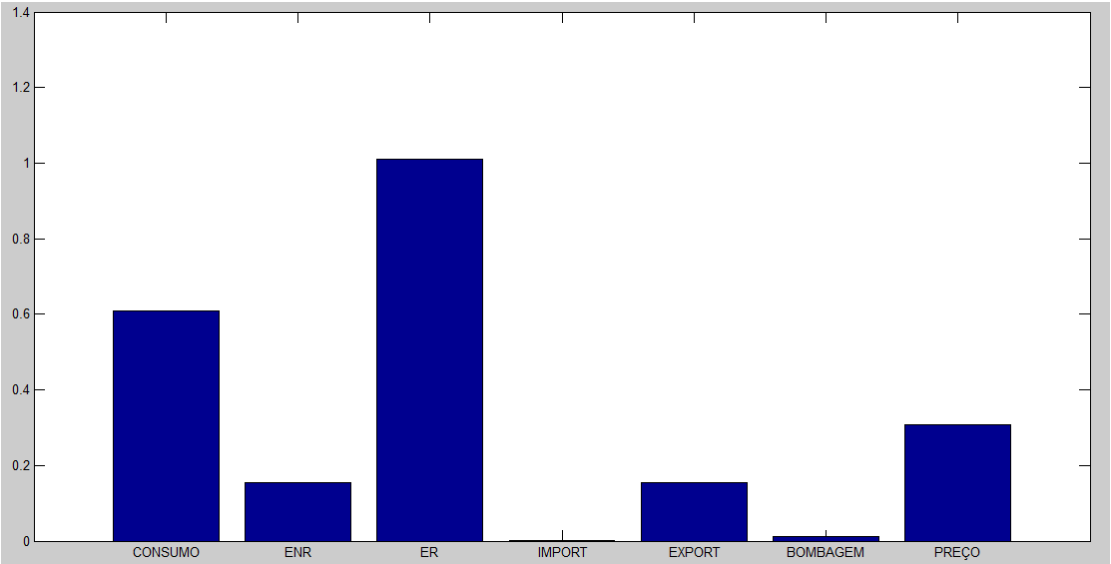


Figura 23 – Protótipo classe 4 (AIS).

Esta classe parece evidenciar o papel das ER no Preço: ER elevadas dão origem a preços mais baixos. Uma quantidade considerável de Exportação também ajuda a esse facto. 610 casos estão atribuídos a esta classe.

Classe 5

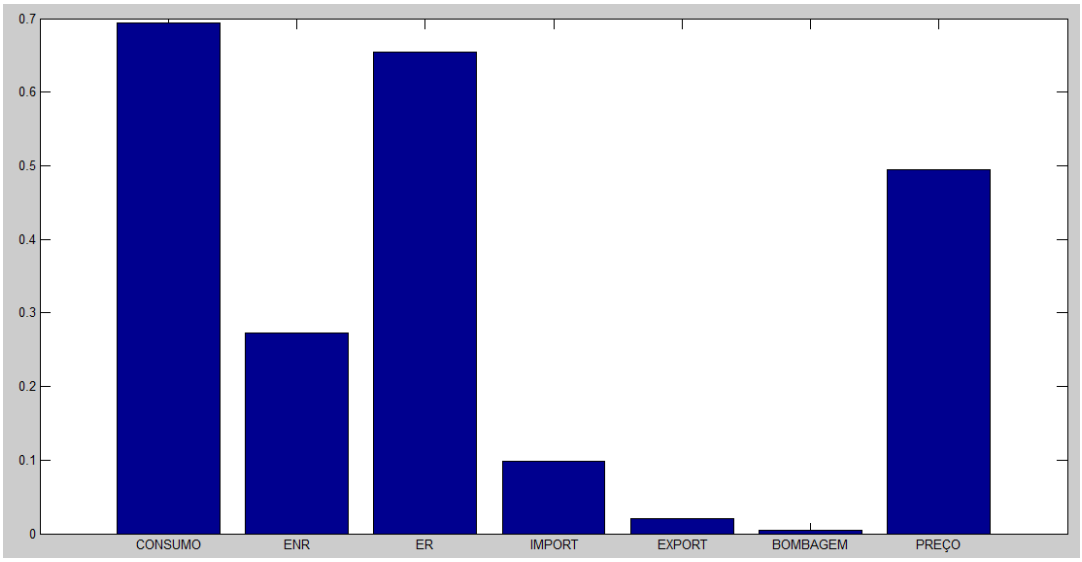


Figura 24 – Protótipo classe 5 (AIS).

