Gustavo Cordeiro Libel
Igor Pardal Latini

Aprendizado Simbólico com ID3

1) Qual conhecimento o ID3 proporcionou que era desconhecido antes de sua execução?

O ID3 proporcionou o conhecimento da relação entre o padrão de cores das frutas e a energia fornecida por ela.

2) Construa um arquivo .arff a partir do dataset fornecido pelo professor. Copie aqui o cabeçalho do arquivo .arff utilizado para treinamento no WEKA (definição dos atributos e da classe de saída).

- Dataset de treinamento:

```
@relation frutaEnergia.symbolic.training

@attribute c0={K,W}
@attribute c1={R,G,B}
@attribute c2={R,G,B}
@attribute c3={R,G,B}
@attribute c4={K,W}
@attribute e={0,2,4}
```

- Dataset de teste:

```
@relation frutaEnergia.symbolic

@attribute c0={K,W}
@attribute c1={R,G,B}
@attribute c2={R,G,B}
@attribute c3={R,G,B}
@attribute c4={K,W}
@attribute e={0,2,4}
```

3) Qual o tamanho do arquivo de treinamento (quantas instâncias)?

O arquivo de treinamento tem 900 instâncias.

4) Qual o número de instâncias por classe?

C0 (K=453, W=447)
C1 (R=295, G=298, B=307)
C2 (R=297, G=307, B=306)
C3 (R=299, G=263, B=338)
C4 (K=466, W=434)
e (0=276, 2=512, 4=112)

5) Qual o valor de entropia para o dataset datasetFrutasEnergia-training.arff em relação aos valores possíveis para a classe de saída E={0,2,4}? Qual a interpretação que você dá ao valor obtido?

A entropia pode ser calculada por:

$$E[D] = \sum_{i=1}^{|C|} -p(c_i).\log_2 p(c_i)$$

D = dataset
C = conjunto de classes de saída
$\mathbf{c_i} = c_i \in C$ (*uma das classes de saída*)
p (c_i) = probabilidade de c_i em D

A probabilidade de cada classe de saída é:
Energia = 0 → 276/900
Energia = 2 → 512/900
Energia = 4 → 112/900

Com isso, a entropia do dataset é 1,36 BIT. É um valor alto de entropia (visto que o máximo para três classes é 1,585). Isso indica que os dados estão bem distribuídos (há grande desordem, incerteza).

6) Qual foi a árvore de decisão gerada pelo algoritmo? Copie e cole aqui.

```
c1= = R
|  c3= = R
|  |  c2= = R: 4
|  |  c2= = G: 2
|  |  c2= = B: 2
|  c3= = G
|  |  c2= = R: 2
|  |  c2= = G: 2
|  |  c2= = B: 0
|  c3= = B
|  |  c2= = R: 2
|  |  c2= = G: 4
|  |  c2= = B: 2
c1= = G
|  c0= = K
|  |  c3= = R
|  |  |  c2= = R: 2
|  |  |  c2= = G: 2
|  |  |  c2= = B: 0
|  |  c3= = G
|  |  |  c2= = R: 2
|  |  |  c2= = G: 4
|  |  |  c2= = B: 2
|  |  c3= = B
|  |  |  c2= = R: 0
|  |  |  c2= = G: 2
|  |  |  c2= = B: 2
|  c0= = W: 0
c1= = B
|  c2= = R
|  |  c3= = R: 2
|  |  c3= = G: 0
|  |  c3= = B: 2
|  c2= = G
|  |  c3= = R: 0
|  |  c3= = G: 2
|  |  c3= = B: 2
|  c2= = B
|  |  c3= = R: 2
|  |  c3= = G: 2
|  |  c3= = B: 4
```

7) Todos os atributos do arquivo .arff foram utilizados pelo ID3 na geração da árvore de decisão? Caso não, quais ficaram de fora?

Não. O atributo c4 (cor 4) não foi usado na geração da árvore de decisão.

8) Para o ramo c1 = B e c2 = R, explique, por meio de cálculo de entropia, o porquê de o ID3 ter escolhido o atributo c3 como sendo o próximo em vez do c4.

