Random Forest Classification Algorithm

By Owen Baines, Amy Chung, Rakesh Raval

University of Leicester

April 2020

Abstract

This paper looks at what a classification algorithm is and examines one such algorithm – random forest. The paper then introduces the topics of entropy, information gain, classification trees and different classification voting methods used in the process. An example is used throughout the paper to demonstrate the concepts and how they link in a clear way and shows a full process used within random forests classification.

Introduction

What is a Classification Algorithm?

Classification algorithms are one of the key components within the discipline of machine learning. They are crafted algorithms which are designed to take an input of numerous correlated or seemingly irrelevant features, and produces a single output representing the common link between the features. Algorithms do this by implementing a mathematical formula to analyse patterns and common features within a known dataset to draw inferences and create rules used to classify unknown data.

As humans, we do not always follow an algorithm to identify certain things, we follow patterns that we have developed throughout our lives. Over time, computers have become more powerful and are able to process large datasets and handle increasingly more complex numerical operations. These advancements increase their ability to not only process operations, but to store and evaluate efficiently. By feeding a computer with a large dataset and programming it to investigate response and explanatory variables, it can investigate potential patterns and mathematically evaluate how much relevant information it adds. The process of teaching a computer how to recognise patterns is the basis of machine learning.[1]

Random Forests

Random forest is an advanced method to analyse a collection of many different classification trees. The name *'random forest'* accurately describes the classification algorithm in the sense that it is the analysis of up to infinitely many classification trees generated at random from a finite set of known, classified data.

Classification Tree

A classification tree is a form of predictive modelling utilised to break down a classification problem into a hierarchy of separate variable-level decisions. The primary purpose of a classification tree is to provide an explanatory insight into patterns and variables that have an impact on a classification. The trees are made up of three main components:

- Root
- Nodes
- Leaves

The root of each tree will represent the initial variable decision which provides the most relevant information to allow the tree to split. Each node forms a new decision point along the tree, with each path of nodes ending in a leaf – a classification outcome. One of the most common algorithms used to build a classification tree is the Iterative Dichotomiser 3 (ID3) algorithm first introduced by Ross Quinlan in his 1986 paper, *'Induction of Decision Trees'*.[2] In order to use the ID3 algorithm, we need to understand entropy and information gain.

What is Entropy?

Entropy, as described by Shannon, is a measure of how much data is produced and how much uncertainty is present in the production. In order to calculate entropy, we need to know the probability of all possible events beforehand – that is, we cannot calculate the entropy of an entirely unknown system. However, we would be able to assume its entropy as 1 since the data would be entirely unpredictable once produced.

An example of a system which is known to always have an entropy of 1 is an unbiased coin – it is not possible to accurately predict whether the coin would land on a head or tail. On the other hand, an entropy of 0 means the dataset is entirely predictable and no new information is added when an event occurs – an example of this is a double-headed coin.[3]

Shannon defined the value of entropy of a discrete system to be:

$$H = -\sum p_i \log_2 p_i$$

where p_i is the probability of the event occurring.

We can extend this further by defining what it means for there to be a conditional entropy such that for two events X and Y, we wish to consider the entropy of Y given event X. In practical terms, this tells us the uncertainty and predictability of event Y assuming event X occurs prior to event Y.

$$H_X(Y) = -\sum p(i)p(j) \log_2 p_i(j)$$

where

- $p(i)$ is the probability of event X occurring,
- $p(j)$ is event Y occurring,
- $p_i(j)$ is the probability of event Y occurring after event X.

To better understand entropy, we will calculate the entropy values of the following random dataset[A1]:

Outlook	Temperature	Humidity	Wind	Play Football
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Table 1 - a table showing a random dataset

We can calculate the entropy of the 'Play Football' response variable as follows:

$$p(Play) = \frac{9}{14}, \qquad p(Not\ Play) = \frac{5}{14}$$

$$H(Play\ Football) = -\sum p_i \log_2 p_i = -\left(\frac{9\left(\log_2 \frac{9}{14}\right)}{14} + \frac{5\left(\log_2 \frac{5}{14}\right)}{14}\right) = 0.9403$$

Since the value 0.9403 is closest to 1, we can conclude that the dataset has a high degree of uncertainty and is not predictable without knowing any other information that may affect the decision to play or not to play (explanatory values).

