DISEASE CLASSIFICATION USING SUPERVISED MACHINE LEARNING METHODS

Po Hsiang Huang
Department of Electrical Engineering
University of California, San Diego
La Jolla, CA 92092
phh003@ucsd.edu

March 22, 2020

ABSTRACT

This paper explores several machine learning algorithms to solve classification problems on different diseases. Five supervised learning methods varying in complexity will be used to classify seven datasets of patients with different diseases varying in size. Comparison between the algorithm and datasets are to provide insight on disease classification using machine learning methods. An important aspect of my study is to see the performance trend of each classifiers on different disease datasets and determine whether or not there is an algorithm that is consistently better at diagnosing diseases given patient information.

1 Introduction

This paper presents the results of an empirical comparison of five supervised learning algorithms using the accuracy performance criteria. I evaluate the performance of SVMs, random forests, decision trees, k-nearest neighbors, and logistic regression on seven binary disease classification problems using an accuracy performance metric. Typical rule of thumb for medical classification is to use ROC, and this paper aims to find, if it exists, the best model for medical classification. For each algorithm I examine common variations, and thoroughly explore the space of parameters. For example, I cross validate each classifier on each dataset, compare at least five hyper parameters for each classifier and find the best parameter that yields the best results, etc.

The datasets are loaded and processed using the Python programming language through an interactive Jupyter notebook. Necessary Python libraries for the experiments include: numpy, pandas, scipy, sklearn, and matplotlib. The classifiers and datasets are used to obtain accuracy from three trials x five classifiers x seven datasets x three partitions. The results are summarized and compared in this paper to find if there is a single optimal supervised learning model for different diseases.

2 Methodology

2.1 Learning Algorithms

I pick five supervised learning algorithms recommended by the instructor. These classifiers have been seen working fairly well on a wide variety of data. All models are imported from sklearn. This section goes over each algorithm with brief discussion of their settings used to generate the result.

Support Vector Machine I use a linear kernel and vary the regularization parameter by factors of ten from 0.001 to 10. After five times cross validation using GridSearchCV, I again train the model on the optimal parameter and fit that best model on the testing data.

Random Forest I use an entropy criterion and vary the maximum depth by one ranging from 1 to 5. After five times cross validation using GridSearchCV, I again train the model on the optimal parameter and fit that best model on the testing data.

Decision Tree I use an entropy criterion and vary the maximum depth by one ranging from 1 to 5. After five times cross validation using GridSearchCV, I again train the model on the optimal parameter and fit that best model on the testing data.

K-Nearest Neighbors I use 5 values of K ranging from K = 1 to K = 11. I choose odd K values so that there is not a situation with tie votes. However, the value for some reason is capped at 12 and therefore does not accept other K values such as 15. After five times cross validation using GridSearchCV, I again train the model on the optimal parameter and fit that best model on the testing data.

Logistic Regression I train the logistic regression model using a liblinear solver because most of the datasets chosen do not have very large number of attributes and instances. I also vary the regularization parameter by factors of ten from 0.001 to 10. After five times cross validation using GridSearchCV, I again train the model on the optimal parameter and fit that best model on the testing data.

Table 1: Description of problems

Problems	Attributes	Instances	Positives
Breast Cancer	10	683	34.5%
Diabetes	20	1151	53.1%
Heart Disease	14	297	54.1%
Hepatitis	20	80	20.6%
Liver Patients	11	579	28.8%
Parkinson's	23	195	75.4%
Tumor	18	132	24.8%

2.2 Data Sets

I pick seven datasets from the UCI machine learning repository [3] with different diseases that people suffer from in order to identify the performance of each supervised learning algorithm on a variety of diseases. Each data is first prepared and reformatted to .arff using Weka for scipy to process. After loading each data and converting categorical labels and multiple labels into binary labels, I clean up all the points by deleting instances with points containing invalid data. I try deleting attributes that contain many invalid data, however, the dimension of some datasets is significantly reduced and therefore I determine deleting instances is a better choice. This section goes over how each dataset is prepared for classification. A summary of the datasets is provided in Table 1.

