A Comparative Analysis of Three Supervised Machine Learning Classification Methods employed on Three Binary-Class Datasets

Sia Khorsand skhorsand@ucsd.edu

University of California, San Diego

Abstract

This project conducts an empirical analysis and comparison of three well-known (supervised) machine learning classifiers–Random Forests, Support Vector Classification/Machines, and Logistic Regression–by employing them on three different datasets from the UCI Machine Learning Repository. To get the best results, model parameters, including train-test splits, were fine tuned and results were cross-validated as a necessary step to keep track of training/testing accuracy and log loss/error. The results of the study suggest that all three of the models can produce good results and their rankings remain relatively consistent throughout the different datasets. Overall, although all models produced really accurate results along with low log loss reports, Logistic Regression seems to be the best option for binary-classification datasets considering its high performance in balance with its simple implementation method and good runtime.

1. Introduction

Machine learning model selection is an important step in a study, or production of a predictive classification result. The right model to choose could highly depend on the type of dataset it is being trained on. Due to the supervised nature of the algorithms chosen for this project, the data has to be cleaned, converted to numerical values, scaled, and relatively balanced for optimal results. After this step, the data is ready to be processed. Each of the models examined in this study works differently in its own way to achieve the same goal. Random Forests, Support Vector Machines, and Linear Regression are all state of the art methods for regression analysis and classification. The goal of this study is to employ each model, fine tune its parameters, and experiment with different test-train ratios to get a sense of the comparison of the performance of each mode on different datasets. The results are then cross-validated to better tune the parameters for even better and more consistent results. The training and testing accuracy and log loss are also reported for each classifier, with the addition of visualizations, to help better understand the performance of the models in comparison to each other. All in all, this methodology ensures that the reported results are consistent over different times.

2. Methodology

Algorithms and Metrics

The entire study is done using Python and its *scikit-learn* library. Visualizations are made to help with readability and understanding using the *matplotlib* library. The models chosen for this study are Random Forest(RF), Support Vector Machine(SVM/SVC), and Logistic Regression(LR).

Firstly, Random Forest is a tree-based supervised learning algorithm that works by building ***n*** number of decision trees using random subsets of the data. The predictions of each tree are then combined to ensure higher accuracy and minimum over-fitting. This makes it ideal for a classification study like this one. The parameters tuned for this study include: **max_depth**: maximum depth

of each tree in the forest, **n_estimators:** the number of trees in the forest,
criterion: the function to measure the quality of each split, **min_samples_split:** the minimum number of samples for each split to happen, **min_samples_leaf:** minimum number of samples to be at a leaf node, and **max_features:** number of features considered for a split. Different values for each of these parameters were considered and tested based on each dataset to get the right results.
Secondly, Support Vector Machines work by identifying the boundary that best separates classes with the maximum margin considered. This makes the model relatively complex. For this model, the parameters **C:** regularization value affecting bias and variance, **kernel:** specifying how the algorithm handles the relationships in the data, **gamma:** the coefficient of the kernel that affects the variance and complexity of patterns, and **probability:** computes probabilities.
Lastly, Logistic Regression is a supervised learning algorithm that predicts the probability of an input belonging to a specific category by applying a logistic function to linear combinations of its features. **C** is used in this model as well as **max_iter:** deciding the max iterations that the model runs in order to converge, **solver:** optimization algorithm, and **penalty:** specifying the regularization method.
To employ each dataset and classifier, GridSearchCV is used to search over the hyper-parameters and cross validation to select the best parameters for each model. The cross validation number used in this specific study was 10, meaning that the validation is 10-fold. Once the best parameters are identified, the best performance of each model becomes known. The next step is to experiment with different train-test sizes. For this experiment, the train test sizes were (20/80, 50/50, 70/30, 80/20, 90/10). To ensure reporting consistent results, the results for each split are added up and averaged out. To record accuracy, log loss is calculated along with the accuracy of each testing and training set.

Datasets

The study focused on three datasets from UCI ML Learning Repository. The models were, in order of analysis, **Credit Approval**, **Tic-Tac-Toe**, and **Breast Cancer** datasets. The Credit approval dataset consists of 690 instances with 15 attributes, both categorical and numerical, aimed to predict the approval or rejection of credit applications. Tic-Tac-Toe includes 958 instances representing various combinations of game states over 9 categorical features, for pattern recognition and strategic analysis, aimed to predict whether it is a winning move or not. Lastly, the Breast Cancer dataset consists of 699 instances and 9 features. With mostly numerical features, the dataset is used to distinguish between malignant and benign breast tumors. The datasets are purposely chosen in three different fields to demonstrate the usage of these supervised algorithms in the real world. Each dataset is cleaned, encoded, scaled, and balanced as needed to prepare for the processes of this experiment.

