Applications of Supervised Machine Learning Algorithms in Finance and Health Science

Taylor Han

Department of Cognitive Science
University of California, San Diego

Abstract

Supervised machine learning algorithms have been acknowledged for demonstrating their strengths in predicting outputs given original inputs. This paper attempts to reproduce the results found by Caruana and Niculescu-Mizil [1] in their study. The datasets considered in this paper pertain to the fields of finance and health science, two of many fields that their success improves our daily life.

I. Introduction

Today, breast cancer is one of the most common cancer that affects women worldwide, with 252,710 invasive cases and taking the lives of 40,610 women in the United States alone in 2017 [2]. Breast cancer represents the majority of new cancer cases and cancer deaths, which makes it hard not to recognize it as a significant public health problem. It is widely understood that early diagnosis of breast cancer remarkably improves the prognosis and the chance of survival, for early treatments are given to the patients. Using the dataset from UCI Machine Learning Repository titled “Breast Cancer Wisconsin (Diagnostic) Data Set,” machine learning algorithms were trained to detect breast cancer from given array of inputs.

One financial fraudulent activity includes forging official currency and using them to purchase goods. The ill-effects that counterfeit money has on society include but not limited to [4]:

- Companies not being reimbursed, weakening their buying power
- Reduction in the value of real currency
- Unwanted inflation due to more currency getting circulated

Furthermore, credit card fraudulent activities hurt financial institutions and their users for identity theft. $24.26B was lost due to frauds and it is reported that unauthorized activity increased by 18.4% in 2018 worldwide [3]. To detect and prevent potential illegal activities, dataset from Kaggle titled “Credit Card Fraud Detection” was used to train machine learning algorithms.

II. Methods

A. Learning Algorithms

As an extension to the 2006 study by Caruana and Niculescu-Mizil, 4 different supervised machine learning classifiers were used to perform on 3 different datasets. Specifically, logistic regression (LGR), random forest (RF), decision tree (DT), and support vector machine (SVM) algorithms are explored using the 20/80 and 80/20 split methods. All of the classifiers except for support vector machine (SVM) were implemented with the Scikit-learn package. Additionally, 5-fold grid search cross validation method was used to optimize the hyper-parameters. The algorithms were ran multiple times until the best set of hyperparameters were discovered.

1) Logistic Regression (LGR)

LIBLINEAR and Newton-CG were used to implement the logistic regression classifier. The former was paired with an L1 penalty and the latter with an L2 penalty.

2) Random Forest (RF)

1024 trees and [0.05, 0.1, 0.2, 0.3, 0.4, 0.6, 0.8, 1] were used to represent the maximum size of feature. It was originally [1, 2, 4, 6, 8, 12, 16, 20], but errors were yielded to all datasets. Thus, each member was divided by 20 and proportional ratio was used to represent the original maximum size of feature. The Weka implementation was followed.

3) Decision Tree (DT)

The criterion used was entropy and the list of max_depth used were [1,2,3,4,5]. The specific parameters for the decision trees were filled in.

4) Support Vector Machine (SVM)

SVM was coded without using machine learning libraries for direct control of C values. C_list [0.1, 1, 10, 100, 1000] represented the list of tradeoff parameters. Higher parameter meant higher weight of training error and lower weight of generalization error in calculating the testing error.

B. Datasets

Table 1. Datasets Information

Dataset	***# of Attributes***	***# of Instances***
Breast Cancer	30	569
Credit Card Fraud	30	284,807
Bank Currency Note	4	1,372

1) Breast Cancer

The breast cancer dataset featured 30 attributes with 569 samples. Features are computed from a digitized image of a fine needle aspirate of a breast mass. They described characteristics of the cell nuclei present in the image. Some of the features are correlated and they were cross compared to the label (M = malignant, B = benign). A visualization of this is available in the appendix.

2) Credit Card Fraud

The credit card fraud dataset featured 30 attributes with 284,807 samples. It was important to note that the dataset is highly unbalanced as only 492 samples were classified as fraudulent, accounting for 0.172% of all transactions. To tackle this, three different training sets were explored:

- "Original Data": Using the training dataset as it is, resulting in the size of 284,807.
- "Half Training Size": Only using half of the original dataset, which only includes 293 samples that are classified as fraudulent. Then, similar number of randomized non-fraudulent classified samples were appended to the training dataset (300), resulting in training dataset total size of 593.
- "Full Training Size": Similar number of randomized non-fraudulent classified samples (500) were appended to the total number of fraudulent classified samples (492). This brought up the training dataset to the size of 992.