Breast Cancer Wisconsin Data Set [6, 7, 8, 13] This dataset contains 699 instances and 11 attributes. The Id attribute is removed because it is not relevant to the classification. 16 instances are removed because they contain invalid data. The original class labels 2 for benign and 4 for malignant are converted to 0 for benign and 1 for malignant.

Diabetic Retinopathy Debrecen Data Set [1] This dataset contains 1151 instances and 20 attributes. All data is complete, but labels are of object type and are therefore converted to float64, 0 for healthy and 1 for detected diabetes.

Heart Disease Data Set [5, 9, 10, 12] This dataset contains 303 instances and 75 attributes collected from four different databases. However, only instances from Cleveland and 14 relevant attributes are used for our problem. 6 instances are removed because they contain invalid data.

Hepatitis Data Set This dataset contains 155 instances and 20 attributes. It has missing data in several attributes, which after removal, yields only 80 instances with 20 attributes.

Indian Liver Patient Dataset This dataset contains 583 instances and 11 attributes. The sex attribute is of object type and is therefore converted to 0 representing male and 1 representing female. 4 instances are removed because they contain invalid data.

Parkinson's Disease [4] This dataset contains 197 instances and 23 attributes. 2 instances are removed because they contain invalid data. This dataset contains the highest number of features compared to all the other selected datasets.

Primary Tumor [11] This dataset contains 339 instances and 18 attributes. It has missing data in several attributes, which after removal, yields 132 instances with 18 attributes. This is a multiple categorical classification problem, so I simplify it to a problem classifying lung tumor vs all other tumors by converting all class labels that are not 1 to 0.

3 Experiments

3.1 Setup

Using a single Jupyter Notebook Python script, I prepare each dataset using the above mentioned methods. All datasets are then randomly split into a training set and a testing set with 80/20, 50/50, and 20/80 ratio to observe the impact of the size of both training and testing data. Each dataset is then trained on each classifiers with different parameters for three trials in order to reduce randomness and obtain an averaged result. For each trial, the validation accuracy is represented through heatmaps shown in the Appendix A and the best parameter found through GridSearchCV is used to fit the testing data.

3.2 Results

Table 2: Binary classification accuracy result (%) on **80/20** split. Accuracy averaged over three trials/repeats. Bold number represent the highest test accuracy seen in each dataset.

	Breast Cancer		**Diabetes**		**Heart Disease**		**Hepatitis**		**Liver Patients**		**Parkinson's**		**Tumor**	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
SVM	97.5	**96.8**	75.2	74.3	85.9	82.2	97.8	79.2	72.2	70.1	87.5	88.9	91.9	84.0
RF	98.4	96.4	71.6	62.5	90.3	78.9	96.6	**87.5**	78.2	69.3	98.5	**91.5**	93.2	84.0
DT	98.5	95.1	70.2	62.9	86.5	78.3	97.1	81.2	71.8	**70.4**	93.9	88.0	94.1	**90.1**
KNN	98.2	96.6	73.0	64.4	71.6	63.3	84.8	77.1	77.1	64.4	93.0	82.9	85.8	86.4
LR	97.3	**96.8**	75.7	**74.5**	86.2	**82.8**	94.4	79.2	72.5	70.1	86.0	86.3	89.4	82.7

Table 3: Binary classification accuracy result (%) on **50/50** split. Accuracy averaged over three trials/repeats. Bold number represent the highest test accuracy seen in each dataset.