Now, if we consider each of the explanatory values separately, we obtain the following:

$$H_{Outlook}(Play\ Football) = -\left(\frac{5}{14} \times \left(\frac{3\log_2 \frac{3}{5}}{5} + \frac{2\log_2 \frac{2}{5}}{5}\right)\right) - \left(\frac{4}{14} \times \left(\frac{4\log_2 1}{4}\right)\right)$$

$$-\left(\frac{5}{14} \times \left(\frac{3\log_2 \frac{3}{5}}{5} + \frac{2\log_2 \frac{2}{5}}{5}\right)\right) = 0.6935$$

$$H_{Temperature}(Play\ Football)$$

$$= -\left(\frac{4}{14} \times \left(\frac{2\log_2\frac{2}{4}}{4} + \frac{2\log_2\frac{2}{4}}{4}\right)\right) - \left(\frac{6}{14} \times \left(\frac{4\log_2\frac{4}{6}}{6} + \frac{2\log_2\frac{2}{6}}{6}\right)\right)$$

$$- \left(\frac{4}{14} \times \left(\frac{3\log_2\frac{3}{4}}{4} + \frac{\log_2\frac{1}{4}}{4}\right)\right) = 0.9111$$

$$H_{Humidity}(Play\ Football)$$

$$= -\left(\frac{7}{14} \times \left(\frac{3\log_2\frac{3}{7}}{7} + \frac{4\log_2\frac{4}{7}}{7}\right)\right) - \left(\frac{7}{14} \times \left(\frac{6\log_2\frac{6}{7}}{7} + \frac{\log_2\frac{1}{7}}{7}\right)\right)$$

$$= 0.7885$$

$$H_{Wind}(Play\ Football)$$

$$= -\left(\frac{6}{14} \times \left(\frac{3\log_2\frac{3}{6}}{6} + \frac{3\log_2\frac{3}{6}}{6}\right)\right) - \left(\frac{8}{14} \times \left(\frac{6\log_2\frac{6}{8}}{8} + \frac{2\log_2\frac{2}{8}}{8}\right)\right)$$

$$= 0.8922$$

What is Information Gain?

In order to build a classification tree, we need to know how much useful information is gained about the response variables from the explanatory variable. This is known as information gain. Information gain can be used to analyse how important or influential an explanatory variable is regarding the response variable – that is, how able are we to predict the state of the response variable knowing the state of this explanatory variable. We can define it in relation to entropy as: [2]

$$IG(Y|X) = H(Y) - H_X(Y)$$

Considering the example given above, we can calculate the following information gain values:

$$IG(Play\ Football|Outlook) = 0.9403 - 0.6935 = 0.2468$$

$$IG(Play\ Football|Temperature) = 0.9403 - 0.9111 = 0.0292$$

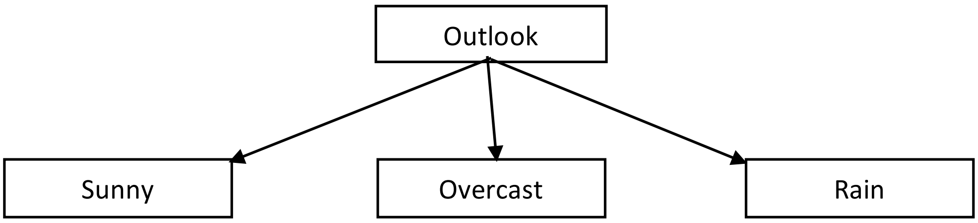
$$IG(Play\ Football|Humidity) = 0.9403 - 0.7885 = 0.1518$$

$$IG(Play\ Football|Wind) = 0.9403 - 0.8922 = 0.0481$$

Building a Classification Tree

Now we have defined entropy and information gain, we use the ID3 algorithm to build a classification tree. To create our root node, we will split the explanatory variable with the highest information gain, which in our example above is ‘Outlook’.

Our classification tree currently looks like:



Next, we need to build new tables split by each outlook value and calculate the new entropy and information gain values for the new datasets.

Play Football split by Outlook: Sunny

Temperature	Humidity	Wind	Play Football
Hot	High	Weak	No
Hot	High	Strong	No
Mild	High	Weak	No
Cool	Normal	Weak	Yes
Mild	Normal	Strong	Yes

