An extensive experimental survey of regression methods

Introduction

The main objective of our opted research paper is to provide a "road map" for researchers who want to solve regression problems and need to know how well the currently available Regression methods works.

In machine learning, Regression methods are designed to predict continuous numeric outputs where an order relation is defined. Regression has been widely studied from the statistics field, which provides different approaches to this problem:

- Linear and Generalized Linear Regression
- Least and Partial Least Squares Regression (LS and PLS)
- Least Absolute Shrinkage and Selection Operator (LASSO) and Ridge Regression
- Multivariate Adaptive Regression Splines (MARS)
- Least Angle Regression (LARS)

Paper Contribution

This research paper was contributed by :

M. Fernández-Delgado, M.S. Sirsat, E. Cernadas[a] , S. Alawadi[a] , S. Barro[a],
M. Febrero-Bande[b]

[a] *Centro Singular de Investigación en Tecnoloxías da Información da USC (CiTIUS), University of Santiago de Compostela, Campus Vida, 15782, Santiago de Compostela, Spain*

[b] *Department of Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela, Campus Vida, 15782, Santiago de Compostela, Spain*

Data pre-processing : Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format. It invloves Data Cleaning, Data Transformation, Data integration, Data Reduction, Data discretization.

Below are the Data pre-processing steps followed :

In the current research paper opted, 48 of 82 datasets is selected (81 because the Air Quality dataset is repeated) listed as regression problems by the UCI Machine Learning Repository.

The remaining 33 datasets were discarded due to the reasons listed in Table 1, out of which 13 of them were discarded whose outputs have few values (specifically, less than 10) -- 17 datasets. Some of the 48 original UCI regression datasets selected for this work generated several regression problems, one for each data column which can be used as output for regression.

The reason to discard the above datasets were, because including them in the dataset collection might favor some regression models with respect to others, thus biasing the results of the study. These datasets should be considered as ordinal classification instead of pure regression problems.

Although several datasets in Table 2 (research paper reference) contain several ten thousand patterns, and even half (buzz-twitter), one (greenhouse-net) and two million patterns (household-consume), the current study is not oriented to large-scale datasets because the available implementations of the majority of regression models would not work on such large datasets due to memory errors or excessive time.

Thus, including large-scale datasets on the current study would bias the results and conclusions, limiting the comparison to those models with implementations that could be run on large data and favoring them over the remaining ones.

The output is pre-processed using Box–Cox transformation in order to make it more similar to a symmetric uni-modal distribution, with the boxcox function (MASS package) of the R statistical computing language. In the “greenhouse-net” and “com-crime-unnorm” datasets, the decimal logarithm of the inputs are used, due to the wide range of many inputs.

Repeated constant and collinear inputs are removed from all the datasets. To calculate the coefficients of the linear model trained the lm function in the stats R package was used on the whole dataset, and the inputs with NA (not available) coefficients in the linear model are removed. This reason leads e.g. the Blog feedback dataset to reduce its inputs from 280 to 13. The rationale behind this is that constant, repeated or collinear inputs lead many models to develop calculations with singular matrices, so it is useful to remove these inputs in order to avoid the subsequent errors. Also, the inputs with discrete values are replaced by dummy/indicator inputs. For each discrete inputs with n values, it is replaced by
n − 1 dummy binary inputs.

Machine Learning Activity

In order to control the execution of each single model, the models were directly run using the corresponding R packages, instead of using train function of caret package.

The model operation is optimized by tuning the set of hyperparameters specified in the caret model list. Almost all the models that were used, have from one to four tunable hyperparameters.

The specific hyperparameter values are calculated by the getModelInfo function of the caret package. For some models (e.g. gprRad) and datasets, this function returns a value list with less items than the number specified in values.txt, and even sometimes just one value is used. In these cases, although the caret model list specifies that hyperparameter as tunable, only one value is used in practice.

The experimental work identifies several outstanding regression models: the M5 rule based model with corrections based on nearest neighbors (cubist), the gradient boosted machine (gbm), the boosting ensemble of regression trees (bstTree) and the M5 regression tree.

Cubist achieves the best squared correlation (R^2) in 15.7% of datasets being very near to it, with difference below 0.2 for 89.1% of datasets, and the median of these differences over the dataset collection is very low (0.0192), compared e.g. to the classical linear regression (0.150). Cubist is slow and fails in several large datasets, while other similar regression models as M5 never fail and its difference to the best R2 is below 0.2 for 92.8% of datasets.

