Heart Disease Diagnosis Prediction

Abstract

In this project, heart disease diagnosis prediction was investigated using 7 kinds of classifiers in machine learning to compared their performance in this task, including ID3 (2 heuristics for tree growing and random pruning), C4.5, logistic regression, multilayer perceptron, support vector machine, bagged and boosted decision stumps. Results showed that except ID3 without pruning, all of the classifiers can perform well in this task. The characteristic of the heart disease dataset used in this research, like the importance of different attributes, was also investigated.

Introduction

Disease diagnosis is done through the analysis of many clues gathered by doctors that specify to one particular person or animal. If we regard the clues as attributes collected, disease diagnosis can be the seen as a binary classification problem in machine learning, so it is natural to think about using machine learning algorithms to help do or improve accuracy the diagnosis.

Tree structured algorithms like id3, C4.5, decision stumps combined with bagging and boosting, logistic regression, multilayer perceptron, support vector machine are popular and powerful machine learning algorithms for classification tasks. Although the precision of diagnosis has a great correlation with the clues collected for diagnosing, as we will see in the later part of this report, it is very important to see the performance of the these algorithms to help the design diagnosis machine.

The following part organized as follows: first the dataset and the performance of different classifiers will be given and compared, then the characteristic of the dataset will be investigated, particularly the importance of different attributes will be analysis both from the aspect of human experience and statics of the dataset, finally collusion will be given.

Dataset Used

It is natural to think about using different datasets from different sources should be used to analysis the performance of different classifiers on the diagnosis task, however, due to time schedule limits, heart disease dataset from Hungarian Institute of Cardiology was used for this task. The following table gives the attribute information in this dataset.

attributr #	name	range
3	age	integer
4	sex	1, 0
9	cp	4, 3, 2, 1, 0
10	trestbps	integer
12	chol	integer
16	fbs	1, 0

19	restecg	2, 1, 0
32	thalach	integer
38	exang	1, 0
40	oldpeak	float
41	slop	3, 2, 1
44	ca	3, 2, 1, 0
51	tha	7, 6, 3
58	num	1, 0

Table 1, attributes in the original dataset

For detailed description of the dataset, please refer to [1].

Experiments

In order to use classifiers, we need first deal with missing values in the dataset, where following strategies are used:

1. If missing values are more than 50%, the discard this attribute;
2. If missing values are less than 50%, then assign the missing value the average value in its class.

#41, #44, and #51 are discarded.

Then the evaluation of the classifiers is done using 2 strategies:

1. Naive approach, random shuffle the dataset to a training set and testing set, fix these two set, do classification using different classifies (use the binary form data, it was generated as described in the following);
2. 10 fold cross validation, on original dataset and the binary form dataset, binary from data set is generated as follow:

   For #3 age, assign 1 for age>average_age, based on classes;

   For #10, assign 1 for cp=1, 2, 3, 4, because 0 represents no chest pains symposium;

   For #12, assign 1 for chol>average_chol, based on classes;

   For #32, assign 1 for thalach>average_thalach, based on classes;

   For #40, assign 1 for oldpeak>average_oldpeak, based on classed;

   Otherwise, assign 0 for attributes above.

The experiments results are giving in the following 2 tables.

Table 2, naïve approach accuracy reports

classifier	Precision (%)
id3+InfoGain	79.31034483
id3+InfoGain+RandomPruning	86.20689655
id3+VarianceImpurity	75.86206897
id3+VarianceImpurity+RandomPruning	86.20689655
C4.5 (J48)	89.65517241
Logistic Regression	87.93103448

Multilayer Perceptron (hidden layer 1, 9 nodes)	86.20689655
Support Vector Machine	87.35632184
Bagged Decision Stumps	87.5862069
Boosted Decision Stumps	87.35632184

Table 3, 10 fold cross validation reports

classifiers	original dataset precision (%)	binary dataset precision (%)
C4.5 (J48)	77.55102041	82.31292517
Logistic Regression	79.59183673	81.63265306
Multilayer Perceptron (hidden layer 1, 9 nodes)	78.79818594	80.83900227
Support Vector Machine	79.16666667	80.95238095
Bagged Decision Stumps	79.45578231	81.02040816
Boosted Decision Stumps	79.47845805	81.06575964

From the experiments results, we can have following conclusions:

1. Pruning is very important for id3 algorithms, as we can see from table 1, after pruning, the performance of id3 improved significantly;
2. C4.5 is a very powerful algorithms when the data is in binary forms, it performed best under two evaluation strategies when applied to binary form data;

Attribute Analysis

Since the dataset are collected in the process heart disease diagnosis, based on human experience, we can investigate a little deeper into the importance of each attributes to the final diagnosis. In fact, in the process of producing binary form dataset, I used the meaning of each attribute needed to be converted to binary forms, like that I assign 1 for all value of #9 cp which corresponding to a chest pain, as we can see from table 3, in the wishes that it can produce a better performance in this dataset.

Using information gain attribute evaluation tool in Weka [5], the ranking of attributes for both original dataset and binary form dataset are produces bellow:

Table 4, attribute ranking

original dataset	binary dataset
cp	exang
exang	oldpeak
oldpeak	sex
thalach	fbs
sex	trestbps
fbs	cp
restecg	chol
chol	restecg
trestbps	thalach
age	age

As we can see from the ranking result, #38 exang is an important attribute for diagnosis of heart

disease in both dataset, it is clearly true because it represents exercise induced angina, it is hard to consider for a health person who accounts this kind of symposium.

An inappropriate procedure in the producing binary forms happens for the assignment of #9 cp, as we can see from Table 4, the importance of #9 reduced after converting to binary form dataset. So it might produce a better dataset for classification if we can investigate the attributes deeper before using machine learning algorithms.

Conclusion

From the experimental part and the attribute analysis part, we can have following conclusion:

1. Decision tree structured algorithms with proper pruning like C4.5 performs good in this designed heart disease diagnosis prediction task
2. Feature engineering is vital for machine learning tasks. Dedicated selected or designed attributes can improve the accuracy.

Thread to Validate

Collect enough data is important to get more general conclusions, while the dataset used in this task is not that large, so the performance of these classification algorithms should be further investigated using more data.

The relation between collected attributes and final classification accuracy should be investigated in more detail. As we can see from the experimental part, the performance increases if we choose the binary form dataset after some preprocessing.

We should account the cost of different prediction results, if we diagnosis a patient to a healthy one, it is usually more unacceptable because it may directly cause the real patient's death if the disease is a deadly disease.

References

[1] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[2] https://en.wikipedia.org/wiki/Computer-aided_diagnosis

[3] Pattern Recognition and Machine Learning (Information Science and Statistics), Christopher M. Bishop, Springer-Verlag New York, Inc. Secaucus, NJ, USA ©2006

[4] Machine Learning, Thomas M. Mitchell, McGraw-Hill, Inc. New York, NY, USA ©1997

[5] http://www.cs.waikato.ac.nz/ml/weka/