

Performance monitoring of an industrial boiler: classification of relevant variables with Random Forests

Matthieu Sainlez ^a, Georges Heyen ^b

^aCRISIA, Haute-Ecole Robert Schuman, Chemin de Weyler 2, B-6700 Arlon, Belgium,
E-mail: matthieu.sainlez@hers.be

^bLASSC, Université de Liège (ULg), Sart Tilman B6A, B-4000 Liège, Belgium,
E-Mail: G.Heyen@ulg.ac.be

Abstract

A data mining methodology, the random forests, is applied to analyze pollutant emission from the recovery boiler of a Kraft pulping process. Starting from a large database of raw process data, the goal is to identify the input variables that explain the most output variations.

Keywords: data mining, random forests, Kraft pulping process, recovery boiler, atmospheric pollutants.

1. Introduction

Data Mining refers to extracting useful knowledge from large amounts of data. Starting from large databases, the main objective is to find interesting latent patterns [1].

Particularly, a random forest [2, 3] is an ensemble of unpruned classification or regression trees, induced from bootstrap samples of the training data, using random feature selection in the CART induction process. Prediction is made by aggregating the predictions of the ensemble. Internal estimates are also used to measure variable importance [2].

Within the framework of a Kraft pulp mill, we analyze the emissions of the recovery boiler, and particularly the nitrogen oxide emission. This kind of boiler acts both as a high-pressure steam boiler and as a chemical reactor with reductive and oxidative zones. Significant perspectives already exist to reduce atmospheric pollutants, and the identification of the most important variables is an interesting byproduct of random forests.

2. The Kraft pulping process

The Kraft process [4] is an alkaline process to produce chemical pulp. A pulp mill can be divided in two main areas: fiber line and chemical recovery loop.

Cellulose fibers are dissociated from lignin by cooking the chips in a solution of sodium hydroxide (NaOH) and sodium sulfide (Na₂S), called white liquor. The residual black liquor is washed from the pulp and treated to recover the cooking chemicals. The black liquor is concentrated and burned in a recovery furnace to yield an inorganic smelt of sodium carbonate (Na₂CO₃) and Na₂S. The smelt is dissolved to form green liquor, which is treated to recycle the calcium carbonate and to regenerate the white liquor.

3. Random Forests methodology

In this paper, we consider a regression problem in which we are trying to predict the value of a continuous variable: pollutant emission of nitrogen oxide (NO_x).

We have a training set $\mathbf{z} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ where each $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ is the i^{th} measurement vector of p input attributes, y_i is the continuous outcome. We fit a model to \mathbf{z} , obtaining the prediction $\hat{f}(x)$ at input x .

Bagging is a general strategy for improving predictor accuracy [1]. The bagging algorithm creates an ensemble of models (by bootstrap sampling) for a learning scheme where each model gives an equally-weighted prediction.

Particularly, random forests are a combination of tree predictors such that each binary tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [2].

3.1. From binary trees to random forests

The binary tree [5, 6] is a widely used framework in data mining; this concept can be applied both to classification or regression problems. Basically, it's a sequence of binary decisions applied to the input variables; each non-terminal node contains a decision involving the comparison of an attribute with a given threshold, which then leads to another node or to a leaf (a terminal node). The root node contains the whole dataset which is recursively splits into two branches at each node. A greedy algorithm selects the attribute and threshold that maximizes a given fitness measure.

A particular tree framework called CART (for "*Classification and Regression Trees*") maximizes the Gini index that selects the split with the lowest impurity at each node; CART was introduced in 1984 (Breiman et al., [7]).

Generally, the resulting tree is easily interpretable (giving a set of decision rules), it works with both numerical and categorical data, and it's a non parametric method (no a priori assumption is made). Unfortunately, trees are sensitive to small changes in the learning sample (Breiman, [8]). Moreover, unstable trees can be stabilized via an ensemble method: we average the predictions of a set of individual models (see for example, [14]).

