

Relatório Intercalar

Redes Neurais para a identificação de Pulsares



Mestrado Integrado em Engenharia Informática e Computação

IART - Inteligência Artificial

Turma 3MIEIC02, Grupo E1_3 :

André Miguel Ferreira da Cruz - 201503776

Edgar Filipe Amorim Gomes Carneiro - 201503784

João Filipe Lopes de Carvalho - 201504875

Faculdade de Engenharia da Universidade do Porto
Rua Roberto Frias, s/n, 4200-465 Porto, Portugal

8 de Abril de 2018

Conteúdo

1	Objetivo	2
2	Descrição	2
2.1	Especificação	2
2.1.1	Descrição e análise do dataset.	2
2.1.2	Pré-processamento dos dados.	3
2.1.3	Modelo de aprendizagem a aplicar: redes neuronais.	3
2.1.4	Redes neuronais: arquitetura, configuração prevista da rede.	4
2.2	Trabalho Efetuado	6
2.3	Resultados esperados e forma de avaliação	6
3	Conclusões	7
4	Recursos	8

1 Objetivo

Este trabalho tem como objetivo a aplicação de Redes Neurais artificiais na identificação de Pulsares.

Os pulsares são um tipo de estrela rara de neutrões que produzem emissões rádio detetáveis na terra. Estes, através do seu feixe emissor e da sua capacidade de rotação, emitem um padrão periódico que os torna descobríveis. Desta forma, fazer a deteção de um pulsar corresponde à procura e reconhecimento dos seus sinais periódicos.

Cada pulsar produz um padrão diferente que varia ligeiramente dos restantes. Assim, podemos detetar um sinal realizando uma análise de alguns parâmetros, como por exemplo o tempo médio que um pulsar demora a reemitir o seu feixe para a terra. À primeira vista concluir se um sinal é realmente um pulsar pode parecer pouco complexo. No entanto, verdadeiramente, a grande maioria das deteções são causadas por interferências de frequências de rádio e de ruído, tornando complicado encontrar sinais genuínos.

Devido às adversidade referidas anteriormente, o processo de deteção de pulsares torna-se bem mais complexo e por recorre-se á utilização de *Machine Learning*. *Machine Learning* é um sub-campo da área da inteligência artificial no qual as aplicações de *software* têm a capacidade de aprender e obter resultados precisos, sem que no entanto sejam explicitamente programadas. Os algoritmos de *machine learning* funcionam através da construção de um modelo a partir de inputs amostrais, com a finalidade de fazer previsões ou decisões orientadas pelos dados e respetivas estatísticas ao invés de seguir estruturas de decisão estáticas.

Assim sendo, neste trabalho será usado um sub-ramo da área de *machine learning*, redes neurais artificiais, como forma de obtenção de resultados para o problema apresentado.

2 Descrição

2.1 Especificação

2.1.1 Descrição e análise do dataset.

Na base de dados fornecida, cada candidato é caraterizado por oito variáveis contínuas e uma classe. As oito variáveis contínuas distinguem-se em dois grupos sendo que as primeiras quatro referem-se a estatísticas relativas ao *folded profile* de um pulsar. Este consiste num *array* de variáveis contínuas que descrevem uma versão pós-análise do pulsar. As variáveis do segundo grupo são obtidas através da análise da curva *DM-SNR* - *Delta Modulation with Signal-to-Noise Ratio*. Esta consiste na modulação de um sinal de analógico para digital ou o seu reverso. Por sua vez, a classe fornecida como última parâmetro tem um binário, indicando se o candidato era de facto um pulsar ou não. Assim, os nove parâmetros analisados são:

1. Média do *folded profile*
2. Desvio padrão do *folded profile*
3. Curtose do *folded profile*
4. Assimetria do *folded profile*
5. Média da curva *DM-SNR*
6. Desvio Padrão da curva *DM-SNR*
7. Curtose da curva *DM-SNR*
8. Assimetria da curva *DM-SNR*
9. Classe resultado

