Processamento de Análises de Doenças Euritermo - Escamosas

David Parreira
Universidade de Évora
l33257@alunos.uevora.pt

Luís Rato
Universidade de Évora
lmr@di.uevora.pt

Resumo

O objetivo deste trabalho encontrou-se um modelo preciso que classifique cada doença dermatológica através das diversas patologias encontradas no paciente. Outro dos objetivos é também construir uma árvore de decisão fácil de interpretar por um dermatologista de forma a ter uma ferramenta de suporte na classificação da patologia do cliente.
Por fim concluímos que é impossível fazer a distinção destas doenças sem análises clínicas.

1 Introdução

As doenças Euritermo Descamativas são um grande desafio na área da dermatologia visto que todas elas partilham sintomas de eritemas e escamamentos com apenas pequenas diferenças, tornando-as difíceis de distinguir. Normalmente é necessário análises clínicas para a se diagnosticar a doença em questão sendo que uma doença pode apresentar sintomas de outra numa fase inicial.
A classificação destas doenças é um trabalho complicado visto que estas partilham sintomas consoante a fase de desenvolvimento podem apresentar comportamentos típicos de outras doenças induzindo em muitos erros de classificação.

Para enfrentar este problema vários modelos foram explorados anteriormente. Do ponto de vista de algoritmos de classificação, certos modelos como redes neurais artificias (ANN) e máquinas de vetores de suporte (SVM) são altamente precisos porém não têm grande aplicabilidade na área da dermatologia visto que não são facilmente analisáveis levando a um aparecimento de estudos que visam melhorar o desempenho de algoritmos que utilizam árvores de decisão. Criando um modelo de árvores de decisão a análise dos dados torna-se simples e rápida.

Este foi o tema escolhido pela maior parte dos estudos envolvendo este conjunto de dados sendo que o mais importante foi efectuado por Demiroz, Govenir e Ilter em Artificial Intelligence in Medicine (1998).

Apresentação dos Dados

Neste grupo de dados possuímos trinta e quatro atributos diferentes por paciente, nos quais doze são características encontradas do paciente à vista humana, vinte e duas são características histopatológicas resultantes da análise da pele do paciente ao microscópio.

O conjunto de atributos do paciente encontram-se avaliados de 0-3 sendo os casos 0 para ausência da patologia, 1 e 2 para casos intermédios no qual existe uma maior presença no caso 2 do que no 1, 3 quando se encontra a maior quantidade possível desta patologia. Existem também dois atributos numéricos distintos que são a idade no qual o valor representa a idade do paciente e o histórico familiar. Este representado de 0-1 representa a ausência ou a presença de qualquer uma das doenças de pele em causa em algum familiar.

Processamento

Algoritmo

Visto que o objetivo de um conjunto de dados como este é a criação de um modelo de classificação fiável e de fácil interpretação o algoritmo escolhido é naturalmente um algoritmo de decision tree que neste caso foi o algoritmo J48.

Procedimento

1. Primeiro passo do procedimento foi transformar os dados de forma a serem utilizáveis pelo algoritmo escolhido. O algoritmo J48 não permite fazer a classificação de numerais, como tal o primeiro passo foi aplicar pré-processamento ao atributo que pretendemos classificar. O filtro escolhido foi NumericToNominal.

2. Neste passo a minha efetuei uma pesquisa sobre as opções de teste do Weka. Após a pesquisa concluí que a melhor opção de teste é cross-validation visto que este método apesar de ser mais robusto e efetuar testes várias vezes permite obter valores mais exatos sobre a precisão do nosso modelo.

3. Após a escolha do tipo de teste iniciei a tentativa de melhoria do algoritmo. Após um primeiro teste sem alterações no algoritmo verifiquei que a árvore obtida era binária logo ativar ou não opção binarySplits do J48 torna-se inútil visto que a árvore já é binária. Por fim acabei a obter melhorias no resultado por ativar a opção unpruned.

Resultados

Após o pré-processamento dos dados e a escolha do tipo de teste efetuei o primeiro teste ao meu conjunto de dados. A Figura 1 apresenta os dados obtidos através do Weka.

