Random Garden: a Supervised Learning Algorithm

Giardino Casuale: un Algoritmo di Apprendimento Supervisionato

Ivan Luciano Danesi, Valeria Danese, Nicolò Russo and Enrico Tonini

Abstract Classification and Regression Trees model and two other tree-based models are considered. These latter tree-based models are the Random Forest and the Random Garden, presented in this work. The feature selection impact on the different algorithms is investigated. The described procedures are applied to 18 Customer Relationship Management data sets constructed in Banking field. The goal is binary classification. Our results show that the best algorithm depends on data set characteristics, as dimensions, proportion of success events and the application of feature selection.

Abstract *Gli alberi di classificazione e di regressione e altri due modelli ad albero sono considerati. Questi ultimi sono la Foresta Casuale e il Giardino Casuale, presentato in questo lavoro. L'impatto della selezione delle variabili è investigato. Le procedure descritte sono applicate a 18 tabelle costruite per la gestione della relazione con i clienti in un contesto bancario. Lo scopo è una classificazione binaria. I nostri risultati mostrano come il migliore algoritmo dipenda dalle caratteristiche dei dati, come dimensioni, proporzione di successi e l'applicazione di procedure di selezione delle variabili.*

Key words: Classification and Regression Trees, Customer Relationship Management, Feature Selection, Random Garden, Random Forest, Tree Bagging.

Ivan Luciano Danesi, Valeria Danese and Nicolò Russo
Data & Analytics Data Science
UniCredit Business Integrated Solutions S.C.p.A
Via Livio Cambi 1, 20151 Milano, Italy, e-mail: ivanluciano.danesi@unicredit.eu, valeria.danese@unicredit.eu, nicolo.russo@unicredit.eu

Enrico Tonini
Independent Author e-mail: enrico.tonini.stat@gmail.com

1 Introduction

In Big Data era, the Customer Relationship Management (CRM) has been receiving a great deal of attention. CRM can leverage on several data sources. During the years, clients' data have been collected in many forms. Today there is the technology for the storage of all of these data in the same place, as well as the tools for merging and using them.

In this work we apply tree-based classification models to different data sets of an important financial institution. Such data sets are constructed in CRM field. The goal of classification applications is to discriminate between success and failure events. The success events are defined as CRM relevant occurrences (*e.g.* claims or product purchases).

Firstly, the features useful for the modeling step are selected. During feature selection step, we consider either a supervised and an unsupervised approach.

Secondly, three tree-based models are applied. A Classification and Regression Tree (CART) model is used as benchmark for our evaluation. The *random forest* algorithm, the most diffused algorithm for bagging and ensembling trees, is then considered. Finally, the *Random Garden*, a CART bagging designed for high dimensional data, is applied.

The tree-based models are estimated on the different data sets, with and without feature selection.

In Section 2 the feature selection techniques and the algorithms are briefly outlined. The results are presented in Section 3 and discussed in Section 4.

2 Section Heading

2.1 Feature Selection

Usually the higher the number of features that can be collected, the more relevant is discriminating between valuable predictors (to be kept) and not (to be discarded). The consequences of including non-informative features in a model may be different, depending on the selected algorithm. Furthermore, in presence of a high number of features, some relevant ones could not be included in trees splitting process enough times to correctly determine the results, as pointed out in [5].

Feature selection procedures can be mainly divided in two categories: wrapper methods and filter methods ([4]). The former involves the use of predictive algorithms while the latter approach analyzes the features one by one and keeps only the ones satisfying a defined rule.

As wrapper methods, we consider the permutation test and as filter one the analysis of variance and correlations predictors. From now on, we refer them as our features selection (FS) procedures through which we generate the datasets.

2.2 Considered Tree-Based Algorithm

2.2.1 Classification and Regression Trees

Classification and Regression Trees (CART) are constructed to generate a response or a class Y using a set of inputs $X_1, \ldots, X_p$ by means of binary splits (see [2] for details).

2.2.2 Random Forest

Random forests (RF) [1] improve predictive accuracy by generating a large number of random trees, then classifying a case using each tree in this new *forest*, and deciding a final predicted outcome by combining the results across all the trees.

More specifically, from the available data a number k of bootstrapped training samples is considered. On each of these k samples, a prediction tree is constructed considering, at each split, a random sample of $m < p$ predictors. This selection of predictors is performed in order to avoid that the strongest predictors would determine the top split in almost all the estimated trees. On the contrary, sampling at each split m predictors, the strong predictor is considered $(p-m)/p$ times. The result is a *forest* composed by *decorrelated* trees. This is done since averaging not-correlated quantities leads to a variance reduction, otherwise higher averaging high-correlated quantities. In random forests the trees are usually fully grown.

2.2.3 Random Garden

In this work, we introduce an algorithm for ensembling fully grown CART in order to generate a forest. The trees are different from each other by randomly selecting both the individuals (bagging) and the features, like random forest. More in detail, the algorithms is composed by the following steps.

1. The features are divided into two groups, based on their impact on the response variable. Features impact on the response variable is evaluated by using the p-value of a correlation test between features and the response variable. The choice of correlation test depends on variable type:

   - $F-$test for binary features;
   - χ^2-test for categorical features with more than 2 levels;
   - Wilcoxon test for numerical features.

