Weka Report

Panagiotis Michael U204N2110

Datasets

Selwood Dataset

Source:

Project instructions

Description:

The provided file contains data for 31 molecules that have been tested in vitro for a particular biological activity. The first line of the file contains its name and is followed by a blank line. The next line contains the name / description of each column data. Each of the next 31 lines corresponds to one molecule. The first column contains the code / name of the molecule, the second the biological activity (Y) as measured experimentally and the next 53 columns the vector describing the molecule. The vector consists of molecular features that have been mainly calculated from the molecular structure. Where the biological effect occurs as <-1* it is considered as -1. (Taken from project description). The class label is considered the "Activity" attribute.

Fish Market Dataset

Source:

Fish Dataset

Description:

This dataset is a record of 7 common different fish species in fish market sales. The dataset consists of 1 nominal attribute called "Species" which denotes the type of the fish, and 6 numerical attributes which hold information about different measurements for each fish. "Weight" is the weight of the fish in grams, "Length1" through "Length3" are vertical, diagonal, cross lengths in cm accordingly, "Height" is the height of the fish in cm and "Width" is the diagonal width of the fish in cm as well.

This is a regression problem, because the target label we are trying to predict is the "Weight" attribute of the fish, which is a continuous numerical attribute.

Pre-processing

For both datasets I removed the nominal attributes and normalized all numerical attributes to a range of 0-1. The pre-processing part was done with the help of RapidMiner for convinience purposes. (I also provided the .rmp files for the processes)

Algorithms

RBFRegressor

Hyperparameters

- NumFuctions -> 6
- Ridge -> 0.1
- Tolerance -> 1e-5
- UseAttributeWeights -> true
- UseCGD -> true
- UseNormalizedBasisFunctions -> true

MLPRegressor

Hyperparameters

- NumFuctions -> 6
- Ridge -> 0.1
- Tolerance -> 1e-5
- UseCGD -> true
- ActivationFunction -> ApproximateSigmoid()
- LossFunction -> ApproximateAbsoluteError()

RandomForest

RandomForest consists of multiple decision tree models, which are then combined to deliver a more accurate prediction.

Hyperparameters

- BreakTiesRandomly -> true
- ComputeAttributeImportance -> true
- MaxDepth -> 4
- NumIterations -> 100

All hyperparameters are chosen with simplicity in mind and were the results of testing and experimenting to get the lowest possible error. The choices were mostly educated trial and error combined with manual hyper-parameter tuning.

