Meta-aprendizagem
Extracção de Conhecimento de Dados II
Mestrado em Inteligência Artificial e Sistemas Inteligentes

António Jesus Monteiro de Castro
Aluno Nº 040594004*

Daria Barteneva
Aluno Nº 040597004†

*frisky.antonio@gmail.com
†dasha@cnotinfor.pt

Faculdade de Engenharia da Universidade do Porto

16 de Maio de 2005

Resumo

Para fazermos este trabalho escolhemos um problema de classificação com o erro como medida de desempenho. Também escolhemos os algoritmos àrvores de decisão e discriminate linear, para podermos aplicar aos problemas escolhidos e, assim, analisar a medida de desempenho de cada um. Com base na informação obtida, identificamos as propriedades dos problemas que afectam a eficácia dos algoritmos seleccionados e definimos uma medida que permite aferir uma nova instância. Em R implementamos a medida definida anteriormente para que, dada uma nova instância, possamos obter informação acerca desse novo problema somente com base nessa medida. Pela nossa análise podemos concluir que as propriedades por nós escolhidas e a respectiva medida implementada, permitem obter resultados correctos.

Palavras-chave: Meta-aprendizagem, Classificação, Árvores de Decisão, Discriminante Linear.

1 Tipo de Problema, *DataSets* e Algoritmos

O tipo de problema escolhido por nós é de **Classificação** e a medida de desempenho será o **Erro**.
Os algoritmos escolhidos para aplicar ao problema são:
Árvores de Decisão e **Discriminante Linear**.

Escolhemos os seguintes problemas (*DataSet's*) retirados do UCI:
Aids2, *Melanoma*, *Painters*, *Rabbit*, *Oats*, *Pendigit*, *Housing*, *Heart Disease*, *German Credit Database*, *Contraceptive Method Choice*, *Iris Plant Database*.
Em anexo a este relatório seguem todos os *DataSet's* em formato CSV para que possam ser utilizados juntamente com o código.

2 Aplicação dos Algoritmos e Identificação das Propriedades

No código em R enviado em anexo é possível encontrar a implementação dos algoritmos utilizados, incluindo a implementação do *Cross Validation* que foi o método utilizado para gerarmos a medida de desempenho escolhida por nós, ou seja, o erro. A tabela 1 foi preenchida com os dados gerados através da execução da função **ml()** do código em anexo. De salientar que fizemos duas avaliações e a tabela 1 reflecte isso mesmo. Na tabela 2 apresentamos as propriedades

Dataset	AD(1)	DL(1)	AD(2)	DL(2)
Aids2	0.9862676	**0.9774648**	0.9848592	**0.9746479**
Melanoma	0.295	**0.28**	0.335	**0.285**
Painters	0.6730769	**0.6346154**	0.7307692	**0.6538462**
Rabbit	**0.3**	0.6833333	**0.2833333**	0.7166667
Oats	**0.9166667**	0.9722222	**0.9027778**	0.9583333
Pendigit	0.1827116	**0.1245678**	0.1863512	**0.1248408**
German	0.27	**0.243**	0.269	**0.246**
Contraceptive	**0.4571429**	0.492517	**0.4510204**	0.4857143
Heart	0.1814815	**0.1555556**	0.2185185	**0.1518519**
Iris	0.05333333	**0.02**	0.08	**0.02**
Housing	**0.9305556**	0.9444444	**0.9285714**	1