Table 2 - a table showing Play Football dataset in relation to Sunny

$$H(Play\ Football) = -\left(\frac{3\log_2\frac{3}{5}}{5} + \frac{2\log_2\frac{2}{5}}{5}\right) = 0.9710$$

$$H_{Temperature}(Play\ Football) = -\left(\frac{2}{5}\times\left(\frac{2\log_2\frac{2}{2}}{2}\right)\right) - \left(\frac{2}{5}\times\left(\frac{\log_2\frac{1}{2}}{2} + \frac{\log_2\frac{1}{2}}{2}\right)\right)$$
$$-\left(\frac{1}{5}\times(\log_2 1)\right) = 0.4000$$

$$H_{Humidity}(Play\ Football) = -\left(\frac{3}{5}\times\left(\frac{3\log_2\frac{3}{3}}{3}\right)\right) - \left(\frac{2}{5}\times\left(\frac{2\log_2\frac{2}{2}}{2}\right)\right) = 0$$

$$H_{Wind}(Play\ Football) = -\left(\frac{3}{5}\times\left(\frac{2\log_2\frac{2}{3}}{3}+\frac{\log_2\frac{1}{3}}{3}\right)\right)-\left(\frac{2}{5}\times\left(\frac{\log_2\frac{1}{2}}{2}+\frac{\log_2\frac{1}{2}}{2}\right)\right)$$

$$= 0.9510$$

$$IG(Play\ Football|Temperature) = 0.9710 - 0.4000 = 0.5710$$

$$IG(Play\ Football|Humidity) = 0.9710 - 0 = 0.9710$$

$$IG(Play\ Football|Wind) = 0.9710 - 0.9510 = 0.0200$$

Our highest information gain value is 0.9710 for the humidity variable. Looking at the entropy of this variable, it is 0. Therefore, the variable is a strong predictor of the response variable. We will terminate the path at this node with the classification dependant on the humidity.

Play Football split by Outlook: Overcast

Temperature	Humidity	Wind	Play Football
Hot	High	Weak	Yes
Cool	Normal	Strong	Yes
Mild	High	Strong	Yes
Hot	Normal	Weak	Yes

Table 3 - a table showing Play Football dataset in relation to Overcast

$$H(Y) = -(\log_2 1) = 0$$

As the entropy is 0, the response variable is highly predictable – this is shown as all the classifications return ‘yes’. Therefore, we have reached the end of this path and terminate at a leaf labelled ‘yes’.

Play Football split by Outlook: Rain

Temperature	Humidity	Wind	Play Football
Mild	High	Weak	Yes
Cool	Normal	Weak	Yes
Cool	Normal	Strong	No
Mild	Normal	Weak	Yes
Mild	High	Strong	No

Table 4 - a table showing Play Football dataset in relation to Rain

$$H(Play\ Football) = -\left(\frac{3\log_2\frac{3}{5}}{5}+\frac{2\log_2\frac{2}{5}}{5}\right) = 0.9710$$

$$H_{Temperature}(Play\ Football) = -\left(\frac{3}{5}\times\left(\frac{2\log_2\frac{2}{3}}{3}+\frac{\log_2\frac{1}{3}}{3}\right)\right)-\left(\frac{2}{5}\times\left(\frac{\log_2\frac{1}{2}}{2}+\frac{\log_2\frac{1}{2}}{2}\right)\right) = 0.9510$$

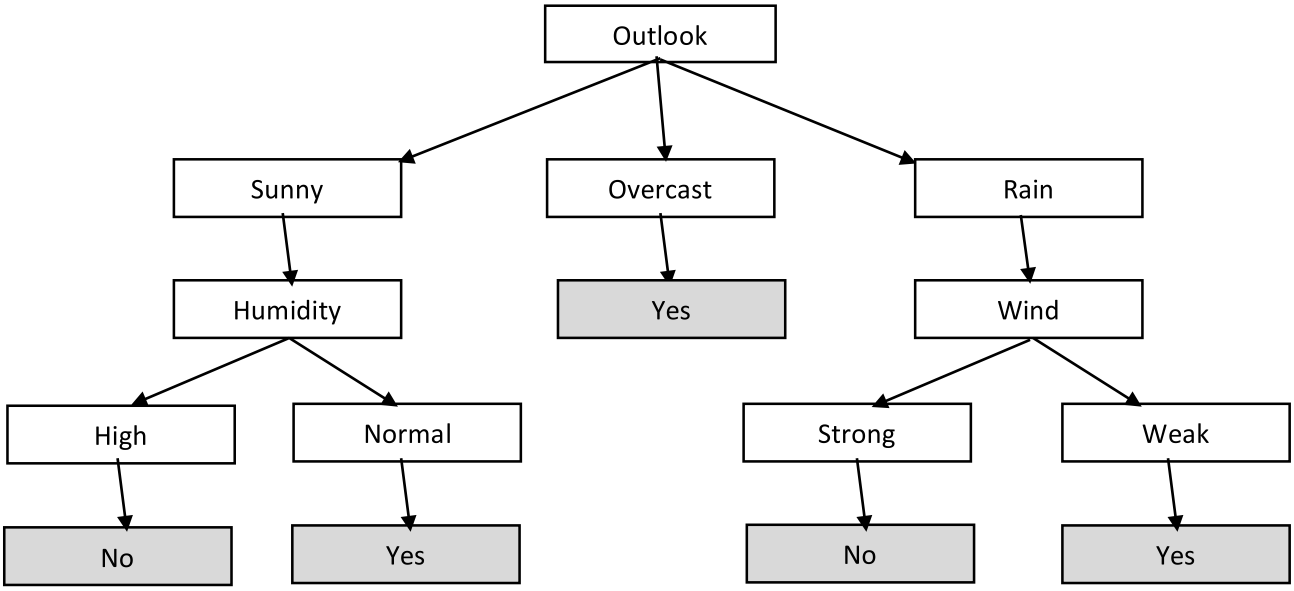
$$H_{Humidity}(Play\ Football) = -\left(\frac{2}{5}\times\left(\frac{\log_2\frac{1}{2}}{2}+\frac{\log_2\frac{1}{2}}{2}\right)\right)-\left(\frac{3}{5}\times\left(\frac{2\log_2\frac{2}{3}}{3}+\frac{\log_2\frac{1}{3}}{3}\right)\right) = 0.9510$$