A classe 5 não parece ser muito ortodoxa: Consumo um pouco alto, valores de ER, ENR e Preço dentro da média. Evidencia que talvez uma proporção equilibrada (nas diferentes escalas) de ER e ENR tenha mais influência no Preço do que o próprio Consumo. Esta classe inclui 840 casos.

4.3 RNs para previsão de preços

4.3.1 RN geral

Primeiramente, foi criada a RN geral para todos os casos (portanto, sem atribuição de classes). Da previsão efetuada resultaram médias de erro demonstradas na seguinte tabela:

MAE geral
5,3876
MAPE' geral
11,9%
RMSE geral
7,3819

Tabela 16 – Erros da RN geral em relação à média dos preços reais.

O MAE foi de cerca de 5.5EUR, enquanto que o MAPE' indica uma média de erros à volta de 12% e o RMSE foi de aproximadamente 7.5EUR.

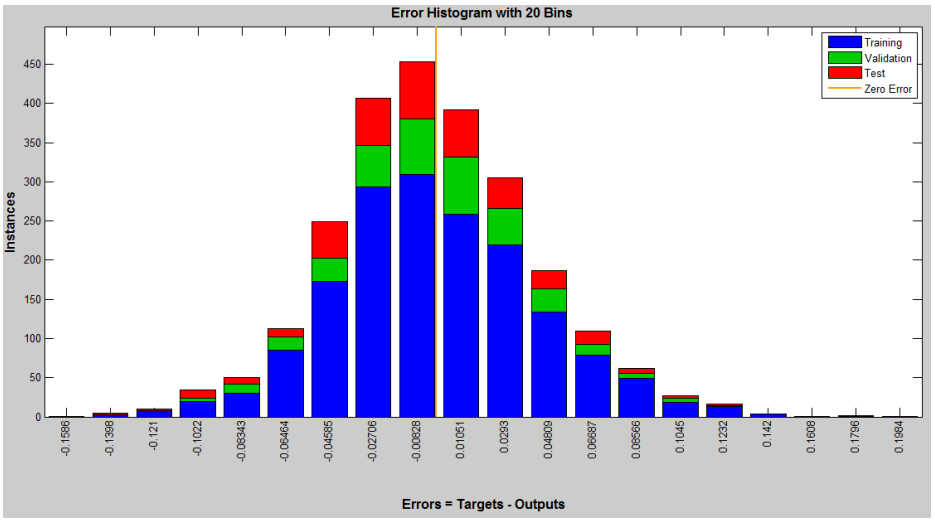


Figura 25 – Histograma de erros da RN geral.

O histograma de erros da RN demonstra que a maior parte dos erros situam-se à volta do 0, o que é indicador de bons resultados.

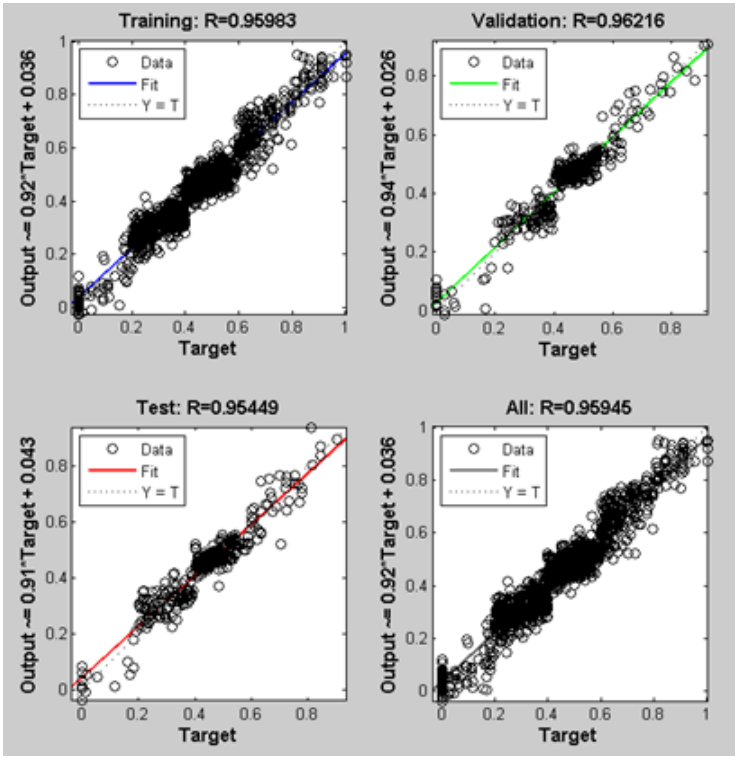


Figura 26 – *Plot regression* da RN geral.

