Generalized Heart Disease Classification: Significant Features and Best Models

William T. Speciale

Department of Computer Science, Saint Louis University

CSCI 4750: Machine Learning

Dr. Jie Hou

May 14, 2020

Abstract

There are two main goals of this study: the first is to determine which of the thirteen primary features (or some combination) of the dataset are the best predictors, and the second is which machine learning algorithm is able to make the best predictions. This project will use Google Colaboratory to organize its findings into a notebook. Data will be imported into the notebook directly rather than being saved locally. Each of the thirteen features in each dataset will be normalized and curated for outliers. The following algorithms will be trained and tested on the data: hinge SGD classifier, Random Forrest classifier, Decision Tree classifier, K Neighbors classifier.

Generalized Heart Disease Classification: Significant Features and Best Models

The data is comprised of four separate subsets, each with the standard thirteen features and labels. Each of these sets originally contained 75 features, but existing literature has focused in on the aforementioned fourteen. All of the sets are composed of individual patient data, although not all of the sets have values for every feature. Labels for these features refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Most experiments with the Cleveland database binarize the labels, instead distinguishing between presence (values 1, 2, 3, 4) from absence (value 0).

While most of the literature focuses on the Cleveland set, for this study all four sets were used. The Cleveland set is the most complete set, however it contains only three hundred patients. Since the focus of this study was to identify features that were strong general predictors, it was decided that pooling the datasets would provide more widely applicable results. This is also a relatively novel approach.

Other studies have used a swath of classification algorithms. Rather than preform an analysis of twenty or more models, a survey of four classifiers was chosen. Each of the five belongs to a different method of classification. In this way, the best style of classification for this dataset could be determined. This dataset is demonstrative of the challenge that healthcare data presents: a mixture of continuous and categorical features with categorical labels. A model that performs well will need to accommodate this. Models selected for this study were all sourced from the scikit-learn Python module (CITATION): hinge SGD classifier, Random Forrest classifier, Decision Tree classifier, K Neighbors classifier.

Methods

Data Retrieval and Processing

Raw data is retrieved from the UCI Machine Learning Repository using the pandas and saved as a csv inside a directory within the project. Each dataset (*processed.cleveland, processed.hungarian, processed.switzerland, processed.va*) is retrieved and saved separately. These set are then loaded as labeled pandas dataframes. Each set is then screened for missing values, and subsequently is fed through the sklearn iterative imputer. This imputer was chosen because of its ability to model each feature with missing values as a function of other features rather than an average value. Using such a technique could seriously distort feature significance and risks homogenizing the data. It should be noted that each of the raw sets is imputed separately, before being combined. This is so that the missing values are not influenced by peers outside of their set. In this way, the final combined dataset will retain distinct sub-trends representative of each of the four supplementary datasets. Once the imputation step is complete, the sets are merged to form a combined main dataset. The combined data is check for missing values and its labels are binarized.

Model Descriptions

The following are brief descriptions for each of the models used in this study. First is the Decision Tree Classifier which uses supervised learning to create a model that makes decisions based on “rules” derived from the features. This model most closely resembles human diagnostic processes. K-Nearest Neighbors Classifier is the second model: using instance-based learning to make its decisions, which is vastly different from current medical diagnostic methods. Model three is the SGD Classifier which utilizes the combination of a traditional linear model and stochastic gradient descent learning to optimize its coefficients. The final model is the Random Forest Classifier, an algorithm that uses a set of decision trees to make its predictions.

Refining Features for Set Formation

At this point in the pipeline, the features for number of major vessels colored (denoted *ca*) and *thal* were dropped. While these features were almost complete in the Cleveland set, in they are sparsely populated in the other sets and for this reason they were dropped. Some consideration was given to dropping the slope feature as well, however there were significantly more instances of this feature.

Once the content of the combined set was finalized it was split into training and testing sets using sklearn train_test_split. Here the data is both stratified and shuffled, with twenty percent of the instances being reserved for testing. There is no validation set reserved at this step, as these folds will be created and used later when selecting models. Since the data has both categorical and continuous features, a column transformer was used to preform transforms on specific features. Categorical features are subject to a MinMax scalar, and continuous features are subject to a MinMax scalar. This transform technique was chosen to minimize the impact raw

features scales would have on the model training, especially the K Neighbors model (which is particularly sensitive to continuous feature scale). The transformer is fit to the training data, and then it is used to transform both the training and testing features.

