



Algoritmos de Aprendizagem e Redes Neurais

Aplicação de ID3 / C4.5 à predição da popularidade de notícias

Relatório Final

Inteligência Artificial

3º ano do Mestrado Integrado em Engenharia Informática e Computação

Elementos

Duarte Alexandre Pinto Brandão | 201007823

Paulo Sérgio Vieira da Costa | 201206045

Pedro Miguel Santos Ferreira | 201103084

21 de Maio de 2017

Índice

1 - Objetivo	3
2 - Especificação.	3
2.1 - Descrição do Problema.	3
2.2 - Abordagem	4
2.3 - Algoritmo ID3	4
2.4 - Algoritmo C4.5	5
2.5 - Estratégia de Resolução	5
2.6 - Função de Avaliação	6
2.7 - Dataset	7
3 - Desenvolvimento.	8
3.1 - Implementação	8
3.2 - Representação de Conhecimento	8
3.3 - Ambiente de Desenvolvimento	8
3.4 - Estrutura do Programa	9
4 - Conclusões.	9
5 - Recursos.	9
5.1 - Bibliografia	9
5.2 - Software	10
5.3 - Elementos do Grupo	10

1. Objetivo

Este projeto teve como objetivo principal a implementação duma plataforma inteligente, capaz de aprender - ou derivar - as regras de classificação para um domínio em análise, a partir de um conjunto de exemplos. No âmbito da sua resolução, também foi esperada a determinação da árvore de decisão que traduz as regras na predição de popularidade de notícias, através de um dataset dado.

2. Especificação

2.1 Descrição do Problema

No processo de aprendizagem das regras de classificação para os elementos que constituem a população do domínio em análise, foram utilizados um conjunto de exemplos, que representam a nossa amostra. Quanto aos parâmetros das regras de classificação, a derivar após o processo de aprendizagem, estes são constituídos pelos elementos do domínio que por sua vez são definidos por um conjunto de atributos. A identificação das variáveis que melhor definem o domínio em análise constitui um passo preponderante para a resolução do problema em questão.

De seguida, o programa é capaz de determinar uma Árvore de Decisão responsável pela tradução das regras de classificação, aplicando o algoritmo C4.5 com base no grupo de dados disponibilizado em <http://archive.ics.uci.edu/ml/datasets>.

Posteriormente, de forma a verificar a eventual necessidade de pré-processamento, o conjunto de dados é extensivamente analisado, de modo a que o modelo obtido seja possível de utilizar na classificação de novos casos que possam surgir futuramente.

Tal como se encontra referido na página da unidade curricular, deste projeto são incluídos os seguintes procedimentos:

- Implementação/aplicação do algoritmo ID3/C4.5.
- Apresentação das regras de classificação, que são retiradas da Árvore de Decisão resultante.
- Medição detalhada de resultados nos dados de treino e de teste.

2.2 Abordagem

As árvores de decisão são representações simples do conhecimento aplicadas posteriormente em sistemas de aprendizagem. São amplamente utilizadas em algoritmos de classificação, como um meio eficiente para construir classificadores capazes de prever classes baseadas nos valores de atributos.

Estas, por sua vez, são constituídas por nós que representam os atributos, de ramos, provenientes destes nós e que recebem os valores possíveis para estes atributos, e de nós folha, que representam as diferentes classes de um conjunto.

Sendo que o problema proposto envolve a predição da popularidade (número de partilhas) de notícias, para efetuar o cálculo desse valor foi utilizado o algoritmo C4.5, que se encontra especificado mais a frente. Para isto, foi utilizado um dataset de várias notícias, com a respectiva classificação de acordo com os atributos definidos inicialmente, de modo a treinar o agente de forma a que aprenda a prever a popularidade de notícias futuras, usando árvores de decisão geradas pelo algoritmo C4.5.

2.3 Algoritmo ID3

O algoritmos ID3, tanto como o C4.5 foram introduzidos para indução de modelos de classificação, mais conhecidos por árvores de decisão. O algoritmo ID3 foi um dos primeiros algoritmos de árvore de decisão, tendo sua elaboração baseada em sistemas de inferência e em conceitos de sistemas de aprendizagem. Este constrói árvores de decisão a partir de um dado conjunto de exemplos, sendo a árvore resultante usada para classificar amostras futuras. O ID3 separa um conjunto de treino em subconjuntos, de forma a que estes contenham exemplos de uma única classe. A divisão é efetuada através de um único atributo, que é selecionado a partir de uma propriedade estatística, denominada ganho de informação, que mede o quão informativo é um dado atributo.