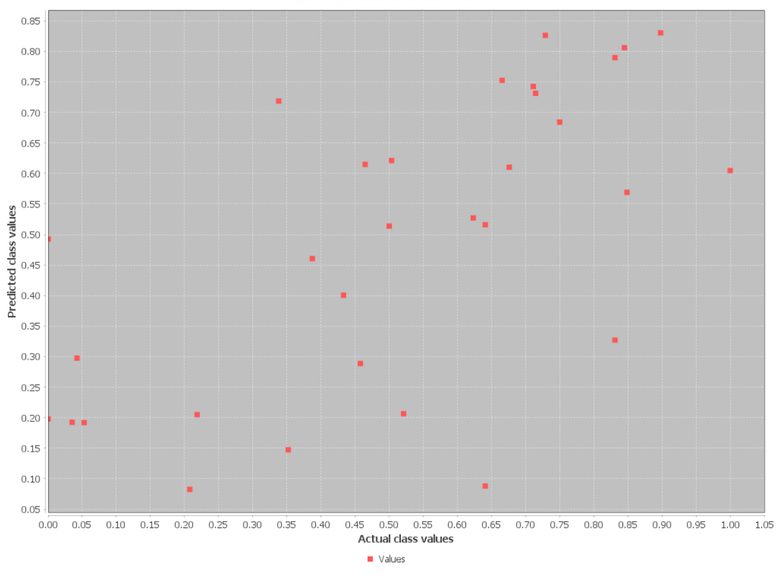
Results

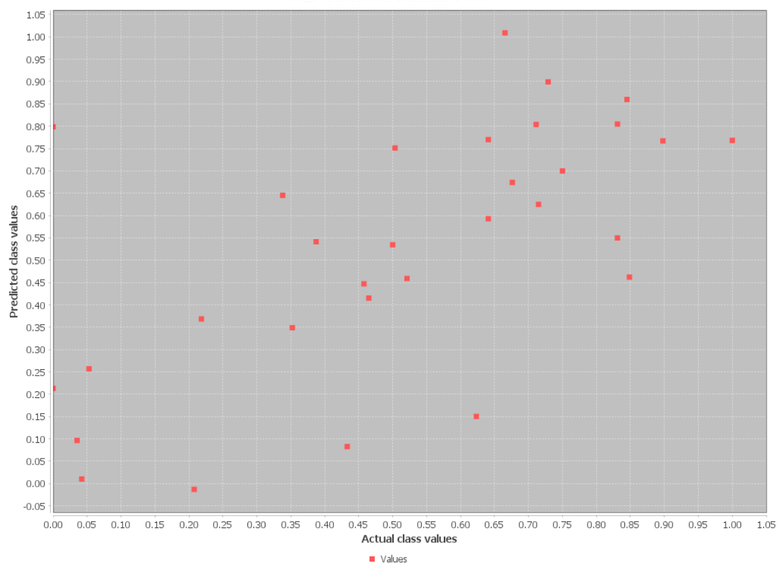
Selwood (10 - Fold Cross Validation)

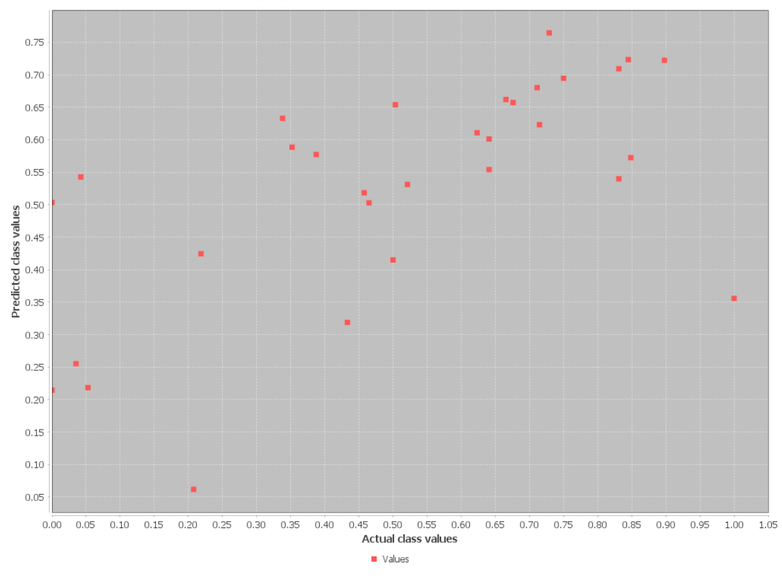
Selwood	RBF	MLP	RandomForest
Correlation coefficient	0.6397	0.6301	0.6192
Mean absolute error	0.171	0.1732	0.1658
Root mean squared error	0.2285	0.2423	0.2255
Relative absolute error	69.7802 %	70.662 %	67.6702 %
Root relative squared error	77.721 %	82.4186 %	76.7174 %
Total Number of Instances	31	31	31

Plots (Actual with Predicted Values)

RBF	MLP	RF
RBF Selwood Scatter Plot	**MLP Selwood Scatter Plot**	**Random Forest Selwood Scatter Plot**





