

APLICAÇÃO DE ALGORITMOS DE ÁRVORE DE DECISÃO SOBRE UMA BASE DE DADOS DA FLOR ÍRIS

Nome estudante: Elitânia de Oliveira Lopes, Nome orientador: Márcia Ferreira Cristaldo

¹Nome completo da Instituição – Instituto Federal de Mato Grosso do Sul- MS

heli.nio@hotmail.com , marcia.cristaldo@ifms.edu.br

Resumo

Uma árvore de decisão é um conjunto de treinamento, ela tem como entrada um conjunto de atributos que compõem um objeto ou situação, e como saída sua “decisão”, ou seja, uma previsão do valor de saída de acordo com a entrada. As decisões são tomadas pela árvore através da realização de uma sequência de testes. Cada nó interno na árvore corresponde a um teste do valor de um dos atributos, e as ramificações a partir do nó são identificadas com os possíveis valores do teste. Cada nó de folha na árvore especifica o valor a ser retornado se aquela folha for alcançada. Neste contexto, este trabalho irá realizar um estudo de caso utilizando o banco de dados Iris do software WEKA.

Palavras-chave: Árvore de decisão, banco de dados.

Introdução

Na área computacional, o processo de construção de uma árvore de decisão é conhecido como indução. A indução de árvores de decisão, tema central do presente trabalho, se trata de um importante tópico de pesquisa em reconhecimento de padrões e um exemplo do aprendizado indutivo, sendo uma das formas mais simples e, ainda assim mais bem-sucedidas, de algoritmos de aprendizagem e classificação. Ela serve como uma boa introdução à área da aprendizagem indutiva, e é de fácil implementação.

Foi realizada então uma análise, mais especificamente, do algoritmo J48, algoritmo do software de mineração de dados WEKA, que é baseado no algoritmo de árvores de decisão C4.5. O estudo realizado concentrou-se no entendimento do conceito de árvores de decisão como um todo e da análise do funcionamento do algoritmo em questão dentro do ambiente WEKA com seus respectivos testes e resultados. Analisou-se o comportamento do software WEKA com o algoritmo J48 para a escolha dos atributos e consequente indução da árvore de decisão com os referidos dados.

Metodologia e desenvolvimento

O presente trabalho tem como entrada os dados do banco de dados Iris, um famoso conjunto de dados reais que foi criado por Fisher no ano de 1936 e doado por Michael Marshall em 01/07/1988. O trabalho de Fisher é um clássico no seu campo e é referenciado com frequência. Este banco de dados é considerado uma das melhores bases de dados conhecidas encontradas na literatura de reconhecimento de padrão e é um domínio extremamente simples. O conjunto

de dados contém 3 classes de 50 instâncias cada, onde cada classe se refere a um tipo de planta da íris. Cada classe possui 4 atributos: o comprimento da sépala, a largura da sépala, o comprimento da pétala e a largura da pétala.

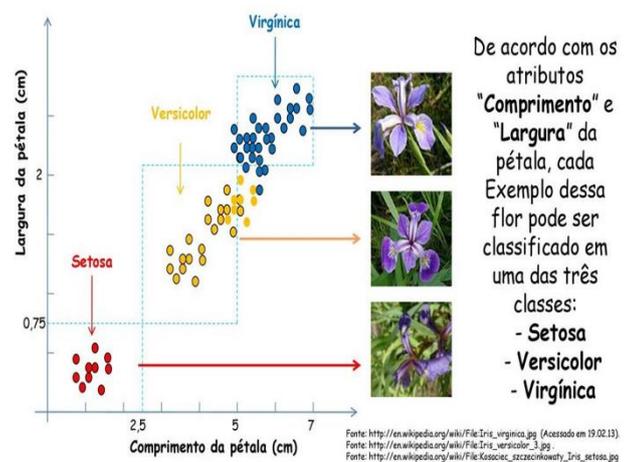


Imagem 1. Exemplo de Árvore de Decisão

Resultados e Considerações Finais

A árvore gerada possui um número de 5 folhas e um tamanho de 9 nós. É exibida na Figura 3 a árvore de decisão gerada pelo J48 com a opção de Poda. Neste trabalho serão realizados testes com base nas 150 instâncias, referentes a flor íris. Para classificarmos essas instâncias serão utilizados o algoritmo de árvore de decisão. Utilizando a base de dados do Weka, será uma forma de conhecer a bibliografia que se encontra no banco de dados, podendo assim usar esses conhecimentos adquiridos em uma nova aplicação tanto nessa área quanto em outras.

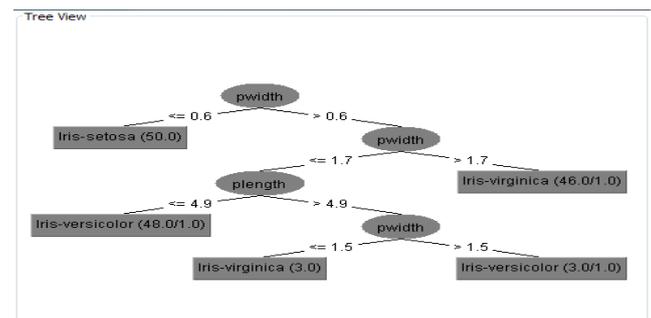


Imagem 3. Árvore de Decisão Induzida

Resultados com os algoritmos classificadores:

Posição	Classificador	Acurária	Curva ROC	Instâncias
1	NaveBayes	94.630	0,997	150
2	K-Star	97.115	0,997	150
3	RandomForest	97.115	0,998	150
4	Logist Boost	97.115	0,994	150
5	Muliclassifier /SL	96.153	0,965	150
6	MultiPerceptron	95.192	0,987	150
7	RandomComittie	92.617	0,967	150
8	J48	96.153	0,977	150
9	SMO	96.153	0,974	150
10	J-RIP	90.384	0,945	150

Tabela1.Comparação dos classificadores com Percentage Split.

Agradecimentos

A minha orientadora Márcia Ferreira Cristaldo por todo apoio e paciência ao longo da elaboração do meu projeto. Também gostaria de deixar um agradecimento especial a instituição de pesquisa Instituto Federal de Mato Grosso do Sul por possibilitar a execução deste trabalho científico.

Referências

- [1]. Data set: Iris.data. Disponível em: <<http://www.statlab.uni-heidelberg.de/data/iris/>>.
- [2]. Mitchell, Tom M. Machine Learning. McGraw-Hill, New York, 1997.
- [3]. MARIN, M. “Indução de Árvores de Decisão para a Inferência de Redes” Gênicas Mitchell, Tom M. Machine Learning. McGraw-Hill, New York, 1997.
- [4]. WAIKATO, U. O. Weka Data Mining Software in Java. Weka - The University of Waikato. Disponível em: <<http://www.cs.waikato.ac.nz/ml/weka/>>.