Como se pode observar nesta última figura, a grande parte dos resultados da RN estão sobre o eixo Y=X e o coeficiente R está muito perto de 1, o que também são bons indicadores de resultados.

Em seguida, iremos verificar se a formação de classes conseguiu reduzir o erro da previsão inicial.

4.3.2 CB

Para testar a utilidade do *clustering básico*, foram treinadas redes neuronais (RN) com os dados separados pelas classes básicas. Depois de obtidas as classes, calculou-se o MAE, MAPE e RMSE:

MAE geral	MAE classe	redução
5,3876	3,9828	26%
MAPE'geral	MAPE'Classe	
11,9%	8,8%	26%
RMSEgeral	RMSEclasse	
7,3825	5,2926	28%

Tabela 17 – Erros calculados para o *clustering* básico (2ª abordagem).

Uma redução em 26% dos erros calculados demonstra que este algoritmo melhorou bastante a eficiência da RN na previsão do PE.

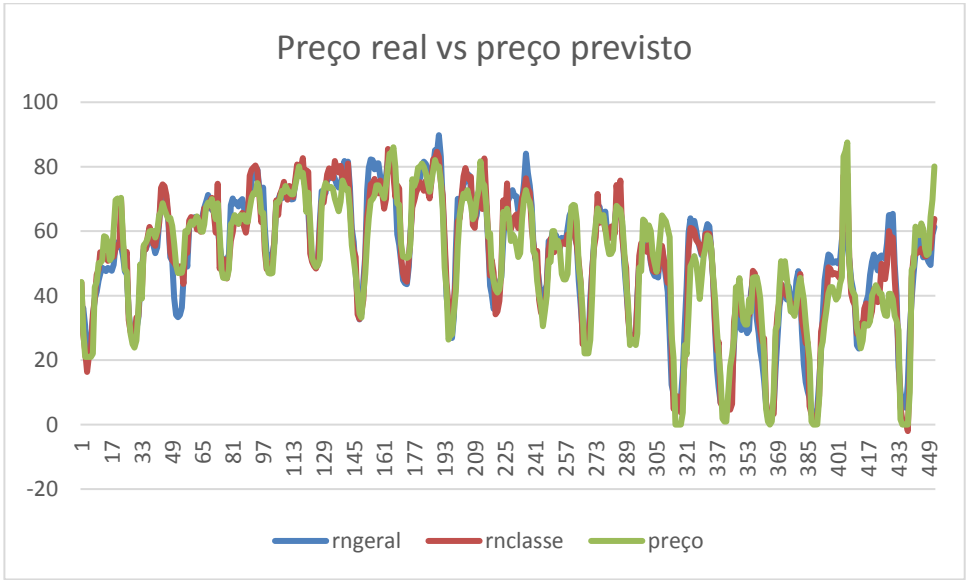


Figura 27 – Preço real vs preço previsto (CB).

4.3.3 AIS

Finalmente, para testar se o algoritmo baseado em AIS é de facto uma ferramenta útil para a previsão do PE, foram treinadas redes neuronais (RN) com os dados separados pelas classes criadas pelos detetores escolhidos.

Para o AIS podemos observar que a sua atribuição de classes fez melhorar a previsão em aproximadamente 18%, comparativamente à previsão efetuada pela RN geral.

MAE geral	MAEclasse		redução
5,3876	4,3973		18,4%
MAPE' geral	MAPE'Classe		
11,9%	9,7%		18,4%
RMSE geral	RMSEclasse		
7,3819	6,0121		18,6%

Tabela 18 – Comparação de resultados (AIS).