Model Evaluation Metrics

Evaluating the models was done in using two methods: balanced accuracy score and a classification report. The first metric measures average recall for each class and is a common evaluator for classification problems. Balanced accuracy was chosen over regular accuracy as it is able to account for imbalances in the data as well as penalize success due to pure randomness in performance. It's the raw accuracy where each sample is weighted by the inverse prevalence of its true class. Classification report was chosen as it provides results for a range of metrics including macro average, weighted average, and sensitivity. Using this will give a holistic view of each model, allowing for a more nuanced evaluation. Performance metrics common to classification tasks for medical purposes commonly use accuracy, sensitivity, and in the case of a multi-class problem specificity. Both of these metrics bear additional significance when choosing a model that is designed for diagnostic purposes.

Results

In order to find the best version of each model for this datasets sklearn GridSearchCV was used. Parameters for each model were prepared, although these lists to not cover all possible model hyperparameters. Choosing a subset of parameters was done to reduce training time and to prevent a highly specialized model from skewing the results. The grid search was then performed for each of the models using the training data. Each search featured ten-fold cross validation with

the models being evaluated using balanced accuracy. Best estimator parameters for each model can be found in Appendix Tables 3-6.

Balanced accuracy scores for each of the chosen models are presented in Table 1.

Table 1

Balanced Accuracy Scores for Final Models

Model	Balanced Accuracy Score
Decision Tree	0.813
KNN	0.791
SGDClassifier	0.785
Random Forest	0.856

By this metric, the Random Forest model leads the field by a modes margin of 4%. It is unsurprising that the SGD Classifier performed the worst, as it is does best on linear datasets. While it clearly is the worst performer, all models are within $\pm 6\%$ of one another, which is a good indicator that a second evaluation metric should be explored.

The second evaluation metric used in this study was the classification report. Presented in Tables 2-5 are the results for the each of the models.

Table 2

Decision Tree Classifier Classification Report

Type	precision	recall	f1-score	support
0 (no disease)	0.793	0.793	0.793	82
1 (disease)	0.833	0.833	0.833	102
accuracy	0.815	0.815	0.815	0.815
macro avg	0.813	0.813	0.813	184
weighted avg	0.815	0.815	0.815	184

Table 4

K-Nearest Neighbors Classifier Classification Report

Type	precision	recall	f1-score	support
0 (no disease)	0.768	0.768	0.768	82
1 (disease)	0.814	0.814	0.814	102
accuracy	0.793	0.793	0.793	0.793
macro avg	0.791	0.791	0.791	184
weighted avg	0.793	0.793	0.793	184

Table 5

SGD Classifier Classification Report

Type	precision	recall	f1-score	support

0 (no disease)	0.707	0.854	0.773	82
1 (disease)	0.859	0.716	0.781	102
accuracy	0.777	0.777	0.777	0.777
macro avg	0.783	0.785	0.777	184
weighted avg	0.791	0.777	0.778	184

Table 6

Random Forest Classifier Classification Report

Type	precision	recall	f1-score	support
0 (no disease)	0.850	0.829	0.840	82
1 (disease)	0.865	0.882	0.874	102
accuracy	0.859	0.859	0.859	0.859
macro avg	0.858	0.856	0.857	184
weighted avg	0.859	0.859	0.859	184

For each model, the f-1 score and accuracy in the classification report closely matches its corresponding balanced accuracy score, indicating that these evaluations are indeed accurate. Precision and recall scores for the Decision Tree and KNN models were within $\pm 4\%$ of one another, while the scores for the SGD and Random Forest models differed by $\pm 17\%$ and $\pm 5\%$ respecivly. The classification report results for the SGD model clearly demonstrates the struggle to linearly separate the data as it precision and recalls scores had the largest separation. Based on these results, it is reasonable to conclude that the Random Forest model was by far in a way the best performing model of those surveyed this study.

This study was also interested in the significance of individual features in their usage for the prediction of heart disease. Based on correlation matrix evaluations of the dataset and prevailing medical knowledge, the strongest predictors are exercised induced angina (chest pain during or as a result of exercise), age and sex as presented in Table 1.A. Since the best performing model was the Random Forest, its feature importance were extracted and are presented in Table 7.

Table 7

Random Forest Feature Importance

Feature Number	Feature Name	Importance Score
6	restecg	0.149
8	exang	0.131
2	cp	0.128
3	trestbps	0.118
5	fbs	0.113
0	age	0.096
4	chol	0.089
1	sex	0.081
7	thalach	0.037
10	slope	0.031
9	oldpeak	0.028

Importance scores from differ significantly from the correlation matrix scores. The highest correlating features were *cp*, *exang*, *oldpeak*, *slope*, and *sex*. In contrast, the features with the highest importance scores were *restecg*, *exang*, *cp*, *trestbps*, and *fbs*. It was surprising to see that only two of the five features were shared between these two metrics. Most notable is the feature *restecg* which had a very weak correlation (0.06) but the highest importance score (0.156).

