KNN is the Best Classifer
Random Forest is the Best Regressor

Girish kumar Kadapa Darshan Dhananjay Aravind Ashoka Reddy Gursimran Singh

Abstract

This report investigates two questions. First, for a given selection of data sets, can we say what is the 'best' classifier or the 'best' regressor in terms of good predictions? How much does the answer depend on the particular selection of data sets? How much does the answer depend on our computational constraints? We investigate these questions using data sets from the UCI repository. Second, we compare the interpretability of a decision tree classifier to that of a convolutional neural network. We compare the decision tree visualization to 'activation maximization', a technique to gain insight into the kinds of inputs that deep neural networks respond to.

1. Introduction

In this project classification, regression and the Convolution Neural Network was used based upon the data-set provided. The main aim of this project was to evaluate the performance of the models and also provide a comparison on those models. By using various techniques for finding the best hyper-parameters for the models a good prediction and accuracy score was achieved. Convolution neural network model is used for the image data-set which as it provides shift-invariant , rotational-invariant and space invariant.

Novelty component is used for one for the credit card dataset where we deep dived into the dataset and looked upon the correlation of the features on the result label. Combination of features was done to create a new feature which would impact more on the label and has high correlation to that.

The final models selected for classification and regression are being done on different metric which will help to determine the good classifier as their cannot be a single best model for a particular data-set. The metric used in the classification are Confusion matrix, Precision , Recall , f1score and accuracy, whereas for regression the root mean square and R2 score has been treated as the metric measured. By using these metric the conclusion was being made on the models that perform well on which data-set.

2. Methodology & Experimental Results

The methodology for selecting best models for classification are different . For the classification models 4 performance metrics (F1Score, Precision, Recall and accuracy) has been used which are standard and being used by many researchers in their experiments. For the classification models R2 score and Root Mean square is being used for evaluating the models.

2.1. Classification Experiments

For classification these are the following data-sets for which different models are being evaluated.

1. Diabetic Retinopathy [2]
2. Default of Credit card Clients [16]
3. Breast Cancer Wisconsin [4]
4. Statlog (Australian Credit approval) [9]
5. Statlog (German Credit approval) [9]
6. Steel Plates Faults [5]
7. Adults [9]
8. yeast [9]
9. Thoraric Surgery Data [17]
10. Sesismic-Bumps [14]

Data pre processing is being done on some of the dataset as there were some missing values. Once the all the na and null values was removed or replaced with the mean or median , the normalization was performed on the dataset to bring down the scale. Different models are used such as Support Vector Machines, Decision Tree, Logistic Regression, K Nearest Neighbours, Random Forest, Ada-Boost, Gaussian Naive Bayes and Neural Networks and the evaluation of these models on-above given data-sets are based upon 4 metrics F1Score, precision, recall and accuracy. In our experiment we have used GridSearchCV for finding the best hyperparamters for the model. By comparing all the models with their best hyperparameters , the metric result are used and final classification technique is being selected.

2.2. Regression Experiments

For Regression, these are the following data-sets for which different models are being evaluated.

1. Wine Quality [7]
2. Communities and Crime [12]
3. Parkinson Speech [13]
4. Facebook Metrics [11]
5. Bike Sharing [10]
6. Student performance [8]
7. Concrete Compression Strength [15]
8. SGEMM GPU Kernel performance! [3]
9. Merck Molecular Activity Challenge [1]

Data are preprocessed as some of the data-sets has null and '?' mark in their attribute and are replaced with the mean and in some cases these rows are removed.Normalization on the data is being done to bring down the scale for all the features. Different models are used such as Support Vector Regression, Decision Tree Regression, Random Forest Regression, Ada-Boost Regression, Gaussian Process Regression, Linear Regression and Neural Networks Regression and the evaluation of these models on-above given data-sets are based upon 2 metrics R2 Score and Root Mean Square error. In our experiment we have used GridSearchCV for finding the best hyperparamters for the model. By comparing all the models with their best hyperparameters, the metric result are used and final regression technique is being selected.

2.3. Interpretability Experiments

The compiled images that were provided from the source website were little distorted. Firstly we tried to augment the data by using different type of transformation such as shifting, rotating the images but the accuracy of the model was not improving substantially. Then second approach was to run by increasing the number of filters to high number such as 512 and compensated the complexity of the model by using random dropout. Performing this kind of approach had a good advantage as this resulted in higher accuracy.Later we combined both the techniques and achieved a greater accuracy then the earlier two models. Looking at the accuracy results , it was a indication that it depends on how much the images are different from each other, so that all forms of the images are being trained and also the the number of convolutions layer and hidden layer play a vital role in the getting the accuracy.