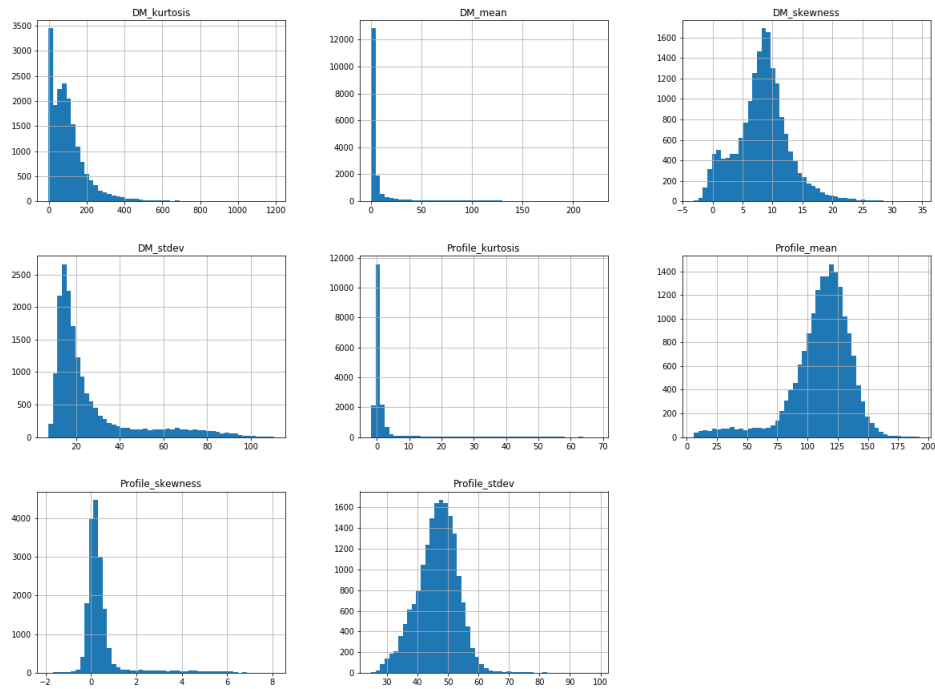


Figura 1: Distribuições relativas às 8 variáveis contínuas

A base de dados fornecida contém um total de 17,898 entradas, sendo que destas, 16,259 tratam-se de dados espúrios causados pelo ruído e as restantes 1,639 tratam-se de dados relativos a exemplos de pulsares reais. Foi ainda possível obter a distribuição associada a cada uma das variáveis contínuas, estando estas presentes na figura 1. Essas mesmas distribuições encontram-se pormenorizadas na figura 2, sob a forma de tabela.

Nas situações de teste, os valores das oito variáveis são usados na análise do candidato de forma a determinar o valor binário que terá a classe passada como nono parâmetro.

2.1.2 Pré-processamento dos dados.

No pré-processamento de dados, é garantido que os dados são todos numéricos e *floats* não havendo qualquer classe para além da classe binária de *output*. É também assegurado que são usadas escalas semelhantes ao longo de todo o trabalho, através de uma normalização dos dados de *input*.

Em relação à classe de *output*, esta é extraída para uma matriz unidimensional em que cada elemento tem um valor binário: se tiver valor 1, significa que se trata de um pulsar, se tiver valor 0 significa que não é pulsar.

2.1.3 Modelo de aprendizagem a aplicar: redes neuronais.

Neste trabalho será usada uma rede neuronal artificial que segue um modelo especial de **aprendizagem supervisionada** denominado *backpropagation*. Assim, torna-se importante esclarecer cada um destes conceitos.

	Profile_mean	Profile_stdev	Profile_skewness	Profile_kurtosis	DM_mean	DM_stdev	DM_skewness	DM_kurtosis
count	17898.000000	17898.000000	17898.000000	17898.000000	17898.000000	17898.000000	17898.000000	17898.000000
mean	111.079968	46.549532	0.477857	1.770279	12.614400	26.326515	8.303556	104.857709
std	25.652935	6.843189	1.064040	6.167913	29.472897	19.470572	4.506092	106.514540
min	5.812500	24.772042	-1.876011	-1.791886	0.213211	7.370432	-3.139270	-1.976976
25%	100.929688	42.376018	0.027098	-0.188572	1.923077	14.437332	5.781506	34.960504
50%	115.078125	46.947479	0.223240	0.198710	2.801839	18.461316	8.433515	83.064556
75%	127.085938	51.023202	0.473325	0.927783	5.464256	28.428104	10.702959	139.309330
max	192.617188	98.778911	8.069522	68.101622	223.392141	110.642211	34.539844	1191.000837

Figura 2: Distribuições relativas às oito variáveis contínuas, sob a forma de tabela

Algoritmos que seguem o modelo da aprendizagem supervisionada são algoritmos nos quais os possíveis *outputs* do algoritmo já são conhecidos e nos quais os dados usados para treinar o algoritmo já se encontram mapeados à resposta correta. Pode-se verificar que este modelo é aplicável ao problema dos pulsares na medida em que a totalidade das entradas na base de dados se encontram mapeados a um resultado, como já foi referido previamente.

Algoritmos de *backpropagation* são algoritmos de aprendizagem supervisionada que usam a noção de gradiente descendente.

O algoritmo do gradiente descendente é um algoritmo iterativo de otimização de primeira ordem para encontrar o mínimo de uma função. Em cada iteração, de forma a encontrar o mínimo da função, o algoritmo avança de forma proporcional ao valor negativo do gradiente naquele ponto da função.