3. Performance Results

Each algorithm is split into 4 train-test splits and fine-tuned to perform well for that test size. Then, the results are recorded as shown below:

Credit Approval Data

Classifier	Split	Train Accuracy	Test Accuracy	Test Loss	Best Param
Random Forest	80/20	95.17%	97.86%	7.24%	depth: 20, leaf: 3, split: 5, Trees: 400
Random Forest	50/50	95.67%	95.43%	7.14%	depth: 20, leaf: 3, split: 5, Trees: 400
Random Forest	20/80	96.43%	95.54%	16.27%	Leaf: 5, split: 6
SVM	80/20	97.86%	96.43%	11.24%	C: 50, gamma: 0.01
SVM	50/50	95.11%	96.00%	11.07%	C:10, Gamma: 0.001
SVM	20/80	92.52%	94.46%	12.14%	C : 5, Gamma: 0.0005
LOGREG	80/20	95.88%	95.00%	11.53%	C: 10, Iter: 100
LOGREG	50/50	94.54%	95.14%	10.12%	C: 10, Iter: 150
LOGREG	20/80	97.86%	94.82%	13.33%	Iter: 200, C: 1.0

Tic-Tac-Toe Data

Classifier	Split	Train Accuracy	Test Accuracy	Log Loss	Best Param
Random Forest	80/20	96.86%	82.28%	18.67%	Depth: 10, Leaf: 3, split: 5, Trees: 200
Random Forest	50/50	92.27%	93.53%	20.49%	Depth: 10, Leaf: 3, split: 5, Trees: 200
Random Forest	20/80	89.21%	96.22%	20.36%	Leaf: 3, split: 5, Trees: 200
SVM	80/20	98.43%	96.86	5.96%	C: 10, Gamma:.01
SVM	50/50	98.12%	98.54%	6.2%	C: 10, Gamma:.01
SVM	20/80	96.89%	98.39%	9.24%	C: 10, Gamma:.01
LOGREG	80/20	97.43%	97.41%	8.31%	Iter: 75 C: 1.5
LOGREG	50/50	98.125	98.54%	10.5%	Iter: 75 C: 1.5
LOGREG	20/80	96.86%	98.39%	14.48%	Iter: 200 C: 1.5

Breast Cancer Wisconsin

Classifier	Split	Train Accuracy	Test Accuracy	Log Loss	Best Param
Random Forest	80/20	95.70%	97.85%	18.67%	Depth: 10, Leaf: 3, split: 5, Trees: 200
Random Forest	50/50	95.10%	95.43%	9.50%	Depth: 10, Leaf: 3, split: 5, Trees: 200
Random Forest	20/80	96.43%	95.54%	11.27%	Leaf: 5, split: 6
SVM	80/20	95.88%	96.43	11.21%	C: 10, Gamma:.01
SVM	50/50	95.11%	95.71%	15.19%	C: 10, Gamma:.01
SVM	20/80	97.52%	94.46%	11.14%	C : 5, Gamma: 0.0005
LOGREG	80/20	94.98%	97.41%	8.31%	Iter: 75 C: 1.5
LOGREG	50/50	97.54%	93.14%	13.09%	Iter: 75 C: 1.5
LOGREG	20/80	97.86%	94.82%	13.33%	Iter: 200, C: 1.0

4. Analysis and Conclusion

After employing Random Forest, SVM, and Logistic Regression on the three datasets, the results are consistent with expectations. All three classifiers performed well with high accuracy and relatively low log loss values. However, their performance fluctuates based on the train/test split ratios.

By a split margin, SVM outperforms the other two when considered overall, demonstrating the best balance between training/testing accuracies and log loss. For the credit approval dataset, SVM achieved the best performance, showing high accuracies and low log loss. Similarly, in the tic-tac-toe dataset, SVM and LR competed really closely for the best performance, while random forest underperformed due to the simplicity of the datapoint features. Also, in the breast cancer dataset, SVM again slightly edged out the other two by maintaining better balance and smaller gap between accuracies.

While SVM showed the most consistent results, this does not mean that RF and LR are less valuable for this use case. In alternative interpretations of the results, if things weren't measured as a whole and other factors were considered, LR remains one of the top choices due to its simplicity, and fast runtime, and strong results, especially for binary classification tasks with linear relationships. On the other hand RF is excellent for identifying non-linear features and their relationships, reducing overfitting, and showing importance of features.

In conclusion, all three models achieved very good performance results, with SVM slightly taking the edge by being most consistent. Its ability to minimize log loss and maintain balance throughout result metrics highlights the algorithm's ability to become more complex, while LR and RF still have valable use-cases.

References

Aha, David. "UCI Machine Learning Repository." *Archive.ics.uci.edu*, archive.ics.uci.edu/dataset/101/tic+tac+toe+endgame.

Breiman, Leo, and Adele Cutle. "Random Forests." *Www.stat.berkeley.edu*, www.stat.berkeley.edu/~breiman/RandomForests/.

Caruana, Rich, and Alexander Niculescu-Mizil. *An Empirical Comparison of Supervised Learning Algorithms*.

Chang, Chi-Chung, and Chi-Jen Jin. "LIBSVM -- a Library for Support Vector Machines." *Www.csie.ntu.edu.tw*, www.csie.ntu.edu.tw/~cjlin/libsvm/.

Quinlan, J.R. "UCI Machine Learning Repository." *Archive.ics.uci.edu*, archive.ics.uci.edu/dataset/27/credit+approval.

Wolberg, William. "UCI Machine Learning Repository." *Archive.ics.uci.edu*, archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original.