	Breast Cancer		**Diabetes**		**Heart Disease**		**Hepatitis**		**Liver Patients**		**Parkinson's**		**Tumor**	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
SVM	97.1	96.6	76.3	73.6	84.9	**83.0**	92.1	80.8	72.7	70.1	88.9	85.7	87.5	83.3
RF	97.9	97.0	75.3	65.8	86.7	77.6	97.9	**86.7**	77.1	71.1	98.3	**86.7**	93.4	**86.9**
DT	96.8	92.6	65.2	63.1	86.7	71.6	98.3	80.8	73.7	70.6	91.6	83.3	87.0	80.8
KNN	97.9	**97.3**	72.9	64.5	74.1	65.8	85	82.5	87.0	65.6	91.5	78.6	86.7	80.3
LR	97.1	96.2	75.8	**74.5**	85.2	82.8	92.3	81.7	73.5	**71.5**	87.8	84.4	89.8	82.8

Table 4: Binary classification accuracy result (%) on **20/80** split. Accuracy averaged over three trials/repeats. Bold number represent the highest test accuracy seen in each dataset.

	Breast Cancer		**Diabetes**		**Heart Disease**		**Hepatitis**		**Liver Patients**		**Parkinson's**		**Tumor**	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
SVM	98.5	95.9	79.2	71.4	86.9	**75.8**	95.5	80.7	76.7	70.5	89.5	79.3	92.3	**80.5**
RF	99.3	95.1	78.8	62.4	91.9	**75.8**	93.7	**83.9**	87.5	70.2	96.6	**81.4**	91.7	77.0
DT	99.8	93.2	76.5	62.7	88.7	73.5	95.4	81.8	74.2	**70.9**	99.4	79.3	91.0	73.3
KNN	100.0	**96.0**	85.4	63.0	83.3	60.2	83.9	82.8	82.8	70.0	91.5	76.1	89.3	77.7
LR	98.8	94.1	79.4	**71.7**	86.7	75.4	86.9	80.2	73.9	70.7	84.4	75.2	88.8	77.0

3.3 Observations

Tables 2, 3, and 4 present all of the results from an experiment. We can easily observe the classifiers with highest accuracy on each dataset from the bold numbers. The performance rankings are also provided in Tables 5, 6, and 7. While one could say that we can determine the best classifiers from these tables, the testing accuracy of classifiers for each dataset and each split is actually very similar to each other. If one classifier performs poorly on a specific disease dataset, it seems that other classifiers also result in similar accuracy score.

One trend observed from these tables is that the performance generally improve as we split the dataset from 20/80 to 80/20, although this is not absolutely true in every case due to randomness when splitting. This is most likely due to more training data providing better classifiers.

Due to the scale of the datasets and number of classifiers, the program takes an average of approximately 1700 seconds (nearly 30 minutes) to complete. Therefore, it was rather difficult to make small adjustments and run the program multiple times. The preset parameter values, C_list in SVM and LR, D_list in RF and DT, and K_list in KNN, provide only a few parameters that the classifiers work well with.

Table 5: Rank order of the classifiers on (80/20) split in comparison.

Classifier	**1st**	**2nd**	**3rd**	**4th**	**5th**
Support Vector Machine	14.3%	57.1%	28.6%	0%	0%
Random Forest	28.6%	0%	28.6%	28.6%	14.3%
Decision Tree	28.6%	14.3%	14.3%	28.6%	14.3%
K-Nearest Neighbors	0%	14.3%	28.6%	0%	57.1%
Logistic Regression	42.9%	14.3%	14.3%	14.3%	14.3%

Table 6: Rank order of the classifiers on (50/50) split in comparison.

Classifier	**1st**	**2nd**	**3rd**	**4th**	**5th**
Support Vector Machine	14.3%	42.9%	14.3%	28.6%	0%
Random Forest	42.9%	28.6%	28.6%	14.3%	0%
Decision Tree	0%	0%	14.3%	42.9%	28.6%
K-Nearest Neighbors	14.3%	14.3%	0%	14.3%	57.1%
Logistic Regression	28.6%	14.3%	42.9%	14.3%	0%

Table 7: Rank order of the classifiers on (20/80) split in comparison.

