AVALIAÇÃO DE ARREDONDAMENTO DE VALORES DE ATRIBUTOS CONTÍNUOS NA INDUÇÃO DE ÁRVORES DE DECISÃO

Lemos, R.N.[1,2] & Baranauskas, J.A.[1]
rnlemos@fmrp.usp.br, augusto@ffclrp.usp.br

Universidade de São Paulo

[1]Faculdade de Filosofia Ciências e Letras de Ribeirão Preto

[2]Faculdade de Medicina e Ribeirão Preto

INTRODUÇÃO

A maior parte das operações para construir uma árvore de decisão cresce linearmente com o número de exemplos de treinamento. Entretanto, o processo de escolha de um atributo contínuo contendo *d* valores distintos requer a ordenação desses valores, crescendo como $d \times \log_2(d)$. Assim, o tempo requerido para construir uma árvore de decisão a partir de um conjunto de treinamento grande pode ser dominado pela ordenação do atributos contínuos (Quinlan, 1993). Assim, pretende-se avaliar neste projeto as conseqüências do arredondamento de valores de atributos contínuos no processo de indução de árvores de decisão.

METODOLOGIA EXPERIMENTAL

Inicialmente foi selecionado o conjunto de exemplos sonar. O problema consiste em discriminar entre sinais de sonar que representam um cilindro de metal daqueles que representam uma rocha ligeiramente cilíndrica. Em resumo, o conjunto contém 60 atributos contínuos, 2 classes, 208 exemplos, com erro majoritário igual a 53,36%.

Todo o conjunto de exemplos original foi submetido ao indutor J48 (Witten & Frank, 1999), obtendo-se uma árvore de decisão. Com base nisso, foram anotados os atributos que apareceram no classificador induzido e, utilizando arredondamento, foram gerados 6 conjuntos derivados do conjunto original, da seguinte forma: 3 dos conjuntos *sonar-t* (*sonar-t-1*, *sonar-t-2*, *sonar-t-3*) tiveram seus valores arredondados (para 1, 2 e 3 casas decimais, respectivamente) para todos os atributos e os 3 conjuntos restantes *sonar-p* (*sonar-p-1*, *sonar-p-2*, *sonar-p-3*) tiveram seus valores arredondados (para 1, 2 e 3 casas decimais, respectivamente) somente para aqueles atributos que apareceram no classificador J48.

Para avaliar o desempenho foi utilizado 10-*fold stratified cross-validation* tanto conjunto original de exemplos (sem arredondamento) como nos 6 conjuntos derivados, obtendo-se média e desvio padrão para tempo de indução, taxa de erro e tamanho do classificador.

RESULTADOS

Na Tabela 1 são mostrados os resultados (média ± desvio padrão) obtidos. Em média, o tempo de indução diminuiu de 0,22 (*sonar*), para 0,16 (*sonar-p*) e 0,13 (*sonar-t*). Isso significa uma redução de 27,27% e 40,91% do tempo de indução, respectivamente.

Conjunto	Tempo de Indução	Taxa de Erro	Tamanho do Classificador
sonar	0,22 ± 0,11	28,83 ± 2,24	29,20 ± 3,58
sonar-t-1	0,11 ± 0,05	23,98 ± 3,65	35,60 ± 3,66
sonar-t-2	0,13 ± 0,02	25,98 ± 3,13	32,40 ± 2,50
sonar-t-3	0,16 ± 0,02	27,40 ± 2,48	29,60 ± 4,12
sonar-p-1	0,16 ± 0,01	23,02 ± 3,62	30,20 ± 2,70
sonar-p-2	0,16 ± 0,01	27,88 ± 2,43	29,40 ± 3,24
sonar-p-3	0,16 ± 0,02	29,95 ± 2,26	29,20 ± 3,82

Tabela 1

Na Figura 1 é mostrada a diferença absoluta em desvios padrões do tempo de indução, ou seja, entre *sonar* e *sonar-p-1*, entre *sonar* e *sonar-p-2* e assim por diante. Quando a barra encontra-se acima de zero significa que o respectivo conjunto cujos valores foram arredondados supera o desempenho do conjunto original (sem arredondamento); se a barra encontra-se abaixo então o conjunto original supera o respectivo conjunto cujos valores foram arredondados. Quando a altura da barra estiver acima (abaixo) de dois significa que o conjunto cujos valores foram arredondados (conjunto original) supera o conjunto original (conjunto cujos valores foram arredondados) significativamente, ou seja, nível de confiança de 95%. Analogamente para taxa de erro e tamanho da árvore mostrados nas Figuras 2 e 3, respectivamente.

Como esperado, o tempo de indução reduziu para todos os conjuntos utilizando arredondamento, embora não de forma significativa (com grau de confiança de 95%).

Em média, a taxa de erro diminuiu de 28,83% (sonar), para 26,95% (sonar-p) e 25,79% (sonar-t). Isso significa uma redução de 6,52% e 10,54% da taxa de erro, respectivamente. Como pode ser observado na Figura 2, a taxa de erro reduziu para todos os conjuntos utilizando arredondamento, embora não de forma significativa.

Em média, o tamanho da árvore aumentou de 29,20 (sonar), para 29,60 (sonar-p) e 32,53 (sonar-t). Isso significa um aumento de 1,35% e 10,24% do tamanho da árvore, respectivamente. Como pode ser observado na Figura 3 o tamanho da árvore aumentou para todos os conjuntos, exceto sonar-p-3 utilizando arredondamento, embora não de forma significativa.

Figura 1

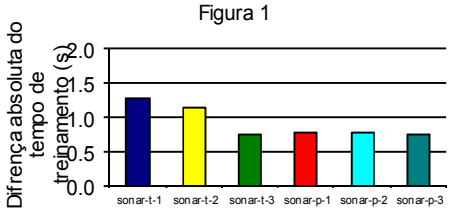


Figura 2

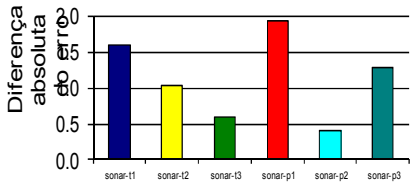
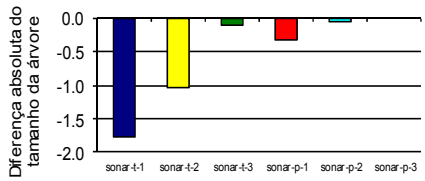


Figura 3



TRABALHOS FUTUROS

Pretende-se dar continuidade sobre o estudo de arredondamento de valores analisando o algoritmo proposto por Weiss & Indurkhya (1998)[Chapter 4].

REFERÊNCIAS

Blake, C. L. & Merz, C. J. (1998). UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/ MLRepository.html.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann. San Francisco, CA.

Weiss, S. M. & Indurkhya, N. (1998). *Predictive Data Mining: A Practical Guide*. San Francisco, CA: Morgan Kaufmann.

Witten, I. H. & Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 1999.