These results are interesting, but speak to one the nuances of this dataset and the applicability of such a model in broader healthcare setting. First, there are serious feature imbalances within the combined dataset, and this impacts how features are used as predictors. One issue is that there are five hundred more male patients than female patients, and the resulting

feature importance of *sex* reflects this. Having separate models for males and females is a consideration for a future project. Another issue is the incompleteness of some of the subsets. Two features that were ultimately stripped were *ca* and *thal* boasted high correlation scores in the imputed Cleveland set. It is possible that these feature are strong predictors, however their relative scarcity in the final combined set prevented their inclusion.

Conclusion

A Random Forest model performing the best was not wholly unexpected. Its use of several decision trees is likely its greatest strength, especially on a dataset that features both continuous and categorical features. As a whole, this study does present machine learning as a viable diagnostic tool. In order for it to be seriously considered in a clinical setting however, there will need to be considerable improvement made to several elements of the pipeline. Expanding the existing datasets will be integral to achieving this end, as presently they are simply too small. Demonstrating efficacy of a model capable of accurate multi-class prediction is also another key improvement that must be made.

References

Aha, D. W. (n.d.). Heart Disease Data Set.

Chen, A. H., Huang, S. Y., Hong, P. S., Cheng, C. H., & Lin, E. J. (2011). HDPS: Heart Disease Prediction System. *Computing in Cardiology*, 557–560. Retrieved from http://cinc.mit.edu/archives/2011/pdf/0557.pdf

Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. *Expert Systems with Applications*, *35*(1-2), 82–89. doi: 10.1016/j.eswa.2007.06.004

Know Your Risk for Heart Disease. (2019, December 9). Retrieved from

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

https://www.cdc.gov/heartdisease/risk_factors.htm

Appendix

Table 1.A

Combined Dataset Feature Correlations with Labels

Feature Name	Correlation with "Num" (Label)
num	1.000
cp	0.472
exang	0.462
oldpeak	0.388
slope	0.340
sex	0.307
age	0.283
fbs	0.128
trestbps	0.104
restecg	0.063
chol	-0.234
thalach	-0.392

Note. Features removed for high NaN values are not shown

Table 2.A

Feature Correlation within each Dataset After Imputation

Feature Names	Cleveland_Data	Hungarian_Data	Switzerland_Data	VA_Data
num	1	1	1	1
oldpeak	0.504	0.546	0.180	0.493
cp	0.407	0.506	0.232	0.168
exang	0.397	0.583	0.132	0.331
slope	0.378	0.696	-0.030	0.128
sex	0.224	0.273	0.090	0.147
age	0.223	0.159	0.051	0.287
restecg	0.184	-0.030	0.047	-0.033
trestbps	0.158	0.139	0.133	0.147
chol	0.071	0.205		0.075
fbs	0.059	0.161	0.081	0.047
thalach	-0.415	-0.333	-0.300	-0.110

Note. Highlighted cells indicate features with a correlation above 0.3

Table 3.A

Decision Tree Classifier Parameters

Parameter	Value
ccp_alpha	0
class_weight	None
criterion	entropy
max_depth	None
max_features	None
max_leaf_nodes	None
min_impurity_decrease	0
min_impurity_split	None
min_samples_leaf	41
min_samples_split	2
min_weight_fraction_leaf	0
presort	deprecated
random_state	None
splitter	best

Table 4.A

K-Nearest Neighbors Classifier Parameters

Parameter	Value
algorithm	auto
leaf_size	5
metric	minkowski
metric_params	None

n_jobs	-1
n_neighbors	10
p	2
weights	distance

Table 5.A

SGD Classifier Parameters

Parameter	Value
alpha	0.0001
average	FALSE
class_weight	balanced
early_stopping	TRUE
epsilon	0.1
eta0	0
fit_intercept	TRUE
l1_ratio	0.15
learning_rate	optimal
loss	hinge
max_iter	1000
n_iter_no_change	5
n_jobs	None
penalty	elasticnet
power_t	0.5
random_state	None
shuffle	FALSE
tol	None

validation_fraction	0.1
verbose	0
warm_start	FALSE

Table 6.A

Random Forest Classifier Parameters

Parameter	Value
ccp_alpha	0
class_weight	None
criterion	entropy
max_depth	None
max_features	auto
max_leaf_nodes	None
max_samples	None
min_impurity_decrease	0
min_impurity_split	None
min_samples_leaf	1
min_samples_split	2
min_weight_fraction_leaf	0
n_estimators	700
n_jobs	None
oob_score	FALSE
random_state	None
verbose	0
warm_start	FALSE
warm_start	FALSE

Figure 1.A

Histogram of Final Combined Dataset: Imputed Separately

Figure 2.A

Histogram of Final Combined Dataset: Imputed Together