3. Conclusions

At the end after performing all the different model experiments on each of the datasets, It has come to the conclusion that for classification the machine learning models that are are being used completely different from the regression models, on the datasets that was provided KNN was one of the good algorithms that performed well on almost all the dataset compared to all other models, taking the F1-Score, Precision and recall metric as reference, where as in regression it is random forest that outperformed on all the other in most of the datasets, taking the R2-Score metric as reference.For CIFAR dataset it was the role of the convolution and the hidden layer (architecture) that plays a vital role in getting the accuracy of the model. Novelty component

A. Detailed experimental results

The full comparison of each of the data set is being done on each of the model and the final table is shown below, for each of the dataset the best model and the f1-score, Precision, recall, training Accuracy and test accuracy is being displayed in table 1 below for classification.

For the regression the same procedure is followed but the metric is different from the classification in this R2-Score is being displayed for the best model in the table 2.

The techniques used for getting the best hyperparameters for the model is by using the GridSearchCV method. This method iterates overall the parameters range that was provided and then best parameters are being chosen which provides a good accuracy. For some of the datasets training took more than 3 minutes to train the model as the data was huge.

For CIFAR dataset the data was trained using the different architectures of the CNN by using various combination of Convolution Layers and the hidden layers Neural Network. The final model was selected based upon the accuracy on the validation set and the test set.The accuracy for our model was around 70 percentage.

B. Overview of project code and data

To execute the code the datasets files must be placed in the same directory as that of the python files. For the classification algorithms the file with the name classification must be executed to run all the datasets for all the machine learning models and finally the results with the best report and the final report is displayed with the details. For regression its the same procedure the filename with regression must be executed during the execution the results of best hyperparameters and r2 Score is being displayed after each model for each dataset.

For CIFAR experiment the data batch file must be placed must be in the same folder and also the model that needs to be loaded in case the training is taking much time to train.

Dataset - Best Algorithm	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score
Diabetic Retinopathy [2] - KNN	0.73	0.63	0.63	0.63	0.63
Breast_Cancer_Wisconsin [4] - KNN	0.98	0.98	0.98	0.98	0.98
Thoracic Surgery Data [17] - Random Forest	0.84	0.85	0.83	0.85	0.80
Seismic-Bumps [14] - Logistic Regression	0.93	0.93	0.90	0.93	0.90
Steel_Plates_Faults [5] - KNN	0.85	0.75	0.76	0.75	0.75
Australian Credit approval [9] - Random Forest	0.89	0.89	0.89	0.89	0.89
German Credit approval [9] - KNN	0.78	0.80	0.79	0.80	0.78
Default of Credit card Clients [16] - Neural Network	0.81	0.81	0.80	0.81	0.79
Adult [9] - Neural Network	0.81	0.81	0.80	0.81	0.79
Yeast [9] - Decision Tree	0.99	0.62	0.60	0.62	0.60

Table 1. Best Classification Algorithm for each dataset.

Dataset - Best Algorithm	Train Accuracy	Test Accuracy	R2 Score	RM square
SGEMM GPU kernel performance [3] - Random Forest	0.73	0.63	0.63	0.63
Merck Molecular Activity Challenge [1] - Random Forest	0.98	0.98	0.98	0.98
Concrete Compressive Strength - SVM Regression [15]	0.96	0.83	0.83	6.476385e+00
Wine Quality [7] - Random Forest	0.55	0.41	0.41	6.326893e-01
Communities and Crime [12] - Support Vector Regressor	0.73	0.63	0.63	1.455227e-01
QSAR Aquatic Toxicity [6]- Random forest	0.72	0.54	0.54	1.256615e+00
Facebook metrics [11]- Decision tree	0.3	0.36	0.36	3.987286e+02
Bike Sharing [10]	0.81	0.81	0.80	0.81
Parkinson Speech [13] - Decision tree, Random forest	1.00	1.00	1.00	0.00
Student Performance [8] - Linear regression	1.00	1.00	1.00	4.375351e-15

Table 2. Best Regression model for each data set.

References

[1] Merck molecular activity challenge, 2012. 2, 3

[2] Bálint Antal and András Hajdu. An ensemble-based system for automatic screening of diabetic retinopathy. *Knowledge-based systems*, 60:20–27, 2014. 1, 3

[3] Rafael Ballester-Ripoll, Enrique G Paredes, and Renato Pajarola. Sobol tensor trains for global sensitivity analysis. *Reliability Engineering & System Safety*, 183:311–322, 2019. 2, 3

[4] Gavin Brown. *Diversity in neural network ensembles*. PhD thesis, Citeseer, 2004. 1, 3

