Random Forests in decisions: Abstract

Author: Craig Wright

Abstract

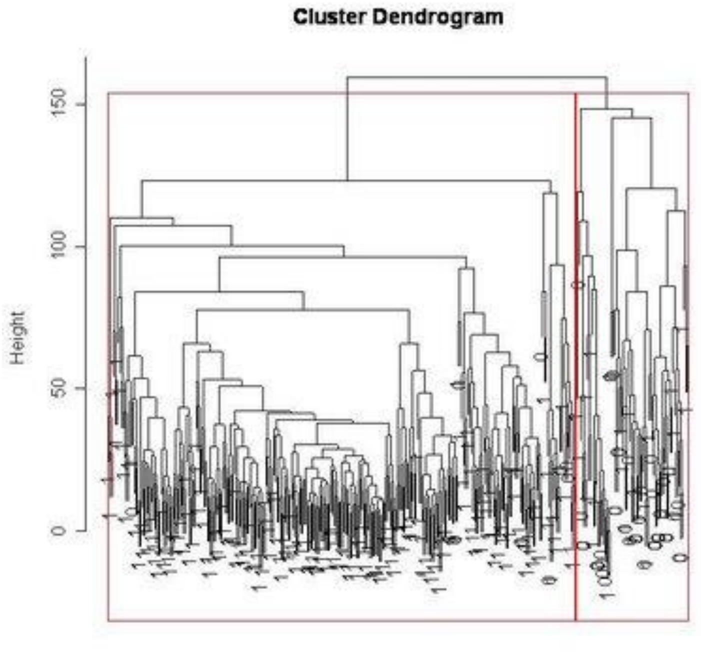
A random forest algorithm is an ensemble of unpruned decision trees. They are commonly deployed where there are extremely large training datasets and an exceedingly large quantity of input variables. In Security and risk, the dimensionality can run into the thousands of input variables. A Random Forest model generally comprises of up to hundreds of individual decision trees.

Keywords: Random Forest, Statistics, Secure Audit

Introduction

A random forest algorithm is an ensemble of unpruned decision trees. They are commonly deployed where there are extremely large training datasets and an exceedingly large quantity of input variables. In Security and risk, the dimensionality can run into the thousands of input variables. A Random Forest model generally comprises of up to hundreds of individual decision trees.

This abstract discusses upcoming research that will be presented in full on completion.



The primary benefit to risk modelling is that Random Forests tend to be very stable in model building. Their relative insensitivity to the noise that breaks down single decision tree induction models makes them compare favourably to boosting approaches while they are generally more robust against the effects of noise in the training dataset. This makes them a favourable alternative to nonlinear classifiers like artificial neural nets and support vector machines.

As the performance is frequently reliant on the individual dataset, it is a good practice to compare several approaches.

Each decision tree in the forest is constructed using a random subset of the training dataset using the techniques of bagging (replacement). Several entities will thus be included more than once in the sample, and others will be left out. This generally lies in the two thirds to one third ratios for inclusion/exclusion.

In the construction of each decision tree model, an individual random subset of the training dataset uses a random subset of the presented variables to decide as to where to partition the dataset at each node. No pruning performed as all decision trees are assembled to their maximum magnitude. The process of building each decision tree to its maximal depth results in a less biased model.

The entirety of the decision tree models taken together form the forest. In this, the forest characterizes the final ensemble model. Each decision tree in this model effectively casts a vote with the majority outcome being classified as the outcome. In the case of regression models, the average value over the ensemble of regression trees is averages to produce the assessment.

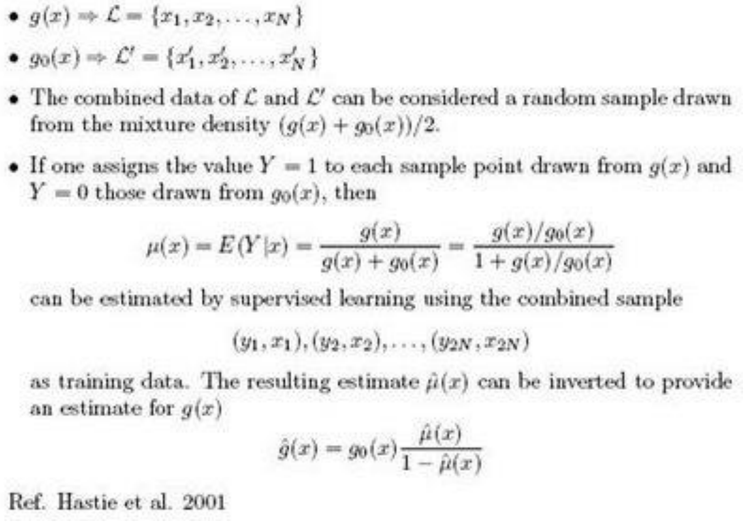
A random forest model is effective for building Security Risk models due to a number of reasons:

1. The amount of pre-processing that needs to be performed on the data is minimal at most,
2. The data does not need to be normalised and the approach is resilient to outliers,
3. Variable selection is generally not necessary the event that numerous input variables are present prior to model building,
4. All of the individual decision trees are in effect independent models. When taken with the multiple levels of randomness that exists within Random Forests, these models tend not to overfit to the training dataset.

The Challenge

Conventional dissimilarity measures that work for simple Risk data may not be optimal in modelling security risk. The use of dissimilarity measures that are based on the intuition of multivariate normal distributions (clusters have elliptical shapes) are generally found not to be optimised in modelling risk. This makes it desirable to have a dissimilarity that is invariant under monotonic transformations of the expressions derived from the risk metrics.

The RF dissimilarity focuses on the most dependent markers whereas the Euclidean distance focuses on the most varying marker.



- $g(x) \Rightarrow \mathcal{L} = \{x_1, x_2, \ldots, x_N\}$
- $g_0(x) \Rightarrow \mathcal{L}' = \{x'_1, x'_2, \ldots, x'_N\}$
- The combined data of $\mathcal{L}$ and $\mathcal{L}'$ can be considered a random sample drawn from the mixture density $(g(x) + g_0(x))/2$.
- If one assigns the value $Y = 1$ to each sample point drawn from $g(x)$ and $Y = 0$ those drawn from $g_0(x)$, then

$$\mu(x) = E(Y|x) = \frac{g(x)}{g(x) + g_0(x)} = \frac{g(x)/g_0(x)}{1 + g(x)/g_0(x)}$$

can be estimated by supervised learning using the combined sample

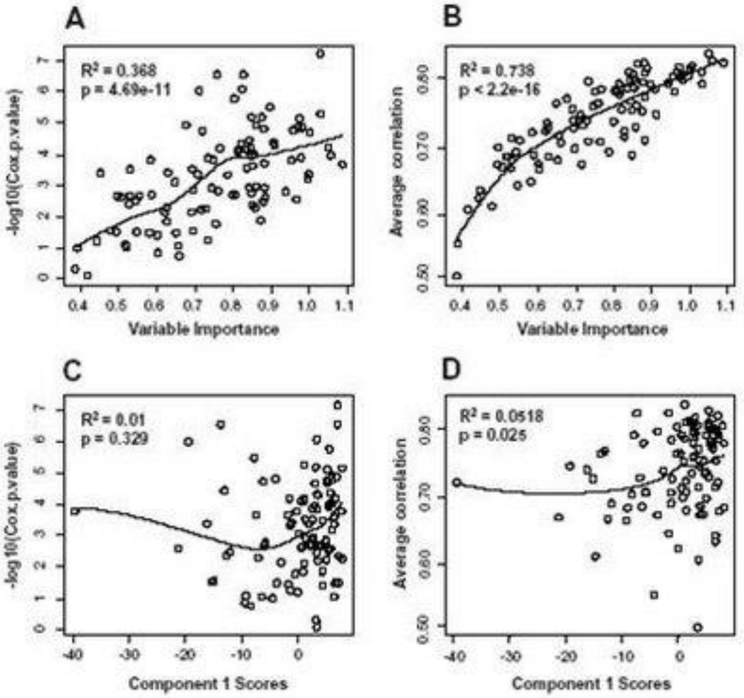
$$(y_1, x_1), (y_2, x_2), \ldots, (y_{2N}, x_{2N})$$

as training data. The resulting estimate $\hat{\mu}(x)$ can be inverted to provide an estimate for $g(x)$

$$\hat{g}(x) = g_0(x) \frac{\hat{\mu}(x)}{1 - \hat{\mu}(x)}$$

Ref. Hastie et al. 2001

Casting an unsupervised problem into a supervised problem



The more important a system is according to RF, the more important it is for survival prediction.

This allows the security risk practitioner to select systems based on quantitative measures rather than perception.

References:

- Ho, Tin Kam (1995). "Random Decision Forest". Proc. of the 3rd Int'l Conf. on Document Analysis and Recognition, Montreal, Canada, August 14-18, 1995, 278-282 (Preceding Work)
- Maindonald, John (Australian National University) Notes from his upcoming publication.