Practically, we have a single training data set, and so we have to find a way to introduce variability between the different models: we use bootstrap data samples [5]. A bootstrap replicate is a random subset of the original dataset, of the same length, taken with replacement [9].

We generate m bootstrap samples and then use each to train a separate copy of a predictive model. This procedure is known as bootstrap aggregating or *bagging* [8].

The aim of aggregating is to create an improved model. We take the average value of each prediction for a given test sample.

For each bootstrap sample \mathbf{z}_i ($i = 1, \dots, m$), we grow a CART tree T_i and we aggregate the ensemble $\{T_i\}_i^m$ (see Figure 1). For a giving prediction $\hat{f}_i(x)$, the bagging estimate is the average of predictions over the m trees.

Bagging is very helpful for reducing variance and, for prediction, Breiman [8] proved theoretically that a bagged predictor will always have improved accuracy over a single predictor.

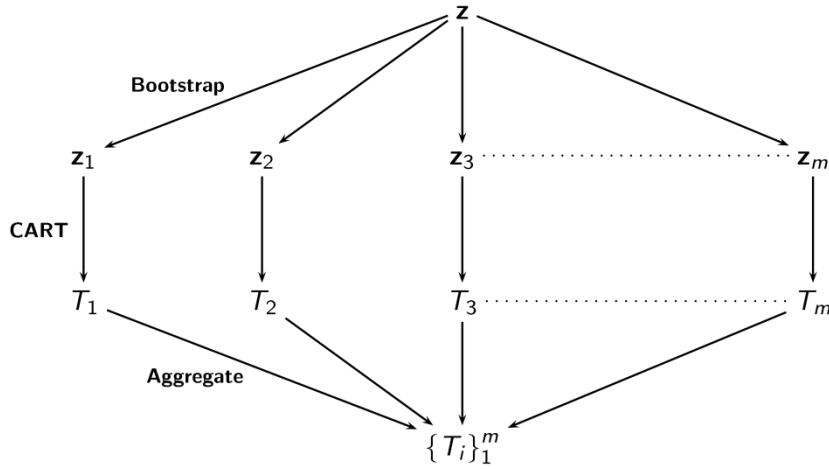


Figure 1 : Bagging : bootstrap aggregating

3.2. Random Forests algorithm

A random forest (Breiman, [2]) is an ensemble of unpruned classification or regression trees, induced from bootstrap samples of the training data, using random feature selection in the CART induction process.

In the random forest methodology [2, 6, 10], a second source of diversity is introduced during the growing of each tree. For each node, the method selects a small random subset of k attributes (from the p input attributes) and uses only this subset to search for the best split. This random selection of features at each node decreases the correlation between the trees in the forest thus decreasing the forest error rate.

We fit each tree on bootstrap sample and we select threshold and attribute at each node from a subset of attributes; the algorithm is described below (Hastie et al., [10]):

-
- For $i = 1, \dots, m$:
 - a) Draw a bootstrap sample z_i of size n from the original sample z .
 - b) Grow a random forest tree T_i to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
 - i) Select k variables at random from the p variables ($k \leq p$).
 - ii) Pick the best variables/split-point among the k .
 - iii) Split the node into two daughter nodes.
 - Output the ensemble of trees $\{T_i\}_1^m$, the prediction at a new point x is given by:

$$\hat{f}_{RF}^m(x) = \frac{1}{m} \sum_{i=1}^m T_i(x)$$

In this study, we take $(k, n_{min}) = (\lfloor \frac{p}{3} \rfloor, 5)$; these are classical values for regression [10]. An analysis of complexity and prediction score helps for selecting appropriate m .

4. Industrial case study

In this paper, we analyze nitrogen oxide (NO_x) emission from a Kraft recovery boiler; the main objective is to find explanatory attributes for predicting NO_x pollutant emissions (we focused this paper on the attributes selection scheme).

The recovery boiler furnace can be considered as consisting of three distinct zones [4]: a drying zone where the black liquor is fired, a reduction zone at the bottom, and the oxidation zone in the turbulent upper section. Air for combustion is introduced from the bottom upward as primary, secondary, tertiary and quaternary air (at different velocities to ensure complete mixing).

