Identificação de spam utilizando Random Forest

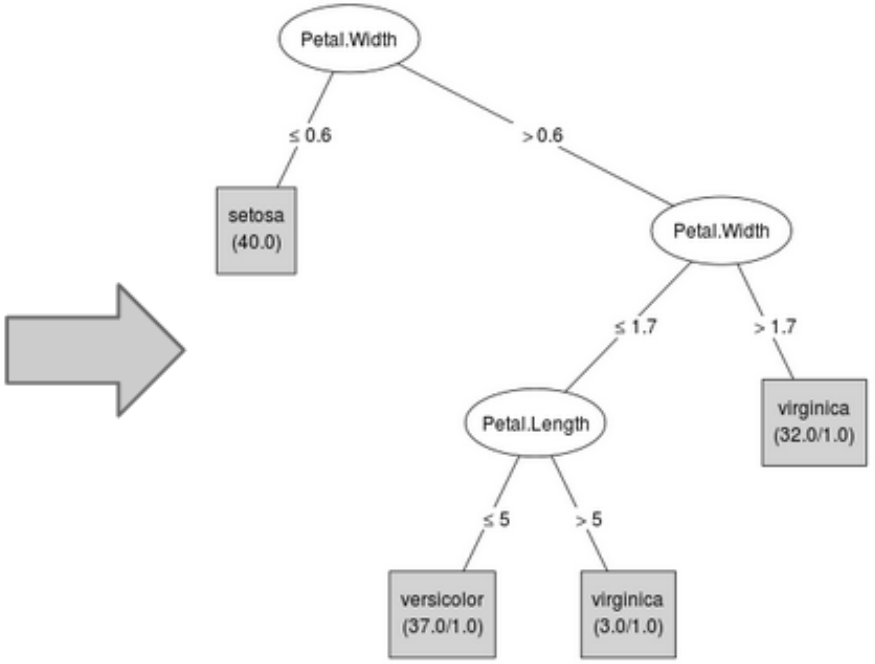
Fabrício J. Barth

Falculdade BandTec e VAGAS Tecnologia

Junho de 2013

Aprendizado de árvores de decisão

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
19	5.7	3.8	1.7	0.3	setosa
30	4.7	3.2	1.6	0.2	setosa
32	5.4	3.4	1.5	0.4	setosa
50	5.0	3.3	1.4	0.2	setosa
73	6.3	2.5	4.9	1.5	versicolor
74	6.1	2.8	4.7	1.2	versicolor
81	5.5	2.4	3.8	1.1	versicolor
89	5.6	3.0	4.1	1.3	versicolor
90	5.5	2.5	4.0	1.3	versicolor
91	5.5	2.6	4.4	1.2	versicolor
104	6.3	2.9	5.6	1.8	virginica
112	6.4	2.7	5.3	1.9	virginica
122	5.6	2.8	4.9	2.0	virginica
126	7.2	3.2	6.0	1.8	virginica
146	6.7	3.0	5.2	2.3	virginica
148	6.5	3.0	5.2	2.0	virginica



Características

- Representação de árvore de decisão:
  - ⋆ cada nodo interno testa um atributo;
  - ⋆ cada aresta correponde a um valor de atributo;
  - ⋆ cada nodo folha retorna uma classificação.

Algoritmo ID3

- O algoritmo ID3 cria uma árvore de uma maneira **top-down** começando com a seguinte pergunta:
  - ⋆ Qual atributo deve ser testado na raiz da árvore?
- Para responder esta questão, cada atributo do conjunto de treinamento é avaliado usando um teste estatístico para determinar quão bem o atributo (sozinho) classifica os exemplos de treinamento.

Entrada: Conjunto de Exemplos E.

Saída: Árvore de Decisão (Hipótese h).

1 Se todos os exemplos tem o mesmo resultado para a função sendo aprendida, retorna um nodo folha com este valor;

2 Cria um nodo de decisão N e escolhe o melhor atributo A para este nodo;

3 Para cada valor V possível para A:

3.1 cria uma aresta em N para o valor V;

3.2 cria um subconjunto E_V de exemplos onde $A = V$;

3.3 liga a aresta com o nodo que retorna da aplicação do algoritmo considerando os exemplos E_V.