$$H_{Wind}(Play\ Football) = -\left(\frac{3}{5}\times\left(\frac{3\log_2\frac{3}{3}}{3}\right)\right)-\left(\frac{2}{5}\times\left(\frac{2\log_2\frac{2}{2}}{2}\right)\right) = 0$$

$$IG(Play\ Football|Temperature) = 0.9710 - 0.9510 = 0.0200$$
$$IG(Play\ Football|Humidity) = 0.9710 - 0.9510 = 0.0200$$
$$IG(Play\ Football|Wind) = 0.9710 - 0 = 0.9710$$

Our highest information gain value is 0.9710 for the wind variable. Looking at the entropy of this variable, it is 0. Therefore, the variable is a strong predictor of the response variable. We will terminate the path at this node with the classification dependant on the wind.

Now, using the above calculations – we can draw the next (and final) stage of the classification tree:



Final Classification with Random Forests

As random forests require infinitely-many classification trees, the final part to consider is 'once a forest has been made and each tree has provided a classification, how do we get a final classification?'. There are many ways to do this, but we will only look at the two most common ways.

The most common method is called 'majority voting'. In this method, each classification tree has an equal vote in the final classification. For example, if we have 10 trees and 4 results in a classification of A while 6 results in a classification of B – the final classification for the new dataset would be B. The other method we will look at is called 'performance voting'. In this method, each tree has a vote weighted to its accuracy to classify a set known as a 'validation set'. In order to use this method, each tree must have individually had its accuracy recorded by running the validation set through its paths. Below is a table which demonstrates the application of both methods.[4]

Tree	Accuracy	Classification
1	75%	A
2	85%	B
3	46%	A
4	39%	B
5	90%	A
6	28%	B
7	65%	B
8	60%	B
9	18%	B
10	87%	A

Table 5 - a table showing the application of both methods

Under our majority voting method, this would result in a classification of B. However, under performance voting, we would get a classification of A.

Conclusion

Random forest is a method to improve the overall accuracy of a singular classification tree algorithm on a dataset. This is done by reducing the potential of bias in a dataset from affecting a singular tree by spreading out the bias across infinitely many trees used in analysis. The method of choosing the final classification as well provides another step-in minimising the potential of bias being introduced by allowing trees to vote based on their accuracy score from a validation test set. Random forests can be used to predict and analyse trends in either numerical, categorical or binary classification datasets with ease and allows ranking of explanatory variables (features of classification) based on their unpredictability, and how much information they pass onto deciding the final classification. More complicated and thorough voting methods can be introduced in the final stage to improve accuracy even further such as voting based upon average entropy of the system – a more predictable dataset has a greater say than a more unpredictable dataset in classification.

References

[1] Alpaydin, E. (2010). *Introduction to Machine Learning*. Cambridge, Mass.: MIT Press

[2] Quinlan, R. (1986). *Induction of Decision Trees.* Machine Learning 1 (81-106). Kluwer Academic Publishers, Boston

[3] Shannon, C. (2001). *A Mathematical Theory of Communication.* Mobile Computing and Communications Review, Vol.5(1), pp.3-55

[4] Rokach, L. Maimon, O. (2008) *Data Mining with Decision Trees: Theory and Applications*. World Scientific Publishing

Appendices

[A1] https://miro.medium.com/max/1096/1*Jr1Qf-m1u-vGzDao6_CxqA.png