Tabela 1: Resultados da aplicação dos algoritmos e respectivos erros

dos problemas já tendo em vista a medida que implementamos e cuja análise apresentamos na secção seguinte, incluindo os refinamentos que fomos fazendo de forma a melhorar a medida por nós encontrada e tida como a ideal. A tabela 2 foi preenchida com os dados obtidos pela função **avaliacao datasets()** que enviamos em anexo no código R. Na tabela 2 a coluna **Larg** contém o número de variáveis, **Tam** o número de observações, **Nom** o número de variáveis nominais, **Num** o número de variáveis numéricas, **AmpCl** a quantidade de valores possíveis que a classe pode ter, **Amp/Tam** é o valor da AmpDS sobre o Tamanho do problema e, finalmente, **AmpDS** é a soma dos valores distintos possíveis para cada variável do *DataSet*. Esta última é calculada pela formúla *length(summary(as.factor(dataset[,i])))* e, dada a sua importância, convém explicar como chegamos a este valor:
Para cada *Dataset* criamos um vector em que cada posição corresponde a uma coluna desse problema (variável). Cada posição do vector contém um valor numérico correspondente ao número de valores possíveis que essa variável possa tomar. Por exemplo, [34 4 4 15 2 2 4 4 2 3]. Isto corresponde a um problema com 10 variáveis em que a primeira tem a possibilidade de ter 34 valores distintos, a segunda 4 e assim sucessivamente. A soma deste vector corresponde

Dataset	Larg	Tam	Nom	Num	AmpCl	AmpDS	Amp/Tam
Aids2	7	2843	4	3	74	290	0.102004924
Melanoma	7	205		7	2	254	1.23902439
Painters	5	54	1	4	8	66	1.22222222
Rabbit	5	60	3	2	5	61	1.016666667
Oats	4	72	3	1	51	64	0.888888889
Pendigit	17	10992		17	10	1604	0.1459224309
German	21	1000	13	8	2	256	0.256
Contraceptive	10	1473		10	3	74	0.05023761
Heart	14	270		14	2	342	1.266666667
Iris	5	150	1	4	3	126	0.84
Housing	5	72	4	1	12	51	0.708333333

Tabela 2: Características dos problemas

aquilo que nós, à falta de melhor, chamamos de Amplitude do *Dataset*.

3 Análise dos Resultados e Explicação da Medida Utilizada

O dataset *Housing* mostrou que é bem avaliado tanto pelas árvores de decisão como pelo discriminante linear. Se analisarmos os *dataset's* com a menor amplitude, o único que não se adapta à regra encontrada é o *Painter*, o que permite concluir que para os *Dataset's* experimentados tivemos uma taxa de acerto de 4/5(0.8) (dos 5 *dataset's* com amplitude inferior a 100, 4 avaliam-se melhor pelas árvores de decisão). Para melhorarmos a regra de escolha de método analisamos a relação entre os atributos nominais e numéricos chegando à conclusão de que a maioria dos *dataset's* com a melhor classificação pelas árvores de decisão têm um número de atributos nominais superior ao de atributos numéricos, embora esta regra se aplicasse também aos *Aids2* e *German*. Neste caso a taxa de acerto será de 3/5(0.6). Fazendo uma intersecção entre as regras de amplitude e de atributos nominais iríamos cobrir apenas 3 *dataset's*, perdendo o *Contraceptive* (que não tem atributos nominais), ou seja, a taxa de acerto seria de 1 no caso de avaliação de *dataset's* como *Rabbit*, *Oats* e *Housing*, mas iríamos errar quando avaliássemos o *Contraceptive*, sugerindo para ele o discriminante linear. Por outro lado reparamos que os *dataset's* com uma aplitude muito baixa em relação ao número de observações, tal como o *Contraceptive* são melhor classificados pelas árvores de decisão. Sendo assim, vamos criar mais uma regra: os *dataset's* cuja relação amplitude/tamanho é inferior a 0.1 e amplitude < 100 são melhor classificados pelas árvores de decisão.

Para este trabalho decidimos implementar a medida usando duas regras: (1) A medida amplitude e a proporção de atributos nominais em relação aos numéricos, (2) A relação de amplitude sobre tamanho do *dataset*.
No código R enviado em anexo, poderá executar a função **sugere.algoritmo(dataset)** para obter o método sugerido para o *Dataset* indicado. Para avaliar vários *Datasets* obtendo a lista com métodos sugeridos, deverá executar a função **rank.datasets()**.

4 Ficheiros Anexos ao Relatório

- Metalearning.R (Ficheiro com o código R)
- Australian.csv
- Contraceptive.csv
- German.csv
- Heart.csv
- Pendigit.csv