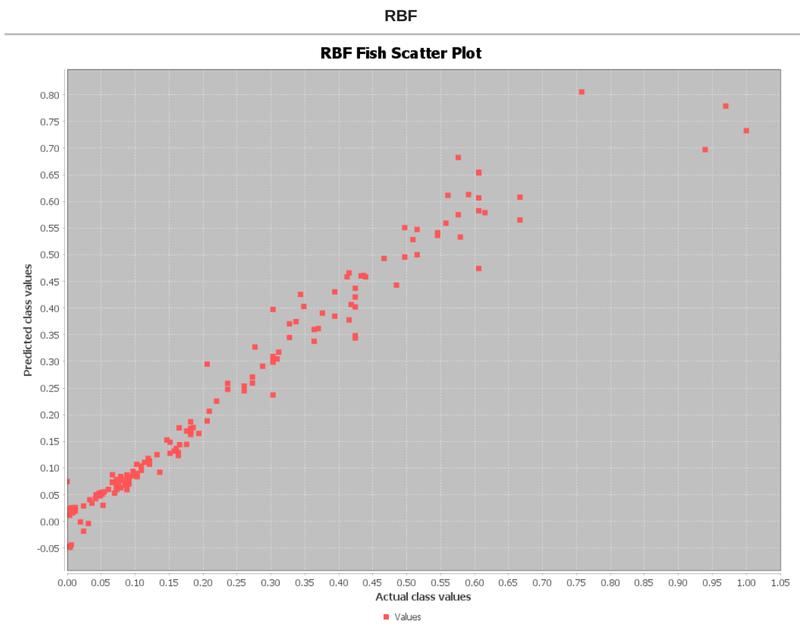


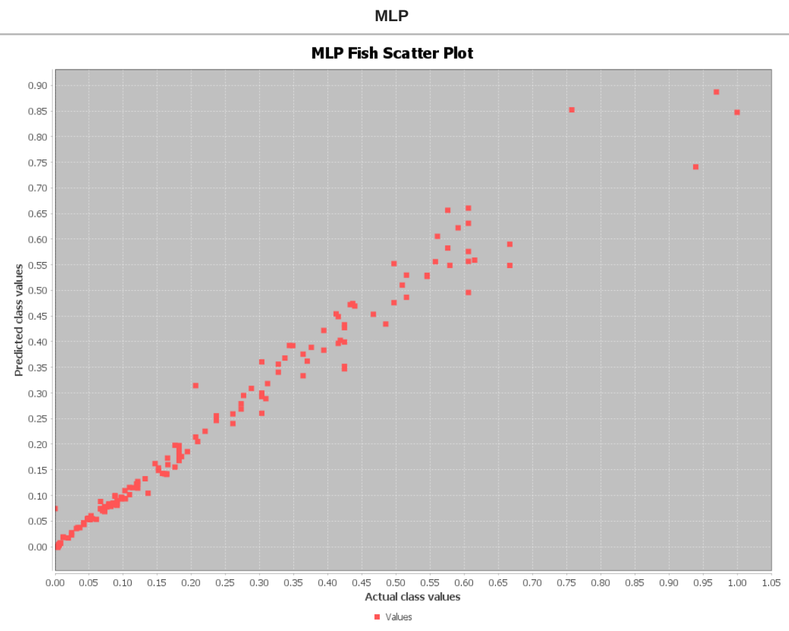
Fish (10 - Fold Cross Validation)

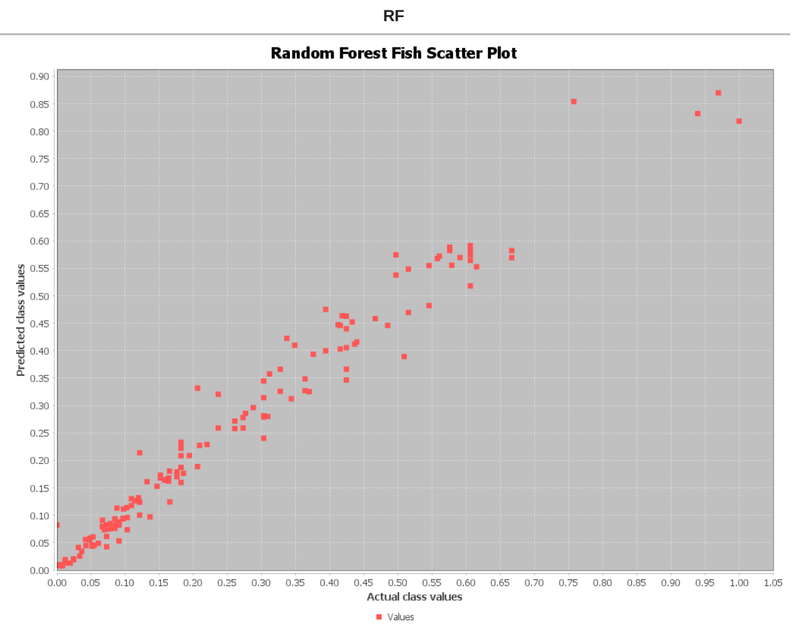
Fish	RBF	MLP	RandomForest
Correlation coefficient	0.9783	0.987	0.9844
Mean absolute error	0.026	0.0198	0.025
Root mean squared error	0.0455	0.0353	0.0389
Relative absolute error	14.4055 %	10.9631 %	13.8852 %
Root relative squared error	20.9037 %	16.2309 %	17.897 %
Total Number of Instances	159	159	159

Plots (Actual with Predicted Values)

RBF	MLP	RF
RBF Fish Scatter Plot	**MLP Fish Scatter Plot**	**Random Forest Fish Scatter Plot**







Observations

In the selwood case RandomForest performed slightly better than the other 2 algorithms, while between the other 2, the performance was close. In the Fish dataset all 3 of them performed exceptionally. Also, the small number of instances and the high dimensionality in the selwood dataset, made the training of the models really hard.

I tried applying attribute selection and even feature engineering to the dataset, but didn't manage to get any better results.

Conclusions

The choice of hyperparameters plays a big role in the performance of regression algorithms, and our choices must precise, even more when our data is very little. In addition, datasets must contain an adequate number of examples in order to correctly train our models, which largely depends on the number of parameters that are generated. Lastly, very large dimensionality in a dataset makes the training algortihm's adaptation (fit) much harder, with oftentimes hindering our results or making the training process harder.