Committee of neural networks (avNNet), extremely randomized regression trees (extraTrees, which achieves the best R2 in 33.7% of datasets), random forest (rf) and ε-support vector regression (svr), but they are slower and fail in several datasets. Fastest regression model is least angle regression lars, which is 70 and 2,115 times faster than M5 and cubist, respectively.

Result analysis with metrics used from paper

This experimental work uses the following methodology:

For each dataset with less than 10,000 patterns, N = 500 random partitions are generated, using the 50% of the patterns for training, 25% for validation (in hyperparameter tuning) and 25% for test. For each dataset with more than 10,000 patterns, a 10-fold cross validation is developed.

Each regression model is trained on the training partitions for each combination of its hyperparameter values, and it is tested on its corresponding validation partition. The performance measures used are the root mean square error (RMSE), the squared correlation (R^2) and the mean absolute error (MAE).

The above three different measures are used in order to give more significance to the results, which in this way can be observed from three different points of view, and also in order to evaluate whether they are coherent suggesting similar conclusions.

Finally, the model is trained on the training partitions using the selected combination of its hyperparameter value and tested on the test partitions. The performance measurements are RMSE, R^2 and MAE between the true and predicted output values concatenated for the N test sets.

R^2 is calculated using the predicted and true outputs for the test patterns, while it is often used to measure the percentage of variance explained by the model on the training patterns. Those regression models which lack tunable hyperparameters are trained on the training partitions and tested on the corresponding test partitions, and the average RMSE, R2 and MAE over the test partitions are the quality measurements.

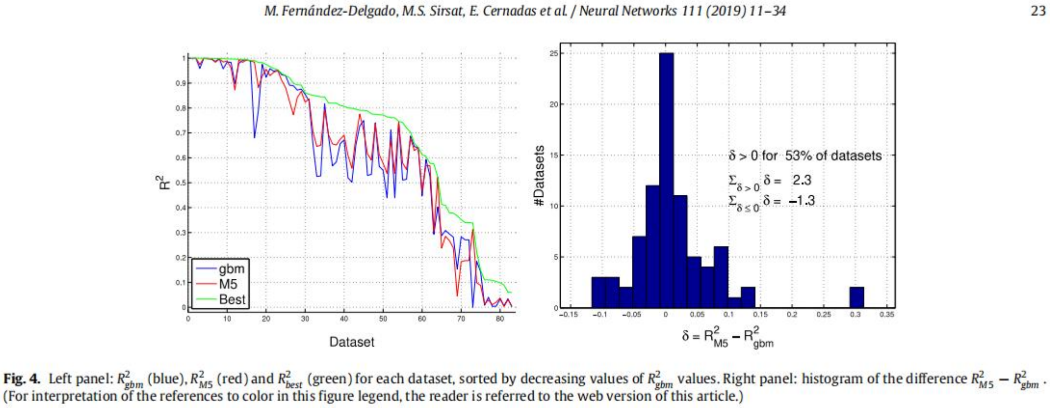
Collinear inputs are removed from the dataset in the initial preprocessing, for certain partitions some inputs in the training set may be collinear despite of being not collinear considering the whole dataset.

The paired T-test gives significant differences, labeled as an asterisk (*), except for the first three models, while the Dunnett, two-sample T and Wilcoxon tests only label few models as statistically different, including svr, penalized and svmRad (the Wilcoxon test also labels rf and grnn as different).

Sign test, which counts the number of datasets where each regression model achieves the best R2, labels all the models as statistically different to cubist excepting extraTrees. Post-Hoc Friedman–Nemenyi test, which develops a comparison of multiple models, identifies as statistically significant the differences with all the models excepting gbm, bstTree extraTrees and svr.

Overall, the number of experiments where the model failed is 1205 and represents 18.85% of the 6391 experiments.

Exploratory Data Analysis / Visualization



M. Fernández-Delgado, M.S. Sirsat, E. Cernadas et al. / Neural Networks 111 (2019) 11–34 23

Fig. 4. Left panel: R^2_{gbm} (blue), R^2_{M5} (red) and R^2_{best} (green) for each dataset, sorted by decreasing values of R^2_{gbm} values. Right panel: histogram of the difference $R^2_{M5} - R^2_{gbm}$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 4 (left panel) plots R2 best and the R2 achieved by M5 and gbm. M5 is near R2 best more often than gbm, and in several cases gbm is clearly below M5, but the former rarely outperforms the latter.

This shows that overall M5 outperforms gbm, although the latter is higher in the global ranking. Cubist, gbm and bstTree fail for some datasets, while M5 never fails.