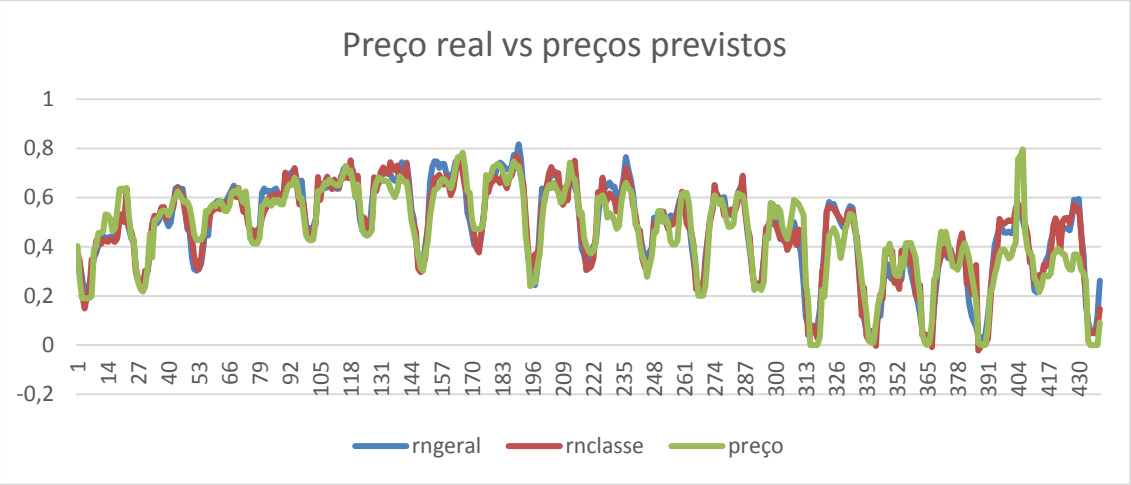


Figura 28 – Excerto de comparação de preços (AIS).

Capítulo 5

Conclusão e trabalho futuro

Este trabalho consiste em duas fases principais: uma primeira está relacionada com a identificação de *clusters* (separação dos casos em classes típicas), enquanto que na segunda pretendeu-se verificar se a divisão dos dados em *clusters* pode ser efetivamente útil na previsão de preços. A identificação de *clusters* permite obter grupos de casos semelhantes, o que é interessante em termos de caracterização de mercado: casos parecidos deverão ter, em princípio, um preço final bastante próximo.

A análise dos protótipos dos *clusters* permitiu identificar as principais características de cada um dos grupos, o que poderá ser útil para interpretar a influência das diferentes variáveis no preço e no comportamento do mercado. Sabendo as previsões de consumo e produção dos diversos tipos de energia, pode-se comparar graficamente com os *clusters* obtidos e concluir uma aproximação do valor do preço do caso previsto.

Tal como esperado, os testes de previsão efetuados permitiram obter ganhos de performance (entre 18% e 26%). Depois de comparar os erros mencionados nesta dissertação para ambos os algoritmos, podemos concluir que o *clustering* básico foi mais eficiente que o AIS na previsão do PE. No entanto, uma maior rigidez em um ou outro parâmetro do AIS poderia perfeitamente contribuir para um aumento da redução de erro em relação à RN criada inicialmente (por exemplo, diminuir a distância da medida de similaridade utilizada).

Logo, uma das possibilidades de continuar o trabalho realizado nesta dissertação será repetir o AIS até obter melhores resultados. Outra sugestão será treinar uma RN para o CB da 1ª abordagem (com os valores nominais) para comparar resultados. Por fim, também se poderá incluir nos estudos a variável "bombagem", que só não foi considerada nesta dissertação pelo seu valor máximo ser muito reduzido (0.28) e não se ter realizado um estudo profundo acerca da influência desta variável.

Referências

[1] Mercado Ibérico de Eletricidade (MIBEL). Disponível em: http://www.mibel.com/index.php?mod=pags&mem=detalle&relmenu=9&relcategoria=1026&idpag=67.

[2] Rafał Weron. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 2014.

[3] Lalit Mohan Saini Sanjeev Kumar Aggarwal e Ashwani Kumar. Electricity price forecasting in deregulated markets: A review and evaluation. *Elsevier Ltd.*, 2009.

[4] Stefan Trueck Adam Misiorek e Rafal Weron. Point and interval forecasting of spot electricity prices: Linear vs. non-linear time series models. *The Berkeley Electronic Press*, 2006.

[5] Xing Yan e Nurul A. Chowdhury. Midterm electricity market clearing price forecasting using two-stage multiple support vector machine. *Journal of Energy*, 2015.

[6] Antonio J Conejo Rosario Espínola Francisco J Nogales, Javier Contreras. Forecasting nextday electricity prices by time series models. *IEEE Transactions on Power Systems*, 2002.

[7] L. Alfredo Fernandez-Jimenez Cláudio Monteiro e Ignacio J. Ramirez-Rosado. Explanatory information analysis for day-ahead price forecasting in the iberian electricity market. *Energies*, 2015.

[8] Jacques Lawarree Massimo Gallanti Andrea Venturini Guang Li, Chen-Ching Liu. State-ofthe-art of electricity price forecasting. *IEEE*, 2005.

[9] Patrick Billingsley. *Probability and Measure*. John Wiley and Sons, 3rd ed. edição, 1995.

[10] Nate Silver. *The Signal and the Noise*. The Penguin Press, 1st ed. edição, 2012.

[11] José P. M. Conde. Previsão de preços de eletricidade no mercado diário e intradiário MIBEL. Tese de mestrado, Faculdade de Engenharia da Universidade do Porto, 2015.