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         709               96.8579 %
Incorrectly Classified Instances        23                3.1421 %
Kappa statistic                          0.9607
Mean absolute error                      0.0113
Root mean squared error                  0.0971
Relative absolute error                  4.2325 %
Root relative squared error             26.6044 %
Total Number of Instances              732

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.996    0.002    0.996      0.996    0.996      0.994    1.000     1.000     1
                 0.934    0.010    0.950      0.934    0.942      0.931    0.995     0.972     2
                 0.958    0.009    0.965      0.958    0.962      0.952    0.997     0.984     3
                 0.918    0.011    0.928      0.918    0.923      0.911    0.993     0.961     4
                 1.000    0.003    0.981      1.000    0.990      0.989    1.000     0.996     5
                 1.000    0.003    0.952      1.000    0.976      0.974    1.000     0.990     6
Weighted Avg.    0.969    0.006    0.968      0.969    0.968      0.962    0.998     0.986
```

Figura 1. Algoritmo J48 sem melhorias

Através Figura 1 podemos observar diretamente que o algoritmo J48 é uma boa escolha para classificar os dados visto que através de indicadores de performance como F-Measure e ROC Area podemos verificar que algoritmo à partida é bastante preciso mesmo sem alterações. Neste caso quanto mais perto do valor 1 os indicadores estiverem melhor o desempenho do algoritmo.

O F-Measure é um indicador de performance que relaciona a precisão do algoritmo com a sua cobertura tornando-se assim um indicador de performance mais fácil de analisar que ambos em separado.

ROC Area é um indicador de precisão do modelo no qual quanto maior for a área abrangida pela curva maior é a precisão do modelo.

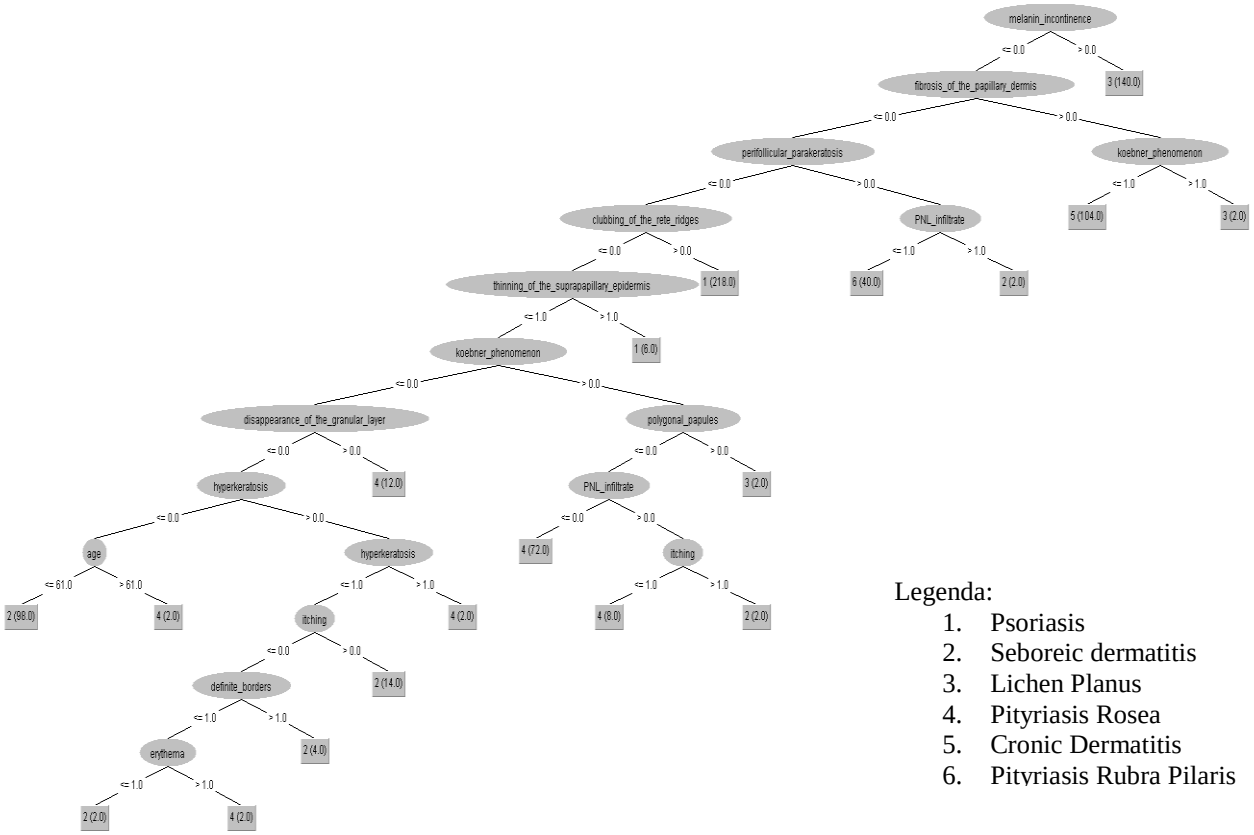
Após a alteração do parâmetro unpruned para true no Weka obtivemos o resultado apresentado na Figura 2.

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         712               97.2678 %
Incorrectly Classified Instances        20                2.7322 %
Kappa statistic                          0.9658
Mean absolute error                      0.0089
Root mean squared error                  0.0869
Relative absolute error                  3.3507 %
Root relative squared error             23.8135 %
Total Number of Instances              732

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.996     0.002     0.996      0.996     0.996      0.994     1.000     1.000     1
               0.934     0.005     0.974      0.934     0.954      0.945     0.998     0.988     2
               0.958     0.009     0.965      0.958     0.962      0.952     0.998     0.986     3
               0.949     0.011     0.930      0.949     0.939      0.930     0.998     0.984     4
               1.000     0.003     0.981      1.000     0.990      0.989     1.000     0.996     5
               1.000     0.003     0.952      1.000     0.976      0.974     1.000     0.990     6
Weighted Avg.  0.973     0.005     0.973      0.973     0.973      0.967     0.999     0.992
```