Após a construção de uma árvores de decisão é importante avaliá-la. Esta avaliação é realizada através da utilização de dados que não tenham sido usados na simulação. Esta estratégia permite estimar como a árvore generaliza os dados e adapta Árvores de decisão a novas situações, podendo também estimar a proporção de erros e acertos ocorridos na construção da árvore.

2.4 Algoritmo C4.5

O algoritmo C4.5 é um aprimoramento do algoritmo ID3, isto devido ao facto de trabalhar com valores indisponíveis, com valores contínuos, podar árvores de decisão e derivar regras. Trabalhar com registos que possuem valores indisponíveis na construção de uma árvore da decisão, pode ser considerado um problema. A falta destes valores, pode ocorrer pelo facto de não terem sido registados no momento de coleção dos dados, ou por não serem considerados relevantes para um determinado caso.

Os registos que possuem valores desconhecidos podem ser simplesmente descartados do conjunto de simulação, ou podem ser classificados pela estimativa da probabilidade dos vários valores possíveis. Este método de classificação foi melhorado, para que lidasse com atributos numéricos, valores em falta e dados com ruído, tendo surgido o algoritmo C4.5 usado neste trabalho, que é uma melhoria do ID3.

2.5 Estratégia de Resolução

O nosso agrupamento de dados resume um conjunto heterogêneo de características sobre artigos publicados pela Mashable num período de dois anos. O objectivo é prever o número de partilhas nas redes sociais (popularidade).

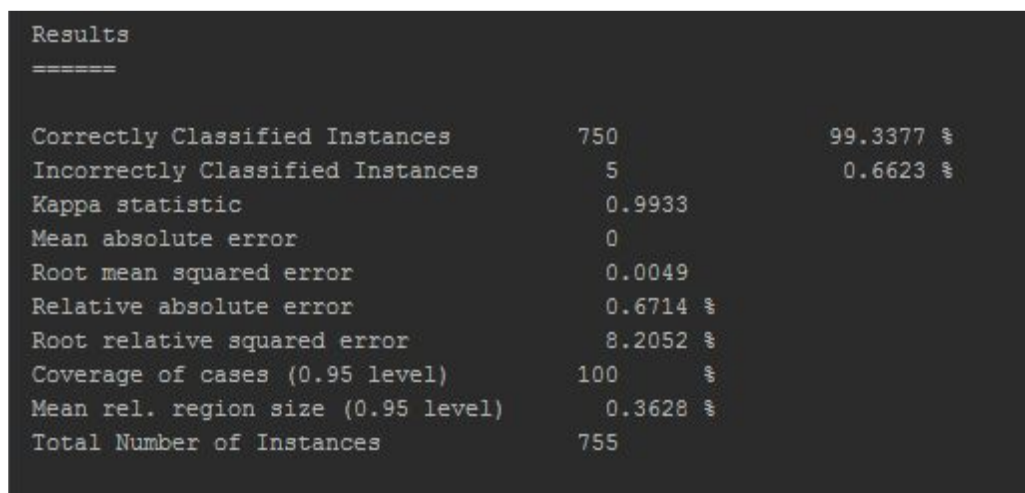
Para a resolução do nosso problema vamos usar o algoritmo C4.5. O algoritmo ID3 foi descartado à partida, devido ao nosso agrupamento de dados conter atributos numéricos, que só é possível classificar com a sua versão melhorada C4.5.

Este algoritmo retorna uma árvore de decisão, com base num agrupamento de dados. O objetivo é que esta árvore esteja na forma mínima, isto é, que as folhas estejam nos níveis mais superiores. Para isto, calcula-se a pureza dos atributos, que é o grau de certeza que um determinado atributo permite classificar uma instância. A unidade de medida da pureza é a entropia (desordem de um sistema, em que um sistema ordenado tem entropia = 0), que é menor quanto maior for a certeza que um atributo fornece à classificação. Os atributos são organizados de forma crescente de entropia na árvore de decisão.