Classifier	**1st**	**2nd**	**3rd**	**4th**	**5th**
Support Vector Machine	28.6%	42.9%	14.3%	14.3%	0%
Random Forest	42.9%	0%	28.6%	14.3%	14.3%
Decision Tree	14.3%	14.3%	14.3%	28.6%	28.6%
K-Nearest Neighbors	14.3%	28.6%	14.3%	14.3%	28.6%
Logistic Regression	14.3%	14.3%	28.6%	14.3%	28.6%

4 Conclusion

The performance results of the five classifiers present in this paper are approximately consistent with Caruana and Niculescu-Mizil's findings in An Empirical Comparison of Supervised Learning Algorithms [2]. The slight variations in the rankings observed from running the experiment a few times are due to randomly shuffling and splitting each dataset and are too small to consider significant changes in the rankings presented in Tables 5, 6, and 7. While we can say that KNN generally performs worse than other supervised learning models on the observed datasets, it is inconclusive if there is a single best model for medical classification. Larger and more datasets as well as additional classifier parameters need to be further experimented in order to determine if such best supervised learning algorithm exists for disease classification.

4.1 Summary Findings

- SVM, RF, DT, KNN, and LR all work well classifying different diseases.
- SVM, RF, and LR have the best and most consistent performances, while KNN on average performs the worst.
- There is no "one best" classifier for classifying different diseases because the results are all very close.

References

[1] Balint Antal, Andras Hajdu: An ensemble-based system for automatic screening of diabetic retinopathy, Knowledge-Based Systems 60 (April 2014), 20-27.

[2] Caruana, Rich Niculescu-Mizil, Alexandru. (2006). An Empirical Comparison of Supervised Learning Algorithms. Proceedings of the 23rd international conference on Machine learning - ICML '06. 2006. 161-168. 10.1145/1143844.1143865.

[3] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[4] 'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)

[5] Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

[6] K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).

[7] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[8] O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.

[9] University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

[10] University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

[11] University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia: M. Zwitter & M. Soklic

[12] V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D.

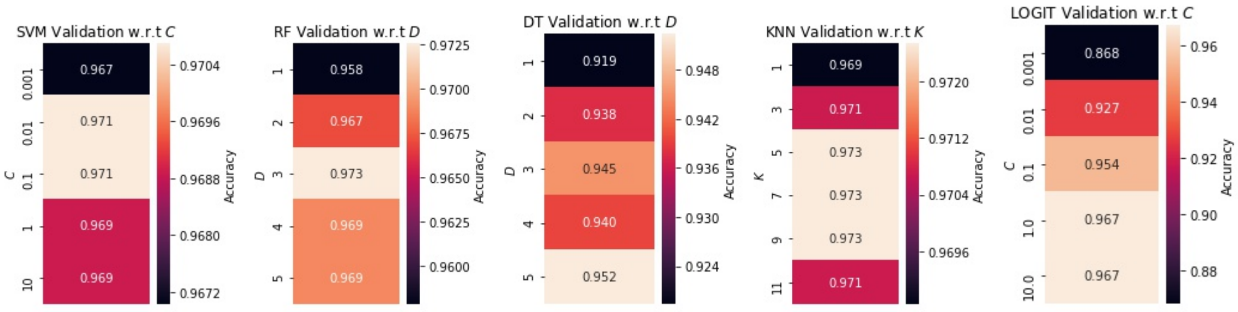
[13] William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