Utilizando a fórmula:

$$E[P] = \sum_{i=1}^{n} \frac{|C_i|}{|C|} * E[C_i]$$

onde C é o conjunto de todos os exemplos e C_i é uma partição de C

Obteve-se os seguintes resultados:

			21	37	0	8	30	0	0	30	0	29	0	0	0	37	0
			K&0	K&2	K&4	W&0	W&2	W&4	R&0	R&2	R&4	G&0	G&2	G&4	B&0	B&2	B&4
	(C1=B, C2=R)								(C1=B, C2=R)								
	C4			0	2	4			C3			0	2	4			
	K	58		0.362069	0.637931	0	1		R	30		0	1	0			
	W	38		0.210526	0.789474	0	1		G	29		1	0	0			
		96							B	37		0	1	0			
										96							
	E[K]	0.570567							E[R]	0							
	E[W]	0.293901							E[G]	0							
									E[B]	0							
	E[C4]	0.461054															
									E[C3]	0							

A entropia calculada para c3 foi 0, enquanto que para c4 foi 0,461054.

Assim, como a entropia usando o atributo c3 é menor (menos incerteza) do que usando o atributo c4, então o ID3 escolheu c3 para ser o próximo.

9) Para o ramo c1 = B, c2 = R e c3 = R, explique porque o ID3 não necessitou incluir mais atributos no ramo.

Como pode ser visto na tabela da questão anterior, c3 = R já classifica todas as instâncias (todas que tem c1 = B, c2 =R e c3 = R tem energia igual a 2), ou seja, a entropia é igual a 0. Assim, não é necessário incluir mais atributos no ramo.

10) Defina o arquivo datasetFrutasEnergia-test.arff (opção supplied test set) como sendo de teste para o modelo aprendido anteriormente. Analise o desempenho do modelo para os exemplos contidos em datasetFrutasEnergia-test.arff com base nas medidas abaixo, explicando o significado e contextualizando-as para a tarefa em questão:

a) matriz de confusão: A matriz de confusão é uma tabela que mostra o número de exemplos classificados correta/incorretamente por classe.

```
=== Confusion Matrix ===
  a     b     c   <-- classified as
 13     0     2 |  a = 0
  3    25     3 |  b = 2
  0     1     3 |  c = 4
```

Através da tabela, percebe-se que houve dois casos de frutas de energia = 0 ser classificada como energia = 4, 3 casos de frutas de energia = 2 ser classificada como energia = 0, 3 casos de frutas de energia =2 serem classificadas como energia = 4 e 1 caso de uma fruta de energia = 4 ser classificada como energia = 2.

b) para cada classe

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,867	0,086	0,813	0,867	0,839	0,767	0,890	0,744	0
0,806	0,053	0,962	0,806	0,877	0,732	0,877	0,895	2
0,750	0,109	0,375	0,750	0,500	0,475	0,821	0,301	4
0,820	0,067	0,870	0,820	0,835	0,722	0,876	0,803	

A última linha é "Weighted Avg."

i) TP rate: "True positives", aqueles exemplos classificados corretamente, ou seja, classificados na categoria que realmente pertencem.

ii) FP rate: "False positives", aqueles exemplos que, para uma determinada classe, foram classificados erroneamente como ela.

iii) precision: Dada uma classe, é o número de exemplos corretamente classificados sobre o total de exemplos classificados como sendo daquela classe. Assim, é a relação entre o número de "true positives" e a soma entre "true positives" e "false positives" para determinada classe.

iv) recall: Dada uma classe, é o número de exemplos corretamente classificados sobre o total de exemplos existentes no arquivo de entrada para aquela classe. Assim, é a relação entre o número de "true positives" e a soma entre "true positives" e "false negatives" para determinada classe.

v) f-measure: média harmônica entre precision e recall. Serve para permitir uma comparação mais direta de diferentes classificadores através de uma única medida. Assim, é igual a duas vezes a precision vezes o recall, divididos por precision vezes recall.

11) Baseando-se nos resultados acima, qual(is) medida(s) indica(m) a probabilidade de o personagem morrer por engano (comer uma furta venenosa por engano) ao utilizar o modelo aprendido? Explique.

A medida de Recall da classe energia = 0 pode ser usada para indicar a probabilidade de o personagem morrer por engano. A probabilidade seria 1-Recall (da classe energia = 0), pois esse valor indica a porcentagem de frutas com energia = 0 classificadas de forma errada.