The original database is a $(n \times p) = (65509 \times 56)$ matrix. The p attributes are mainly physical flow rates, pressures, and temperatures of black liquor, fuel, air...

We use a Matlab R13 implementation of Breiman's random forest algorithm for regression ([11], based on Breiman and Cutler's original Fortran code version 3.3).

4.1. A data mining approach for modeling

Firstly, the data need to be preprocessed to make it appropriate for the study [12]. Then, the given original data set is partitioned into two independent sets [1], a training set (70% of the data) and a test set (the remaining 30%). The training set is used to derive the model, whose accuracy is estimated with the test set (see Figure 2).

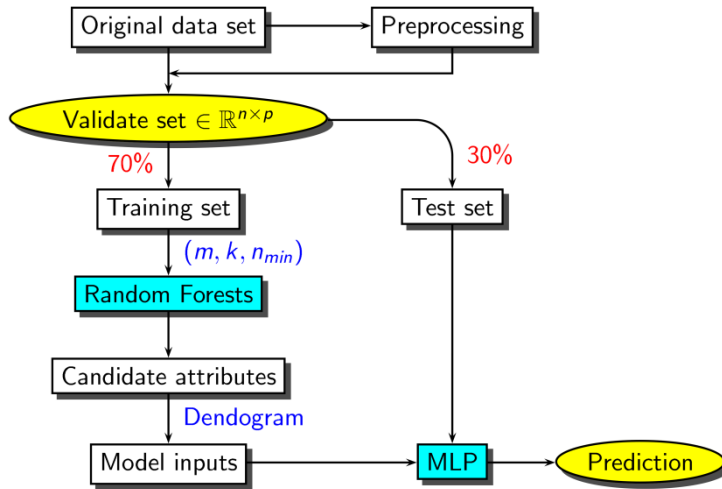


Figure 2: A data mining approach for prediction.

We use random forests to rank input attributes according to their importance measure. Candidate attributes correlations are also analyzed (e.g., with a dendrogram) to avoid redundancies. The resulting attributes are the inputs of the model.

A successful model is the feed-forward neural network [5], known as multilayer perceptron (MLP). In the end, the model relevance is assessed by its performance for predicting new observations.

4.2. Attributes selection scheme

In many data mining applications, only a few input variables have substantial influence on the response. It is often useful to learn the relative importance or contribution of each input variable in predicting the response.

Random Forests use the out-of-bag (OOB) samples to a variable importance measure. On average, 37% of the samples will not be present in a given bootstrap replicate [6, 9]: they are called OOB sets. When a tree in the forest is grown, the OOB samples are passed down the tree, and the prediction accuracy is recorded.

Then, one at a time, each attribute values are randomly permuted in the OOB samples, and the accuracy is again computed. The decrease in accuracy as a result of this permuting is averaged over all trees, and is used as a measure of the importance of a variable in the random forest [10, 13].

4.3. Results

Attributes are ranked according to this importance measure (expressed as a percentage of the overall importance, see Figure 3). For a fixed number of trees m , a variable with a larger importance score relative to other variables indicates that the variable is important for regression. This hierarchy presents only the first 25 relevant attributes.

