CMPT 459 E100 Data Mining

Assignment 1

Report

Prof: Martin Ester

Student Name: Zeyong Jin

SFU ID: 301353174

Computing ID: zeyongj@sfu.ca

Date: February 7th, 2021

Contents

Part 1 Pseudocode 3

Part 2 Answers of Task 1 And 3 6

Task 1 Average Accuracy 6

Task 3 Handle Missing Values 6

Part 3 Acknowledgements 7

1 Reference 7

2 Read Me 7

Part 1 Pseudocode

funtion entropy (D) [1]

input: an attribute-valued dataset D

output: a double-type entropy

log2 (x) = log (x) / log (2)

entropy = 0

dictionary = summarizeDataset (D, targetAttribute)

for key k in dictionary

    proportion = dictionary[k] / total number of D

    entropy = entropy - proportion * log2 (proportion)

return entropy

function informationGain (dataset, attribute, entropyOfSet) [1]

input: an attribute-valued dataset D, attribute A, entropy of set E

output: a double type informationGain

informationGain = E

for value v in attributeValues (D, A)

    sub = subset (D, A, v)

informationGain -= (number in sub) / (total number of D) * entropy (sub)

return informationGain

procedure growDecisionTree (dataset) [2]

// Grow Algorithm, known as C4.5 Algorithm, works on both categorical and numerical data.

```
input: an attribute-valued dataset D
output: a decision tree T
identify continuous attributes
if there are some values = "?"
    predict missing values // In order to handle missing values
T = {}
if entropy (D) == 0 then
    terminate // All training examples corresponding to a leaf node belong to the same class.
for all attribute A in D
    compute gain ratio if we split on A
maxA = attribute A with the highest information-theoretic criteria (gain ratio)
T.root = maxA
subD = induced sub-datasets from D based on maxA
for all subD
    subT = grow(subD)
    Attach subT to the corresponding branch of T
return T
```

procedure errorReductionPruning (Tree, SplitRatio) [3]

input: an unmodified decision tree Tree, a double value of split ratio

// Split ratio is to determine the percentage of data that would be leaving for test

output: a modified decision tree Tree'

SplitExamples(SplitRatio, Tree, TrainingSet, TestingSet)

```
Theory = SeparateAndConquer(TrainingSet)

loop

NewTheory = BestSimplification(Theory,TestingSet)

if Accuracy (NewTheory,TestingSet) < Accuracy (Theory,TestingSet)

exit loop

Theory = NewTheory

return(Theory)
```

Part 2 Answers of Task 1 And 3

Task 1 Average Accuracy

The average accuracy in 5-fold cross validation is 0.7851107226107226.

And the accuracy of final decision tree is 0.7952762533075034.

Task 3 Handle Missing Values

The grow function allows attribute values to be marked as "?" for missing. And missing attribute values are simply not used in gain and entropy calculations. [4]

To be precise, this algorithm will return the probability of belonging to the positive or negative class for the set of a certain attribute where the current attribute is N/A. And then, I called the function to return the prediction for the set of attribute X and each of set of attributes on the list X. [5] [6]

Part 3 Acknowledgements

1 Reference

[1] L. Meeden. [Online]. Available: https://www.cs.swarthmore.edu/~meeden/cs63/f05/id3.html. [Accessed 8 February 2021].

[2] "otnira golb!," [Online]. Available: http://www.otnira.com/2013/03/25/c4-5/. [Accessed 8 February 2021].

[3] P. A. f. R. Learning, "JOHANNES FURNKRANZ," Machine Learning, no. 27, pp. 139-172, 1997.

[4] H. G. Gaurav L. Agrawal, "Optimization of C4.5 Decision Tree Algorithm for Data Mining," International Journal of Emerging Technology and Advanced Engineering, vol. 3, no. 3, pp. 341-345, March 2013.

[5] Chirag, "Stack Overflow," 27 February 2017. [Online]. Available: https://stackoverflow.com/questions/42219073/c4-5-algorithm-missing-values#:~:text=Don'tPlay%20Play-,The%20C4.,which%20the%20value%20is%20missing.. [Accessed 8 February 2021].

[6] Sergi, "GitHub," 16 November 2020. [Online]. Available: https://github.com/AtenrevCode/DecisionTreeClassifier/blob/master/tree.py. [Accessed 10 February 2021].

2 Read Me

1. The codes are inspired and modified by the website: https://github.com/AtenrevCode/DecisionTreeClassifier/blob/master/. And I would like to express my thank to the original author.

2. The three functions in the 2.2 section, grow function starts from line #241, prune function starts from line #268, test function starts from line #117.

3. The predictions could be reproduced and the accuracy value would not be changed.