[5] Massimo Buscema, Stefano Terzi, and Marco Breda. Using sinusoidal modulated weights improve feed-forward neural network performances in classification and functional approximation problems.. *WSEAS Transactions on information science and applications*, 3(5):885–893, 2006. 1, 3

[6] Matteo Cassotti, Davide Ballabio, Viviana Consonni, Andrea Mauri, Igor V Tetko, and Roberto Todeschini. Prediction of acute aquatic toxicity toward daphnia magna by using the ga-k nn method. *Alternatives to Laboratory Animals*, 42(1):31–41, 2014. 3

[7] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009. 2, 3

[8] Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. 2008. 2, 3

[9] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. 1, 3

[10] Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2-3):113–127, 2014. 2, 3

[11] Sérgio Moro, Paulo Rita, and Bernardo Vala. Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. *Journal of Business Research*, 69(9):3341–3351, 2016. 2, 3

[12] Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002. 2, 3

[13] Betul Erdogdu Sakar, M Erdem Isenkul, C Okan Sakar, Ahmet Sertbas, Fikret Gurgen, Sakir Delil, Hulya Apaydin, and Olcay Kursun. Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*, 17(4):828–834, 2013. 2, 3

[14] Marek Sikora and Łukasz Wróbel. Data-driven adaptive selection of rule quality measures for improving rule induction and filtration algorithms. *International Journal of General Systems*, 42(6):594–613, 2013. 1, 3

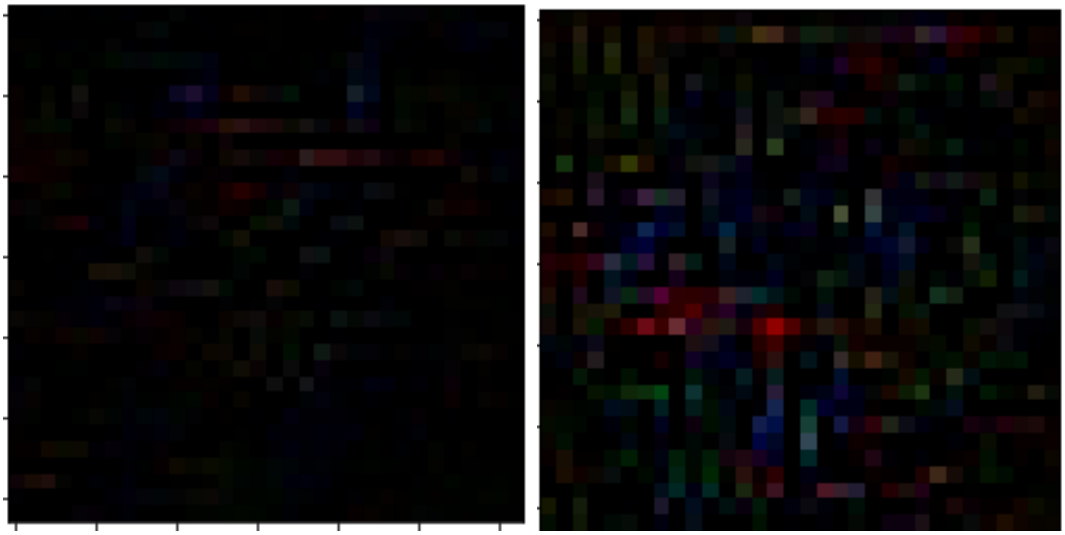

Figure 1. Images of Activation Maximation of Aero Plane(LEFT) and Car(RIGHT)

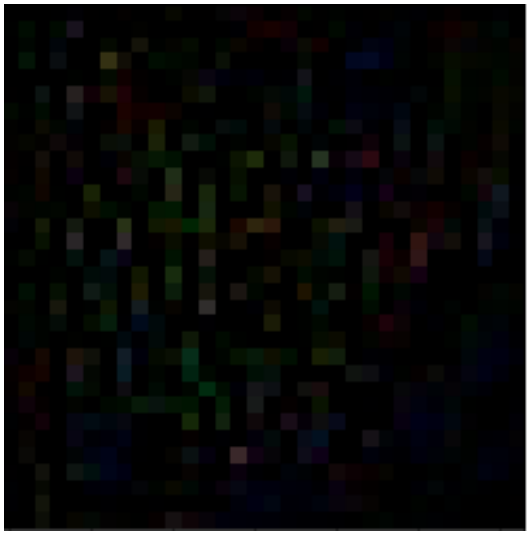

Figure 2. Image of Frog Activation Maximization.

[15] I-C Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research*, 28(12):1797–1808, 1998. 2, 3

[16] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009. 1, 3

[17] Maciej Zikeba, Jakub M Tomczak, Marek Lubicz, and Jerzy 'Swikatek. Boosted svm for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied Soft Computing*, 2013. 1, 3