[12] "2.5 Evaluating forecast accuracy | OTexts". Disponível em: https://www.otexts.org/fpp/2/5.

[13] Hyndman, R. and Koehler A. (2005). "Another look at measures of forecast accuracy".

[14] Tofallis (2015). "A Better Measure of Relative Prediction Accuracy for Model Selection and Model Estimation", *Journal of the Operational Research Society*, 66(8),1352-1362.

[15] Rafal Weron e Adam Misiorek. Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. *International Journal of Forecasting*, 2008.

[16] Z. Xu J.H. Zhao, Z.Y. Dong e K.P. Wong. A statistical approach for interval forecasting of the electricity price. *IEEE Trans on Power Systems*, 2008.

[17] *Clustering*. Disponível em: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/.

[18] "Cluster Analysis: Basic Concepts and Algorithms". Disponível em: https://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

[19] Estivill-Castro, Vladimir (20 June 2002). "Why so many clustering algorithms — A Position Paper". *ACM SIGKDD Explorations Newsletter*.

[20] Rosie Cornish (2007). "Cluster Analysis". Disponível em: http://www.statstutor.ac.uk/resources/uploaded/clusteranalysis.pdf

[21] Karimpour, A.; and Pariz, N.; Sokhanvar, Kh.; "Electricity Price Forecasting Using a Clustering Approach". *2nd IEEE International Conference on Power and Energy (PECon 08)*. 2008, Johor Baharu, Malaysia.

[22] Li Xie, Hua Zheng, Lizi Zhang. "Electricity price forecasting by clustering-LSSVM". *Power Engineering Conference, 2007. IPEC*, 2007.

[23] de Castro, Leandro N.; Timmis, Jonathan (2002). *Artificial Immune Systems: A New Computational Intelligence Approach*. Springer.

[24] Abbod, M.F.; Al-Enezi, J.R.; Alsharhan, S. (2010). "ARTIFICIAL IMMUNE SYSTEMS – MODELS, ALGORITHMS AND APPLICATIONS". Disponível em: http://dspace.brunel.ac.uk/bitstream/2438/4643/1/Fulltext.pdf

[25] Jason Brownlee. "Clever Algorithms: Nature-Inspired Programming Recipes". Disponível em: http://www.cleveralgorithms.com/nature-inspired/immune.html

[26] Forrest, S.; Perelson, A.S.; Allen, L.; Cherukuri, R. (1994). "Self-nonself discrimination in a computer". *Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy*. Los Alamitos, CA.

[27] AISWeb: The Online Home of Artificial Immune Systems. Disponível em: http://www.artificial-immune-systems.org/algorithms.shtml

[28] Bessa, R.B.; Fidalgo, J.N.; Lima, F.P.A.; Minussi, C.R.; " A modified negative selection algorithm applied in the diagnosis of voltage disturbances in distribution electrical systems". *18th International Conference on Intelligent System Application to Power Systems, ISAP 2015*.

[29] Paulo Cortez e José Neves. *Redes Neuronais Artificiais*, 2000.

[30] V. M. F. Mendes L. F. M. Ferreira J. P. S. Catalão, S. J. P. S. Mariano. Short-term electricity prices forecasting in a competitive market: A neural network approach. *Electric Power Systems Research*, 2007.

[31] S. N. Sivanandam e S. Sumathi. *Introduction to Neural Networks Using MATLAB 6.0*, 2005.

[32] Georgilakis PS. Market clearing price forecasting in deregulated electricity markets using adaptively trained neural networks, 2006.

[33] Mandal P, Senjyu T, Funabashi T. Neural networks approach to forecast several hour ahead electricity prices and loads in deregulated market. *Energy Conversion and Management, 2006*.

[34] Li C, Wang S. Next-day power market clearing price forecasting using artificial fishswarm based neural networks. *Advances in Neural Networks - ISNN 2006*.

[35] Gao F, Cao X, Papalexopoulous A. Forecasting power market clearing price and quantity using a neural network method. *IEEE Power Engineering Society Summer Meeting*, 2000.

[36] Wang A, Ramsay B. A neural network based estimator for electricity spot pricing with particular reference to weekend and public holidays. *Neurocomputing 23*, 1998.
[37] Yamin HY, Shahidehpour SM, Li Z. Adaptive short-term electricity price forecasting using artificial neural networks in the restructured power markets, 2004.
[38] Szkuta BR, Sanabria LA, Dillon TS. Electricity price short-term forecasting using artificial neural networks. *IEEE Transactions on Power Systems*, 1999.
[39] Gareta R, Romeo LM, Gil A. Forecasting of electricity prices with neural networks, 2006.