3) Bank Currency Note

The bank currency note dataset featured 4 attributes with 1,372 samples. While authentic currency note sample size was 610, fraudulent currency note sample size was 762. Adjusting the training dataset like for the credit card fraud dataset deemed unnecessary as those two groups were similar in size.

III. Experiments and Data

A. 20/80 Split & 80/20 Split

To obtain a good representation of result, two trials with different partitions of train and test set were performed. Then, the accuracy of each training, validation, and testing datasets were computed. The first trial (Partition 1) only used 20% of the training dataset to have the algorithms trained, reserving the remaining 80% for testing. The second trial (Partition 2) on the other hand used 80% of the training dataset to train the algorithms, leaving the remaining 20% for testing. All of the algorithms used 5-fold cross validation.

B. Partition 1 (20/80 Split) Accuracy Results

Table 1A: Breast Cancer (BC)

Classifier	***Train***	***Validation***	***Test***
LGR	0.91595543	0.91150442	0.938596491
RF	1.0	0.92035398	0.951754385
DT	0.95151731	0.91150442	0.907894736
SVM	0.91150442477	0.92121598	0.868421052

Table 1B: Credit Card Fraud (Original Data) (CCFOD)

Classifier	***Train***	***Validation***	***Test***
LGR	0.99907832	0.99891154	0.999148547
RF	1.0	0.99936799	0.999482106
DT	0.99969277	0.99917487	0.999429439
SVM	0.99935043275	0.99742131	0.999416272

Table 1C: Credit Card Fraud (Half Training Size) (CCFHT)

Classifier	***Train***	***Validation***	***Test***
LGR	0.94903846	0.86440678	0.91789473684
RF	1.0	0.87288136	0.89684210526
DT	0.90673077	0.87288136	0.891421321
SVM	0.97457627	0.84783151	0.84842105263

Table 1D: Credit Card Fraud (Full Training Size) (CCFFT)

Classifier	***Train***	***Validation***	***Test***
LGR	0.96057089	0.95425868	0.94339622641
RF	1.0	0.95268139	0.94968553459
DT	0.94321386	0.93690852	0.91194968553
SVM	0.94479495	0.95307635	0.93710691823

Table 1E: Bank Currency Note (BCN)

Classifier	***Train***	***Validation***	***Test***
LGR	0.99084249	0.98905109	0.98542805100
RF	1.0	0.98540146	0.98724954462
DT	0.9981685	0.94525547	0.97267759562
SVM	0.992700729	0.92642179	0.99180327868

Table 1F: Rank Order of the Classifiers

Dataset	***1st***	***2nd***	***3rd***	***4th***
BC	RF	LGR	DT	SVM
CCFOD	RF	DT	SVM	LGR
CCFHT	LGR	RF	DT	SVM
CCFFT	RF	LGR	SVM	DT
BCN	SVM	RF	LGR	DT

There were more instances of Random Forest having the highest testing accuracy across the datasets. There was really no other trend that could be spotted.

C. Partition 2 (80/20 Split) Accuracy Results

TABLE 2A: BREAST CANCER (BC)

Classifier	*Train*	*Validation*	*Test*
LGR	0.98899878	0.96043956	0.97368421052
RF	1.0	0.95604396	0.96491228070
DT	0.98569478	0.92747253	0.94736842105
SVM	0.940659340	0.91949023	0.95614035087

TABLE 2B: CREDIT CARD FRAUD (ORIGINAL DATA) (CCFOD)

Classifier	*Train*	*Validation*	*Test*
LGR	0.99923852	0.99921438	0.99919244408
RF	1.0	0.99953916	0.99952599978
DT	0.99961816	0.99941188	0.99945577753
SVM	0.999398714	0.99942121	0.99935044415

TABLE 2C: CREDIT CARD FRAUD (HALF TRAINING SIZE) (CCFHT)

Classifier	*Train*	*Validation*	*Test*
LGR	0.96057089	0.95425868	0.94339622641
RF	1.0	0.95268139	0.94968553459
DT	0.94321386	0.93690852	0.91194968553
SVM	0.94479495	0.93973174	0.93710691823

TABLE 2D: CREDIT CARD FRAUD (FULL TRAINING SIZE) (CCFFT)