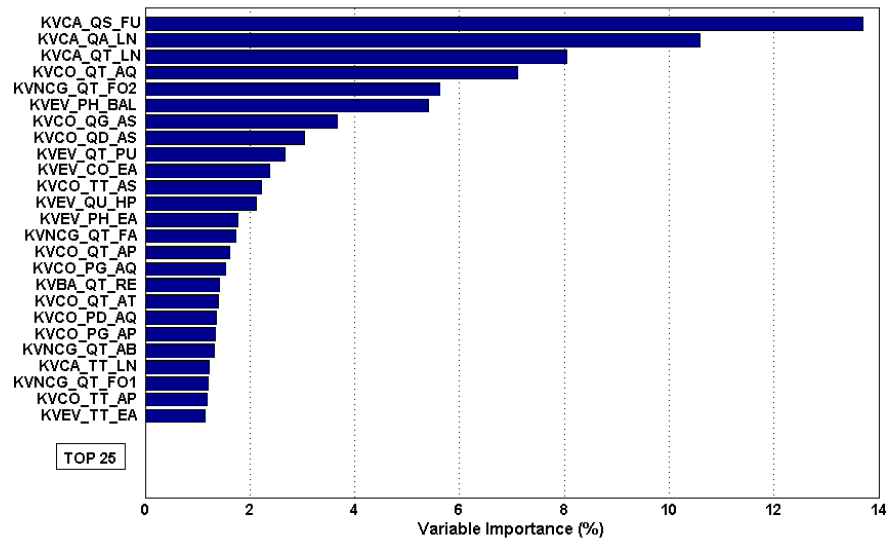


Figure 3: Random Forest attributes score importance - $(m, k, n_{min}) = (80, 18, 5)$

After a breakdown, during a boiler start up or consequent to high variation in steam demand: the furnace is brought to the right temperature by burning heavy fuel. We can observe that, logically, the fuel rate introduces at the lower of the furnace (QS-FU) is the more relevant attribute for predicting NO_x emissions.

The entering black liquor flow rate (QA-LN) is a part of the total black liquor flow rate (QT-LN) which is circulated in a loop around the liquor guns; these variables are highly correlated. Then, quaternary air (QT-AQ) was especially designed to control NO_x emissions and non-condensable gas (NCG-QT) coming from the process are incinerate into the furnace.

5. Conclusions

This paper briefly traced the application of random forests to a pulp mill atmospheric pollutant. We talked only about the attributes selection scheme; these selected attributes can also be used for prediction with a neural network (e.g., a multilayer perceptron). Random Forests handle very large database and its internal variable importance measure is very helpful for understanding complex interactions between attributes and discovering latent patterns. This method is easy to use and quite fast, requiring only a little tuning on parameters.

Acknowledgements

This project was supported by the Walloon Region (FIRST Program, PHOEBUS). The authors would like to thank PEPITe S.A. (Liège, Belgium) for the expert opinion on the subject.

References

- [1] Jiawei Han, Micheline Kamber. *Data Mining : Concepts and Techniques*. Morgan Kaufmann Publishers, 2006
- [2] Leo Breiman, Random Forests. *Machine Learning*, 45: 5-32, 2001.
- [3] Yongheng Zhao, Yanxia Zhang, Comparison of decision tree methods for finding active objects. *Advances in Space Research*, 41: 1955–1959, 2008.
- [4] Gary A. Smook, *Handbook for Pulp & Paper Technologists - Third Edition*. Angus Wilde Publications, 2002.
- [5] Christopher M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006
- [6] P.M. Granitto a, F. Gasperi, F. Biasioli, E. Trainotti, and C. Furlanello. Modern data mining tools in descriptive sensory analysis: A case study with a Random forest approach. *Food Quality and Preference*, 18: 681–689, 2007.
- [7] Breiman, L., Friedman, J., Olshen, R. and Stone, C. *Classification and Regression Trees*, Wadsworth, 1984.
- [8] Leo Breiman, Bagging Predictors. *Machine Learning*, 24:123-140, 1996.
- [9] B. Efron, Estimating the error rate of a prediction rule: some improvements on cross-validation. *Journal of the American Statistical Association*, 78: 316–331., 1983.
- [10] Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction - Second Edition*. Springer Series in Statistics, 2009.
- [11] Ting Wang, Package random forests for Matlab R13, available at <http://lib.stat.cmu.edu/matlab/>
- [12] Sigurdur Olafsson, Xiaonan Li, and Shuning Wu. Operations research and data mining. *European Journal of Operational Research*, 187: 1429–1448, 2008.
- [13] Kellie J. Archer, Ryan V. Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52: 2249 – 2260, 2008.
- [14] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely Randomized Trees. *Machine Learning*, 36: 3-42, 2006.