A preparação dos dados para o processamento e para a utilização de técnicas de aprendizagem automática, consome com frequência uma porção significativa do esforço total aplicado. O primeiro passo para a utilização desses dados é o processamento prévio.

2.6 Função de Avaliação

Com o objectivo de obter a melhor eficácia e os melhores resultados com o nosso programa fizemos algumas experiências com os métodos de avaliação que tínhamos a disposição, concluindo que aquele que produzia os melhores resultados era o de Cross-Validation. Também foram efetuados testes com o método de Percentage-Split, embora os resultados obtidos tenham sido consideravelmente piores. Posteriormente decidimos fazer um teste à sua eficácia, tendo obtido os seguintes resultados:



```
Results
=====

Correctly Classified Instances      750           99.3377 %
Incorrectly Classified Instances     5           0.6623 %
Kappa statistic                    0.9933
Mean absolute error                 0
Root mean squared error             0.0049
Relative absolute error             0.6714 %
Root relative squared error         8.2052 %
Coverage of cases (0.95 level)     100          %
Mean rel. region size (0.95 level)  0.3628 %
Total Number of Instances          755
```

Correctly Classified Instances	750	99.3377 %
Incorrectly Classified Instances	5	0.6623 %
Kappa statistic	0.9933	
Mean absolute error	0	
Root mean squared error	0.0049	
Relative absolute error	0.6714	%
Root relative squared error	8.2052	%
Coverage of cases (0.95 level)	100	%
Mean rel. region size (0.95 level)	0.3628	%
Total Number of Instances	755	

figura z - Resultados de classificação utilizando o método de Cross-Validation

Com isto, foi possível concluir que a função de avaliação Cross-Validation é que retorna uma maior taxa de acertos, sendo portanto a de maior relevância para o nosso projeto.

2.7 Dataset

Os dados usados para o treino da Rede foram do site <http://archive.ics.uci.edu/ml/datasets>, onde o dataset foi posteriormente transformado no ficheiro arff, porque é uma maneira fácil e eficiente de ser representado, sendo composto por instâncias independentes e sem qualquer ordem. O ficheiro de dados encontra-se então dividido em duas partes distintas, a definição dos atributos e nome da relação, e a parte dos dados onde estes atributos tomam valores, como se pode verificar no exemplo dado em baixo.

Este ficheiro contém 39643 casos diferentes com a devida avaliação passada como último argumento. Neste caso esse valor refere-se ao número de partilhas de cada notícia.

```
@relation 'news'
@attribute n_tokens_title numeric
@attribute n_tokens_content numeric
@attribute n_unique_tokens numeric
@attribute n_non_stop_words numeric
@attribute n_non_stop_unique_tokens numeric
@attribute num_hrefs numeric
@attribute num_self_hrefs numeric
@attribute num_imgs numeric
@attribute num_videos numeric
@attribute average_token_length numeric
@attribute num_keywords numeric
@attribute data_channel_is_lifestyle numeric
@attribute data_channel is entertainment numeric
```

$$(\dots)$$

```
@data  
12,219,663.594467,99999.9992,815.3846091,4,2,1,0,4680.365297,5,0,1,0,0,0,0,0,0,0,0,0,  
0,496,496,496,1,0,0,0,0,0,0,0,500.3312041,378.2789296,40.0046751,41.26264773,40.1225435,521.  
.6171455,92.56198347,45.66210046,13.69863014,769.2307692,230.7692308,378.6363636,100,700,-3  
50,-600,-200,500,-187.5,0,187.5,593  
9,255,604.7430806,99999.9993,791.9463034,3,1,1,0,4913.72549,4,0,0,1,0,0,0,0,0,0,0,0,0,  
0,0,0,1,0,0,0,0,0,0,0,799.7556874,50.0466754,50.09625181,50.10067342,50.00071194,341.24579  
12,148.9478114,43.1372549,15.68627451,733.3333333,266.6666667,286.9146006,33.33333333,700,-  
118.75,-125,-100,0,0,500,0,711  
9,211,575.1295307,99999.9992,663.8655406,3,1,1,0,4393.364929,6,0,0,1,0,0,0,0,0,0,0,0,0,  
0,918,918,918,1,0,0,0,0,0,0,0,217.7922885,33.334457,33.35142493,33.3335358,682.1882937,702.  
2222222,323.3333333,56.87203791,9.478672986,857.1428571,142.8571429,495.8333333,100,1000,-4  
66.6666667,-800,-133.3333333,0,0,500,0,1500
```

(...)