Classifier	*Train*	*Validation*	*Test*
LGR	0.94640602	0.9407314	0.95477386934
RF	1.0	0.93568726	0.95979899497
DT	0.96469013	0.91424968	0.90954773869
SVM	0.94829760	0.92672078	0.92964824120

TABLE 2E: BANK CURRENCY NOTE (BCN)

Classifier	*Train*	*Validation*	*Test*
LGR	0.99179705	0.99088423	0.98545454545
RF	1.0	0.99179581	0.99272727272
DT	0.9958979	0.97812215	0.97818181818
SVM	0.9917958	0.97028379	0.98545454545

TABLE 2F: RANK ORDER OF THE CLASSIFIERS

Dataset	*1st*	*2nd*	*3rd*	*4th*
BC	LGR	RF	SVM	DT
CCFOD	RF	DT	SVM	LGR
CCFHT	RF	LGR	DT	SVM
CCFFT	RF	LGR	SVM	DT
BCN	RF	LGR	SVM	DT

TABLE 2G: CARUANA AND NICULESCU-MIZIL'S RANK

Dataset	*1st*	*2nd*	*3rd*	*4th*
N/A	RF	SVM	DT	LGR

VI. CONCLUSION

Although the findings of this paper did not completely have an accord with the findings laid out by Caruana and Niculescu-Mizil in 2006 (Table 2G), the random forest classifier supported such claims. There may be many explanations to discrepancies between the data shown in Table 1F, Table 2F, and Table 2G. They may include:

- The experiments conducted in this paper are relatively small compared to the 2006 study.
- The data explored in this paper are not similar to another.
- The data explored in this paper and the classification tasks are relatively simpler compared to the ones used in the 2006 study.
- Different libraries incorporated may show varying results, in addition to human error.

However, there were multiple conclusive findings from this paper. First, unbalanced dataset led to high accuracy in the testing time. This came as no surprise since having a lot more on one side of the label allows the classifiers to learn better and with accuracy. This phenomenon was demonstrated with the credit card fraud dataset; using the full original dataset, which is unbalanced, yielded close to 1 accuracy. As soon as the training dataset was adjusted and balanced, the accuracy started to drop.

Furthermore, having the higher ratio of training dataset compared to testing dataset resulted in higher testing accuracy. This was shown in the credit card fraud dataset as well; Table 1C shows the classification results of using 20% of training data while Table 2C uses the same training data but 80% of it. The higher testing accuracy in Table 2C was due to having more data to learn from and being tested less with the trained classifier. Although this phenomenon is generally accepted, it was not always the case that this trend held—the bank currency note classifiers were indifferent to the varying ratios.

To conclude, it was important to acknowledge that there is no "one-classifier-fits-all" classification approach to the dataset explored in this paper, let alone in Caruana and Niculescu-Mizil's study. The uniqueness of every dataset makes it necessary to try multiple classifiers and find the best hyperparameters through optimization to solve the question at hand.

REFERENCES

[1] Caruana R. Niculescu-Mizil A. (2006, June). An empirical comparison of supervised learning algorithms. The Proceedings of the 23rd International Conference on Machine Learning (pp. 161 – 168).

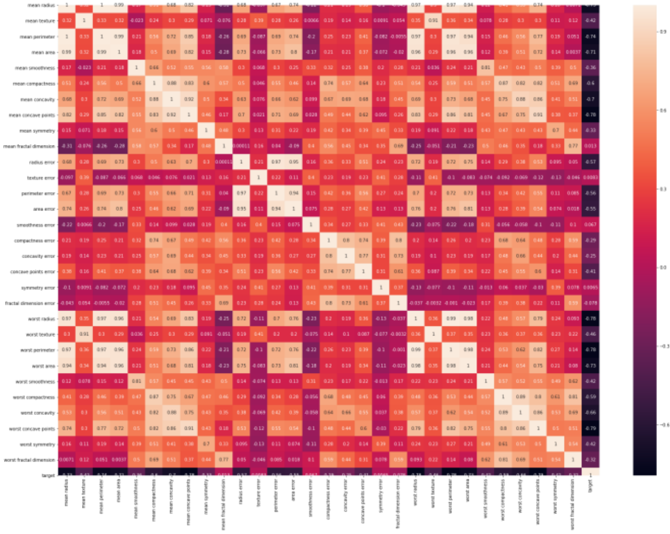
[2] https://shiftprocessing.com/credit-card-fraud-statistics/