A Appendix

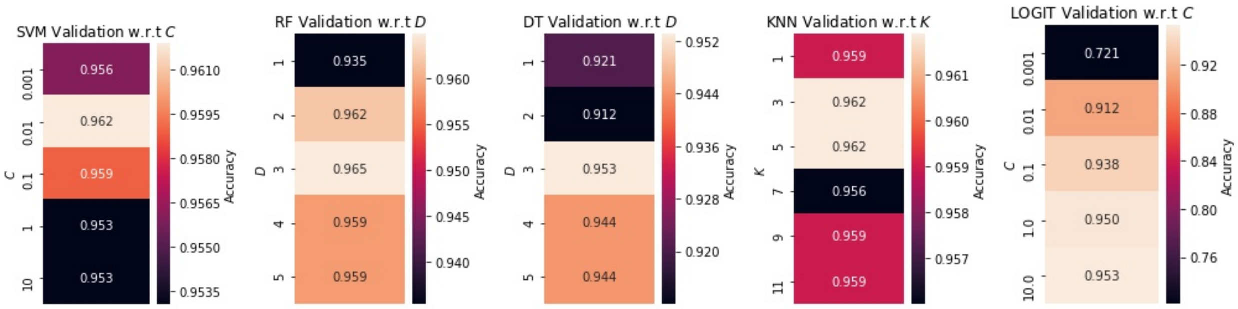
A.1 Validation Heatmaps

There are a total of 315 (3*5*7*3) validation heatmaps collected. 35 samples with 80/20 split from one trial are provided here.