3 Desenvolvimento

3.1 Implementação

Para a implementação, usamos a Weka API, uma ferramenta de data mining desenvolvida em Java pela Universidade de Waikato, na Nova Zelândia, que facilita a criação e treino de redes neurais através do algoritmo C4.5.

3.2 Representação de Conhecimento

As redes neurais artificiais são compostas por uma série de nós ligados entre si com objetivo de emular uma rede neuronal biológica. Os nós representam os neurónios e as arestas entre eles representam as sinapses. Nesta rede encontramos 3 tipos de nós: de entrada, de saída e nós intermédios.

Aos nós de entrada são fornecidos os dados que descrevem o objeto a ser classificado. Os nós de saída são neste caso o número de partilhas previstas para cada uma das notícias dadas. Os nós intermédios estabelecem a correspondência entre nós de entrada e de saída e é nestes que é feito o processamento dos dados.

Através do algoritmo C4.5 é feita a construção de árvores de decisão, a partir de um conjunto de dados de treino, da mesma forma que o algoritmo ID3 faria, utilizando o conceito de Entropia. O conjunto de dados de treino é um conjunto de amostras já classificadas representadas no dataset. Cada amostra consiste de um vetor p-dimensional onde é feita a representação de valores de atributos ou características da amostra, assim como a categoria à qual esta pertence.

3.3 Ambiente de Desenvolvimento

O trabalho foi desenvolvido em Windows 10, utilizando o IntelliJ IDEA, IDE de Java. Foi integrada a biblioteca de software Weka, que contém uma série de estruturas de dados e algoritmos usados para data mining, classificação de atributos implementação de redes neurais.

3.4 Estrutura do Programa

De modo a efetuar uma melhor organização do projecto foi feita a decisão de o separar em projecto módulos: módulo lógico e módulo gráfico.

O módulo lógico é responsável pela obtenção de resultados e pela classificação/previsão da popularidade de uma notícia. O programa encontra-se então dividido em duas classes: ClassifyData: É a estrutura que contém a lógica para o treino e teste de redes neuronais, através da chamada de métodos da API Weka. O treino da rede é feito posteriormente à leitura do ficheiro arff que por sua vez contém a amostra dada. As classes usadas do Weka para fornecer à rede neuronal o conjunto de dados de treino são a DataSource e Instances. A classe Interface contém a implementação da interface com o utilizador em swing e utiliza os métodos definidos na classe ClassifyData.

4 Conclusões

Através da implementação deste projeto foi-nos permitida a obtenção de conhecimentos imprescindíveis ao desenvolvimento de classificadores que utilizam o algoritmo C4.5, assim como à seleção correta de atributos para que sejam fornecidos ao algoritmo dados em qualidade e quantidade satisfatórias para a produção de uma árvore de decisão. No que toca à previsão das notícias dadas, foi possível obter ~99% de instâncias corretamente classificadas com a utilização da função de avaliação Cross-Validation . A falta de dados para notícias com atributos nos extremos do espectro, diminuem a capacidade do algoritmo de produzir previsões corretas para esses valores. Assim a eventual utilização de um dataset com ainda mais informação possibilitaria uma ainda maior taxa de acerto para os casos verificados.

5 Recursos

5.1 Bibliografia

<http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity>
<https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm>
https://fenix.tecnico.ulisboa.pt/downloadFile/3779571252917/licao_17.pdf
http://www.saedsayad.com/decision_tree.html
<http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html>
<https://cis.temple.edu/~giorgio/cis587/readings/id3-c45.html>
<https://pt.slideshare.net/aorriols/lecture5-c45>

http://saiconference.com/Downloads/SpecialIssueNo10/Paper_3-A_comparative_study_of_decision_tree_ID3_and_C4.5.pdf

5.2 Software

<http://weka.sourceforge.net/doc.stable/weka/classifiers/trees/J48.html>

5.3 Elementos do Grupo

Este projeto foi desenvolvido por Duarte Alexandre Pinto Brandão, Paulo Sérgio Vieira da Costa e Pedro Miguel Santos Ferreira, tendo sido o esforço aplicado na sua implementação distribuído de forma equitativa pelos três elementos.