   Features with p-value less or equal to critical value 0.05 are considered highly relevant, otherwise less relevant. The former and the latter group constitute the sample F_R and F_N of features respectively. Clearly $F_R \cup F_N = F$, where F is the complete set of features.
2. Each tree is constructed by

a. Sampling with a replacement of a number of records equal to the total amount of observations, thus considering on average 63.21 % different observations in each sample (see [2]).
b. Sampling *mtry* number of features, set by default equal to the square root of the dimension of F. This sample is constructed in a stratified way from F_R and F_N sets. From F_R are extracted q_R features, where q_R is the proportion of F_R dimension with respect to F dimension. Conversely, from F_N are extracted q_N=*mtry*-q_R features.
c. Growing a fully grown CART (not pruned).

3. Considering the mean of all CART outputs as predicted response variable.

The result is a forest of trees with a lower number of freedom degrees than Random Forest trees. Since the trees are constructed one by one and well-finished, we name this algorithm as Random Garden (RG).

The procedure described above for RG is different from RF mainly due to the random feature selection, since it is performed in a stratified way. An example of RF definition with stratified random feature selection is given by xRF algorithm introduced in [5]. In this study our aim is to stress the differentiation of the trees. As a matter of fact, we do not extract the features at each split as in xRF, but we sample features once for each tree during RG construction.

3 Application

3.1 The Data

The application is performed on 18 datasets collected in a financial institution [1]. Datasets are related to Bank customers and are constructed by merging different sources of data and this has been the first step of the analysis. The response variable is binary (0 or 1, for failure and success respectively). Data sets characteristics are in Table 1.

By looking at Table 1, we can identify three clusters of datasets with respect to Ncust and Pev characteristics. The clusters index is indicated in Table 1. In particular we observe three clusters.

Cluster 1. Population with higher number of customers and lower proportion of success events.
Cluster 2. Population with lower number of customers and lower proportion of success events.
Cluster 3. Population with lower number of customers and higher proportion of success events.

[1] More details about the datasets are not provided due to Legal and Compliance issues.

Table 1 Index number of data set (DS), number of features (Nfeat), number of customers (Ncust), number of events (Nev), proportion of success events in percentage (Pev) and relative cluster (Cluster) for the 18 data sets.

DS	Nfeat	Ncust	Nev	Pev	Cluster
1	311	576546	1983	0.344	1
2	256	158229	463	0.293	2
3	224	171134	412	0.241	2
4	295	94456	1381	1.462	2
5	317	100024	1193	1.193	2
6	264	95229	939	0.986	2
7	689	220598	910	0.413	2
8	376	227578	500	0.220	2
9	423	39620	4760	12.014	3
10	440	35377	706	1.996	2
11	382	12957	353	2.724	2
12	388	24892	2021	8.119	3
13	443	223276	25163	11.270	3
14	257	70257	2376	3.382	2
15	824	423929	12624	2.978	1
16	642	129791	2031	1.565	2
17	932	421887	8903	2.110	1
18	578	11801	515	4.364	2

3.2 Results

Datasets are divided in training and test sets. The training set is used for model estimation and is composed by the 30% of the observations, while the testing set is used for model performance evaluation and is composed by the remaining 70% of the observations. We apply Monte Carlo cross-validation. This procedure creates multiple splits of data into train and test sets and each split is randomly performed from the full dataset. The number of random splits is set equal to 50.

For every combination of dataset, Monte Carlo split and algorithm/feature selection configuration, the model is trained and tested for measuring the performance indicators. Overall, 5400 different models are trained. We choose Area Under ROC Curve (AUC) as performance metric due to its relevance in this application ([3]). As a matter of fact, we are interested in *ranking* the customers according to the model output. We report a summary for AUC values in Table 2.

4 Discussion

The approaches show good performances applied to the different datasets in almost all the trials. As expected, results can be very different among the datasets.

Algorithms performances, measured as AUC score, seems to be related to the cluster where each dataset belongs to. By way of example, in Clusters 1 and 2 which

Table 2 AUC values on average for the three dataset clusters for the different algorithm and FS configurations.

FS	Algorithm	Cluster 1	Cluster 2	Cluster 3
no	CART	0.7136	0.7568	0.8702
	RF	0.7977	0.7833	0.8827
	RG	0.8074	0.7995	0.8359
yes	CART	0.7470	0.7359	0.8589
	RF	0.7935	0.7690	0.8767
	RG	0.8026	0.7923	0.8403

are the ones exhibiting the lowest percentages of positive targets, RG algorithm performs better, RF one performs worse although much better than CART. On the other hand, on Cluster 3 datasets which are the ones with a smaller customer size and more balanced in terms of positive targets proportion, RF is the algorithm performing better, followed by CART and RG. Regarding the impact of the FS procedure, the more valuable improvement is observed only for CART.

Declaration of Interest

The views and opinions expressed in this paper are those of the authors only, and do not necessarily represent the views and opinions of UniCredit Business Integrated Solutions S.C.p.A. or any other organization. All the computations have been conducted on anonymized data on UniCredit servers by UniCredit employee. The results have been observed only in aggregated form.

References

1. Breiman, L.: Random forests. Machine Learning **45(1)**, 5–32 (2001)
2. Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning. Springer series in statistics, Springer, Berlin (2001)
3. Hanley, J.A. and McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology **143(1)**, 29–36 (1982)
4. John, G., Kohavi, R., Pfleger, K.: Irrelevant Features and the Subset Selection Problem. Proceedings of the Eleventh International Conference on Machine Learning **129**, 121–129 (1994)
5. Nguyen, T.T., Huang, J.Z., Nguyen, T.T.: Unbiased Feature Selection in Learning Random Forests for High-Dimensional Data. The Scientific World Journal (2015) doi: 10.1155/2015/471371