Árvores de decisão – algoritmos ID3 e C4.5

Simone C. Garcia*
Luis O.Alvares**

Resumo

As árvores de decisão são representações simples do conhecimento e têm sido aplicadas em sistemas de aprendizado. Elas são amplamente utilizadas em algoritmos de classificação, como um meio eficiente para construir classificadores que predizem classes baseadas nos valores de atributos. Assim, podem ser utilizadas em várias aplicações como diagnósticos médicos, análise de risco em créditos, entre outros exemplos.

As árvores de decisão consistem de nodos que representam os atributos, de arcos, provenientes destes nodos e que recebem os valores possíveis para estes atributos, e de nodos folha, que representam as diferentes classes de um conjunto de treinamento [HOL 94].

Uma árvore de decisão tem a função de particionar recursivamente um conjunto de treinamento, até que cada subconjunto obtido deste particionamento contenha casos de uma única classe. Para atingir esta meta, a técnica de árvores de decisão examina e compara a distribuição de classes durante a construção da árvore. O resultado obtido, após a construção de uma árvore de decisão, são dados organizados de maneira compacta, que são utilizados para classificar novos casos [HOL 94][BRA 99].

Os algoritmos ID3 e C4.5 foram introduzidos por Quinlan [QUI 93] [DAN 97] 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. Ele 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 treinamento em subconjuntos, de forma 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 quanto informativo é um atributo [DAN 97].

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 no treinamento. Esta estratégia permite estimar como a árvore generaliza os dados e se adapta

* carboni@atlas.ucpel.tche.br
** alvares@inf.ufrgs.br

a novas situações, podendo, também, se estimar a proporção de erros e acertos ocorridos na construção da árvore [BRA 99].

O algoritmo C4.5 é um aprimoramento do algoritmo ID3, isto devido ao fato de trabalhar com valores indisponíveis, com valores contínuos, podar árvores de decisão e derivar regras [ING 96] [CAB 97].

Trabalhar com registros que possuem valores indisponíveis na construção de uma árvore da decisão, pode ser considerado um problema. A falta destes valores, pode ocorrer pelo fato de não terem sido registrados no momento da coleta dos dados, ou por não serem considerados relevantes para um determinado caso. Os registros que possuem valores desconhecidos podem ser simplesmente descartados do conjunto de treinamento, ou podem ser classificados pela estimativa da probabilidade dos vários valores possíveis [ING 96].

O tratamento de atributos com valores contínuos envolve a consideração de todos os valores presentes no conjunto de treinamento, fazendo com que estes valores sejam ordenados de forma crescente e, após esta ordenação, seja selecionado o valor que favorecerá na redução da informação necessária. O uso de valores contínuos pode tornar-se lento se o número de valores for muito elevado, demandando grande tempo de ordenação [BRA 99]. Também se deve considerar que trabalhar com estes valores envolve um número substancial de computações [ING 96].

O resultado do particionamento recursivo, utilizado na construção de árvores de decisão, pode ser uma árvore muito complexa, de acordo com o seu conjunto de treimamento. O método de podar árvores de decisão é realizado substituindo uma subárvore por um nodo folha. Este método é realizado se uma regra de decisão estabelecer que a taxa de erro prevista na subárvore é muito grande, em relação a utilização de um único nodo folha. A substituição de partes da árvore é realizada considerando que estas não contribuem à exatidão da classificação em determinados casos, produzindo algo menos complexo e assim mais compreensível [QUI 93] [ING 96].

A partir de uma árvore de decisão é possível derivar regras. As regras são escritas considerando o trajeto do nodo raiz até uma folha da árvore. Estes dois métodos são geralmente utilizados em conjunto. Devido ao fato das árvores de decisão tenderem a crescer muito, de acordo com algumas aplicações, elas são muitas vezes substituídas pelas regras. Isto acontece em virtude das regras poderem ser facilmente modularizadas. Uma regra pode ser compreendida sem que haja a necessidade de se referenciar outras regras [ING 96].

Referências

[BRA 99] BRAZDIL, P. Construção de Modelos de Decisão a partir de Dados. Disponível por WWW em: http://www.ncc.up.pt/~pbrazdil/Ensino/ML/DecTrees.html, 1999.

[CAB 97] CABENA, P. et al. Discovering Data Mining from Concept to Implementation. Upper Saddle River, New Jersey: Prentice Hall, 1997.

[DAN 97] DANKEL, D. The ID3 Algorithm. Disponível por WWW em: http://www.cise.ufl.edu/ ~ddd/cap6635/Fall-97/Short-papers/2.htm, 1997.

[HOL 94] HOLSHEIMER, M. and SIEBES, A. Data Mining: the search for knowledge in databases. Disponível por FTP anônimo em ftp.cwi.nl no arquivo /pub/CWIreports/AA/CS-R9406.ps.Z, 1994.

[ING 96] INGARGIOLA, G. Building Classification Models: ID3 and C4.5. Disponível por WWW em: http://www.cis.temple.edu/~ingargio/cis587/readings/id3-c45.html, 1996.

[QUI 93] QUINLAN, J. R. C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann Publishers, 1993.