Inteligencia na Web e Big Data - Arvore de Decisão

Carlos Reynaldo Portocarrero Tovar
UFABC - São Paolo, Brasil

Abstract

O Arvore de Decisão e uma técnica de Classificação muito usada na Mineração de Dados, neste artigo usaremos-la pra criar um modelo de classificação pra determinar o autor de pinturas.

Keywords: Mineração de dados, Técnica, Classificador, Arvore Decisão
2017 MSC:

1. Classificação com Arvores de decisão

Conceito. Classificação e a tarefa de organizar objetos em uma entre diversas categorias pre-definidas, e um problema universal que engloba muitas aplicações diferentes.[1]

Conceito. Os arvores de Decisão são muito usados na Mineração de Dados no Problema de Classificação , são modelos estatísticos que utilizam um treinamento supervisionado para a classificação e previsão de Dados.[1] Em outras palavras, em sua construção é utilizado um conjunto de treinamento formado por entradas e saídas. Estas últimas são as classes. Arvore de Decisão e uma estrutura de fluxo parecida com uma árvore, Nós internos denotam um teste de atributo, Ramos representam um resultado do teste e Nós folhas representam rótulos de classe ou distribuição de classes.

Composição. Geração de Árvores de decisão consiste de 2 fases : Construção da árvore, Poda de Arvore. A Construção, no início, todos os exemplos de treinamento estão na raiz e Particiona exemplos recursivamente baseando-se

nos atributos selecionados. A poda e Identificar e remover ramos que refletem ruído ou outliers.

2. ¿Como construir Arvores de decisão? - Algoritmo de Hunt

Conceito. No algoritmo de Hunt una arvore de decisão cresse de forma recursiva pelo particionamento de registros de treino em sucessivos subconjuntos mais puros.

Composição. Suponhamos um conjunto D que contem os registros de treino que estam associados ao nodo t e y = *{y1,y2...yn}* sejam os rótulos das classes logo temos esta definição recursiva. Passo 1, Se todos os registros em D pertencem à mesma classe y_t então t e um nodo folha rotulado como y_t. Passo 2, Se D ,contiver registros que pertençam a mais de uma classe, uma condição de teste atributo e selecionada para particionar os registros en subconjuntos menores. Um nodo filho e criado para cada resultado da condição de teste e os registros de D_t são distribuídos para os filhos baseados nos resultados. O algoritmo e então aplicado recursivamente a cada nodo filho.[1]

3. Métricas para Selecionar a Divisão

Conceito. Entropia é o cálculo do ganho de informação baseado em uma medida utilizada na teoria da informação. A entropia caracteriza a impureza dos Dados: em um conjunto de Dados, é uma medida da falta de homogeneidade dos Dados de entrada em relação a sua classificação

$$Entropia(t) = -\sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t)$$

Conceito. Ganho de informação e um critério que pode ser usado para determinar la qualidade de uma divisão, neste caso precisamos comparar o grado de impureza do nodo pai com o grau de impureza dos nodos filhos, quanto maior a diferencia , melhor a condição do teste, a Entropia e usada para representar o grado de impureza.

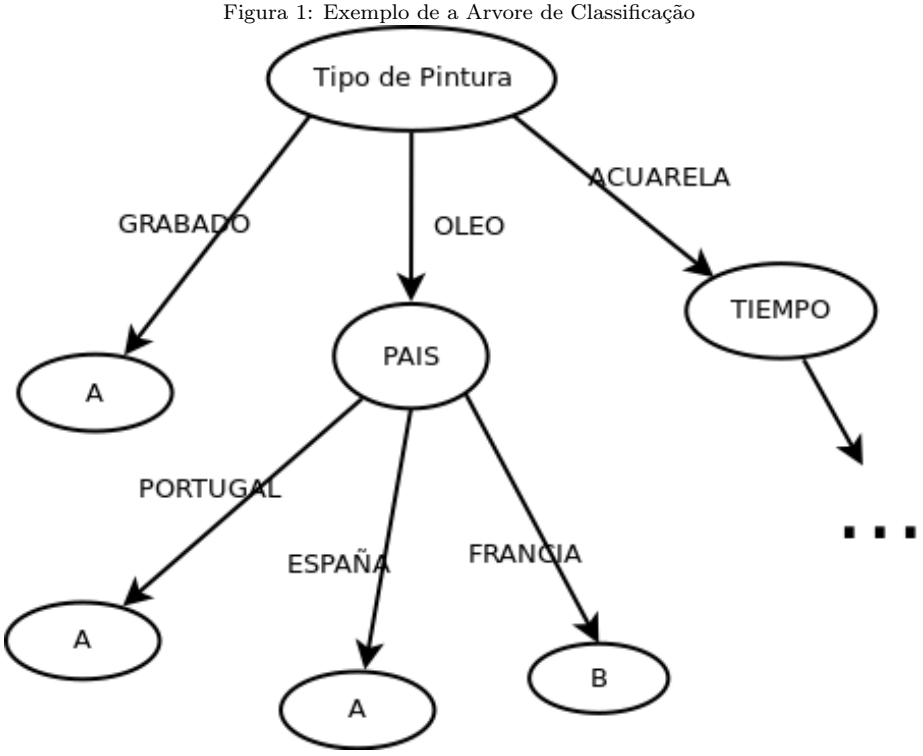
Item	Tipo	Pais	Tempo	Marco	Autor
E1	grabado	espana	moderno	si	A
E2	óleo	Portugal	moderno	no	A
E3	óleo	França	moderno	si	B
E4	óleo	espana	moderno	no	A
E5	acuarela	espana	clássico	no	A
E6	acuarela	França	clássico	si	B
E7	acuarela	espana	moderno	si	A
E8	acuarela	Portugal	clássico	si	B

Tabela 1: Tabela de Dados de Trenamento.

4. Planteamento do Problema

. Tomamos uma data de Pinturas e de seus pintores e buscamos formar um modelo de trenamento para conhecer o possível pintor de novas pinturas ,na data temos os atributos:tipo, pais, tempo, marco e a coluna autor e a classe que queremos classificar, cada atributo tem possível valores que pode tomar.

Figura 1: Exemplo de a Arvore de Classificação



5. Aplicação do Método

Descrição. O que faremos é agrupar todos os elementos na raiz da arvore, depois sacar a a entropia geral, e sacar a entropia pelo cada coluna atributo , então comparamos as entropias de cada atributo , o coluna que tem o melhor entropia e feita classificador ,como siguente passo geramos um novo no pelo cada tipo de atributo dessa coluna e repetimos os pasos pra cada no, o proceso termina quando ja no tem mais columnas o uma columna tem todos o mesma clase que se esta pesquisando.

References

[1] M. S. . V. K. Pang-Ning Tan, Introdução ao Data Mining - Mineração de Dados, 2nd Edition, Vol. 1, Softcover, 2006.