[3] https://web.archive.org/web/20070616172012/http://wfhum mel.cnchost.com/counterfeiting.html

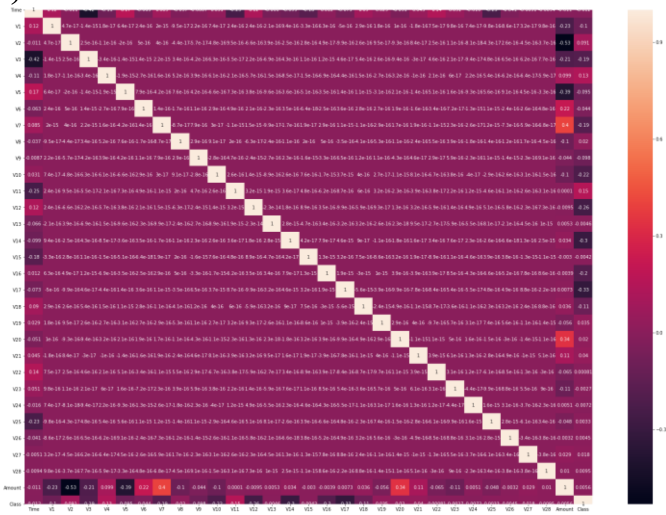
APPENDIX

A. Heatmaps of the dataset

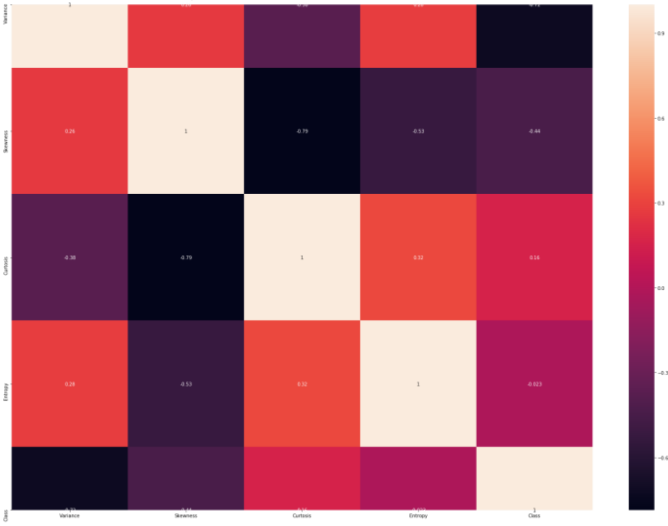
1) Breast Cancer Features



2) Credit Card Fraud Features

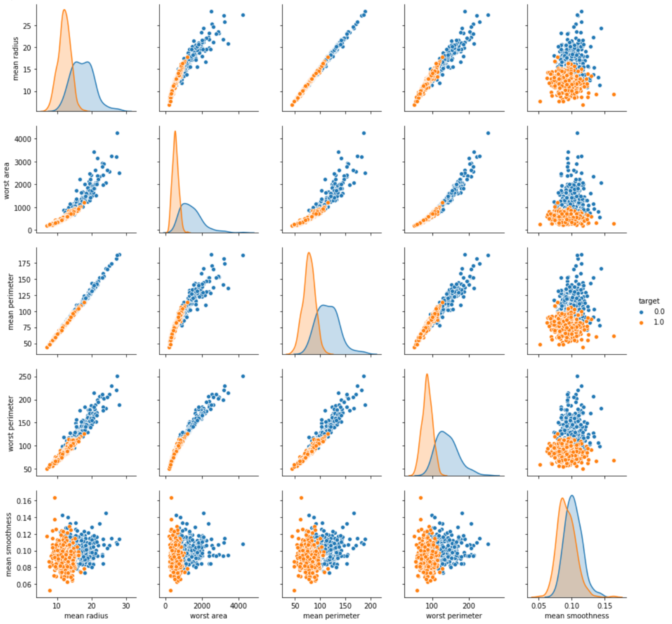


3) Bank Currency Note Features



B. Other Data Visualization

1) Breast Cancer Features Across Label



ACKNOWLEDGEMENT

The author of this paper would like to share a feeling of appreciation to Dr. Zhuowen Tu and his teaching team for introducing the concepts of multiple supervised machine learning algorithms, not limited to the ones used in this paper.