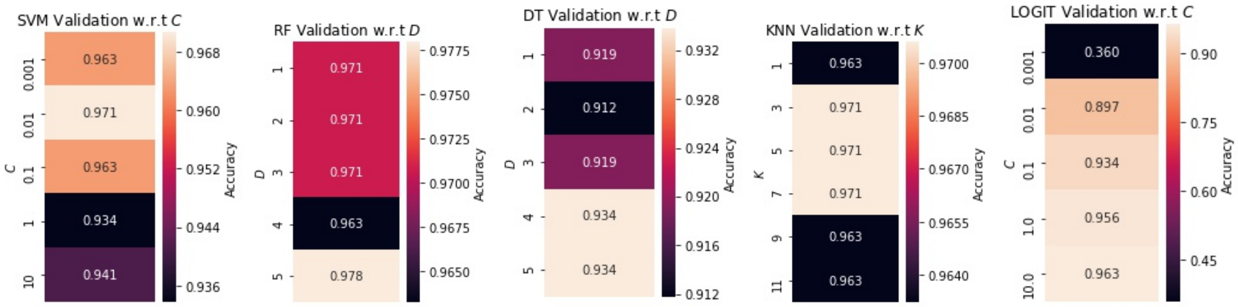
Breast Cancer



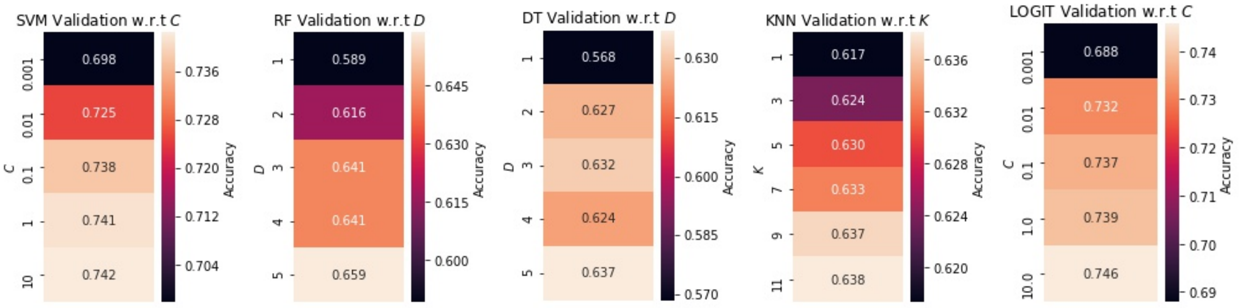
Diabetes



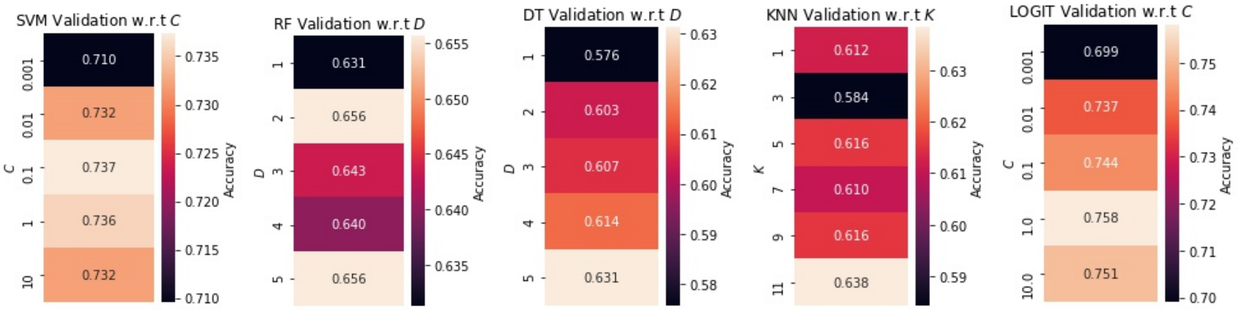
Heart Disease



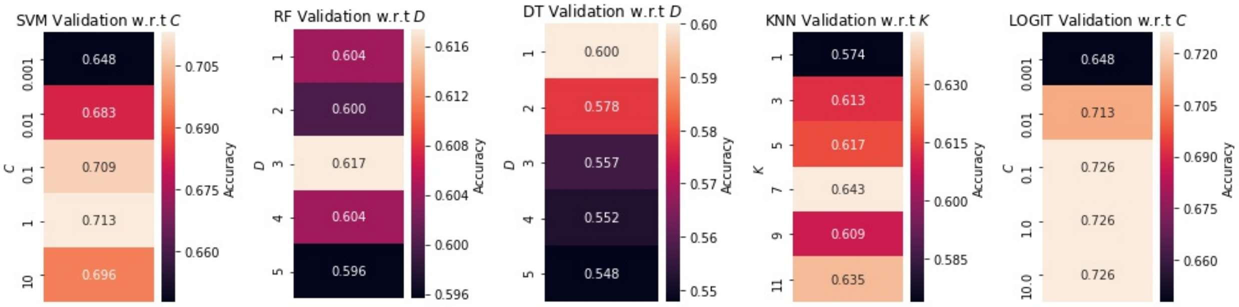
Hepatitis



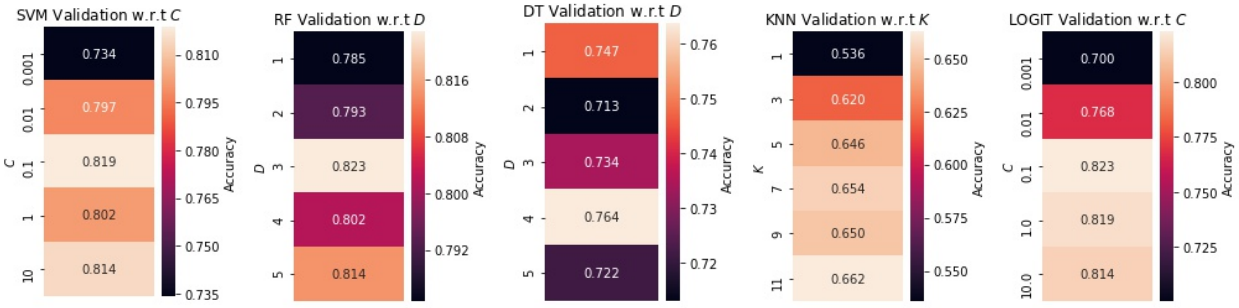
Liver Patients



Parkinson's Disease



Primary Tumor


SVM Validation w.r.t C
C
Accuracy
RF Validation w.r.t D
D
Accuracy
DT Validation w.r.t D
D
Accuracy
KNN Validation w.r.t K
K
Accuracy
LOGIT Validation w.r.t C
C
Accuracy