Figura 2 J48 Após melhorias

Na Figura 2 podemos observar uma melhoria geral na performance do algoritmo sendo que o valor médio dos indicadores de performance subiu. Este aumento de performance dá-se ao facto de a opção pruning remover nós ou "pedaços" da árvore cujos a sua remoção não afete drasticamente a performance do algoritmo resultando numa árvore mais simples e diminuindo possíveis erros de overfitting. Ao desligarmos esta opção, o resultado obtido é uma árvore ligeiramente maior mas ainda assim perfeitamente legível e com uma maior precisão de classificação.

O resultado final da construção deste modelo é uma árvore de classificação apresentada na Figura 3.



Legenda:

1. Psoriasis
2. Seboreic dermatitis
3. Lichen Planus
4. Pityriasis Rosea
5. Cronic Dermatitis
6. Pityriasis Rubra Pilaris

Figura 3 Árvore Modelo

Conclusão

Podemos concluir que com o modelo que foi criado um modelo fiável de classificação no qual a tarefa de distinção das doenças de pele torna-se bastante mais simples. Conseguimos também concluir que o diagnóstico destas doenças é feito obrigatoriamente por análises de laboratório visto que o nó de raiz da árvore é uma patologia que só pode ser identificada através de análise ao microscópio. Porém com as respetivas análises facilmente um dermatologista é capaz de identificar a doença através das suas características.

Para terminar poderia efetuar-se também englobar num software de medicina um algoritmo que percorra a árvore obtida de forma a indicar a doença encontrada consoante as patologias encontradas no doente.

Referências

Dados:
https://archive.ics.uci.edu/ml/datasets/Dermatology

Paper mais relevante:
https://www.ncbi.nlm.nih.gov/pubmed/9698151

Papers que referenciam este conjunto de dados:
https://www.ncbi.nlm.nih.gov/pubmed/9698151

Informação sobre o algoritmo utilizado:
https://weka.wikispaces.com/J48-Weighter+patch

Informação sobre as doenças referenciadas neste artigo

Psoriasis:
https://www.mayoclinic.org/diseases-conditions/psoriasis/symptoms-causes/syc-20355840

Seboreic Dermatitis:
https://www.mayoclinic.org/diseases-conditions/seborrheic-dermatitis/symptoms-causes/syc-20352710

Lichen Planus:
https://www.mayoclinic.org/diseases-conditions/lichen-planus/symptoms-causes/syc-20351378

Pityriasis Rosea:
https://www.mayoclinic.org/diseases-conditions/pityriasis-rosea/symptoms-causes/syc-20376405

Cronic Dermatitis:
https://www.mayoclinic.org/diseases-conditions/atopic-dermatitis-eczema/symptoms-causes/syc-20353273

Pityriasis Rubra Pilaris:
https://reference.medscape.com/article/1107742-overview