4 Os passos 1, 2 e 3 são aplicados recursivamente para cada novo subconjunto de exemplos de treinamento.

Exemplo de classificação de Spam usando J48

O objetivo deste exercício é demonstrar a criação de um modelo preditivo no formato de árvore de decisão para identificar spam. Para tanto, será utilizado o dataset disponibilizado em *http://archive.ics.uci.edu/ml/datasets/Spambase*.

http://rpubs.com/fbarth/classificacaoSpamJ48

Aprendizado de florestas de árvores de decisão

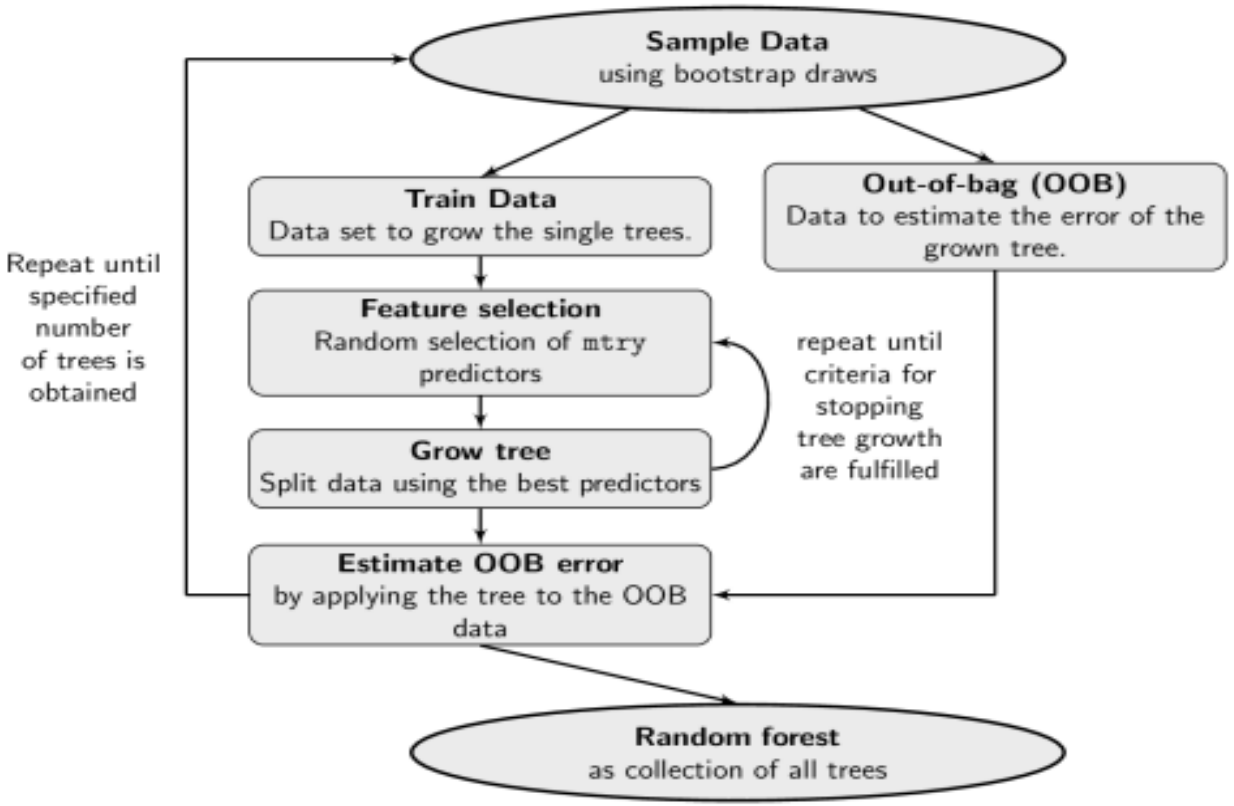


Figure 1: Random Forest Algorithm

Exemplo de classificação de Spam usando RandomForest

http://rpubs.com/fbarth/classificacaoSpamRandomForest

Material de **consulta**

- Tom Mitchell. Machine Learning, 1997. (Capítulo 3)
- Russel e Norvig. Inteligência Artificial, 2a. edição, capítulo 18.
- **Weka** no **R**: http://cran.r-project.org/web/packages/RWeka/RWeka.pdf.

- Yanchang Zhao. R and Data Mining: Examples and Case Studies. (Capítulo 4): http://cran.r-project.org/doc/contrib/Zhao_R_and_data_mining.pdf
- Exemplo de uso de algoritmos indutores de árvore de decisão. http://rpubs.com/fbarth/arvoreDecisao. Acesso em 14 de junho de 2013.

- Package 'randomForest'. http://cran.r-project.org/web/packages/randomForest/randomForest.pdf. Acessado em 14 de junho de 2013.
- Breiman, Leo (2001). "Random Forests". Machine Learning 45 (1): 5-32.
- H. Costa, F. Benevenuto, L. Merschmann. Detecting Tip Spam in Location-based Social Networks. In Proceedings of the ACM Symposium on Applied Computing (SAC'13). http://homepages.dcc.ufmg.br/ fabricio/download/sac2013.pdf