Dada uma rede neuronal artificial e um função de erro o algoritmo de *backpropagation*, também conhecido por algoritmo da propagação de erros para trás (*backward propagation of errors*), calcula o gradiente da função de erro considerando sempre o peso das arestas na rede neuronal. Este cálculo é propagado para trás na rede neuronal, com o gradiente da última camada de pesos a ser calculado primeiro e o gradiente da primeira camada de pesos em último lugar (daí o seu nome *backpropagation*). Este fluxo invertido da informação de erro permite uma computação eficiente do gradiente em cada camada, ao invés da abordagem ingénua de calcular o gradiente de cada camada separadamente.

O algoritmo desenvolve-se em duas fases. Numa primeira fase, propaga-se através da rede de forma a chegar aos valores *output*. De seguida, calculo o custo da propagação para cada output, sendo que esta estará intrinsecamente relacionada com a função de erro usada no algoritmo do gradiente descendente. Por fim, propaga-se em sentido contrário de forma a gerar um valor de erro para cada um dos neurónios. Na segunda fase, os pesos das arestas são constantemente atualizados conforme o resultado desempenhado por cada um dos candidatos. Na atualização das arestas é usado uma percentagem previamente definida como fator de atualização dos pesos. Quanto mais elevado for este fator, mais rapidamente será o treino da rede neural. Por outro lado, quanto mais baixo for este fator, mais preciso será o treino.

2.1.4 Redes neuronais: arquitetura, configuração prevista da rede.

As redes neuronais, surgiram com base na perceção biológica das interligações presentes no cérebro humano. Sendo o cérebro humano constituído por neurónios, os quais recebem impulsos nervosos como entrada e saída de dados. Um neurónio tem a capacidade de estimular os seus neurónios vizinhos, quando este é propriamente estimulado (função de ativação). Posto isto, podemos definir matematicamente um neurónio como sendo:

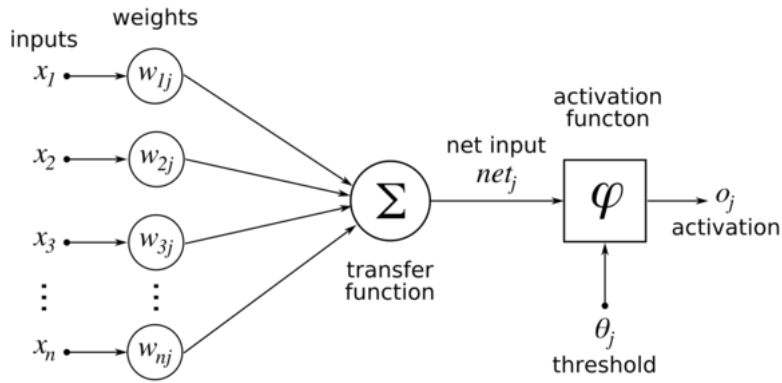


Figura 3: Modelo Matemático de um Neurônio

$$y = f_{ativação}(\sum_i w_i x_i - \theta)$$

Figura 4: Representação Matemática da Saída de um Neurônio

Como o nome indica, é uma rede, por isso a sua consistência detêm vários neurónios conectados na forma de um grafo acíclico. As redes neuronais mais comuns, *feedforward*, são redes em que o *output* de certos neurónios é o *input* de outros. Desta forma é possível separar a rede em diferentes camadas:

1. Uma **Input Layer**, primeira camada na rede neuronal, responsável por receber os dados, determinar o tamanho do *dataset* e apresentar os padrões de reconhecimento. São apenas constituídas por neurónios de *input*.
2. Uma ou mais **Hidden Layer**, é nesta(s) camada(s) que é acontece toda a "magia", isto é, é nesta(s) camada(s) que é feito praticamente todo o processamento dos dados introduzidos na nossa rede neuronal, fazendo a extração das características dos dados.
3. Uma **Output Layer**, última camada da rede, responsável por apresentar os dados, esta camada é constituída com um número de neurónios igual aos valores dos resultados desejados. É uma camada que não contém função de ativação e os neurónios presentes nela são exclusivamente neurónios de *output*.

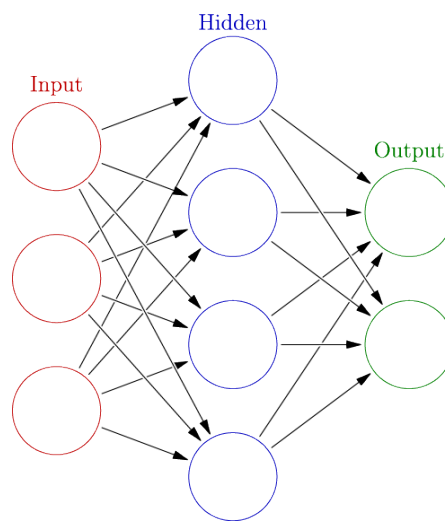


Figura 5: Representação Abstrata da Rede Neural

2.2 Trabalho Efetuado

2.3 Resultados esperados e forma de avaliação

3 Conclusões

4 Recursos