

Assignment No 03: Decision tree

CSE-0408 Summer 2021

TASNOVA TASNIM

Department of Computer Science and Engineering
State University of Bangladesh (SUB)
Dhaka, Bangladesh
email:tasnimprity12@gmail.com

Abstract—Decision Trees are considered to be one of the most popular approaches for representing classifiers. Researchers from various disciplines such as statistics, machine learning, pattern recognition, and Data Mining have dealt with the issue of growing a decision tree from available data. This paper presents an updated survey of current methods for constructing decision tree classifiers in a top-down manner. The chapter suggests a unified algorithmic framework for presenting these algorithms and describes various splitting criteria and pruning methodologies.

Index Terms—python

I. INTRODUCTION

A decision tree is a flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret.

II. LITERATURE REVIEW

compared the decision tree classification algorithm and developed the Weka method and it is based on choosing the file and selecting attributes to convert .csv file to flat file. The decision tree algorithms are presented and achieved a high rate of accuracy for classify the data into the correctly and incorrectly instances. Anshul Goyal et .al [7] background study a performance evaluation of Naïve bayes and J48 classification algorithms. J48 gives more classification accuracy for bank dataset having two values Male and Female. The result shows that j48 and Naïve Bayes gives better accuracy.

ShwetaKharya et.al [8] examined various data mining approaches that have been applied for breast cancer diagnosis and prognosis. Decision tree is search to be the best forecaster with 93.62on benchmark dataset and also on SEER data set. Abdullah H. Wahbeh et. al [9] had presented a performance evaluation of Naïve Bayes, J48 classification, sequential Minimal Optimization (SMO) classifier. Compared these three

classification techniques based on two main aspects such as accuracy and execution time. In term of accuracy, results showed that the Naïve Bayes classifier achieved the highest accuracy, followed by SMO and J48 classifier. In term of execution time, results showed that the SMO model takes less execution time followed by the NB model and J48 classifier.

III. PROPOSED METHODOLOGY

The main objective of the study is to find the best decision tree based classification algorithms from five algorithms namely ID3, C4.5, C5.0, PART and Bagging CART. The classification algorithms are validated based on the performance measures such as precision, recall, f-measures, accuracy and kappa statistic

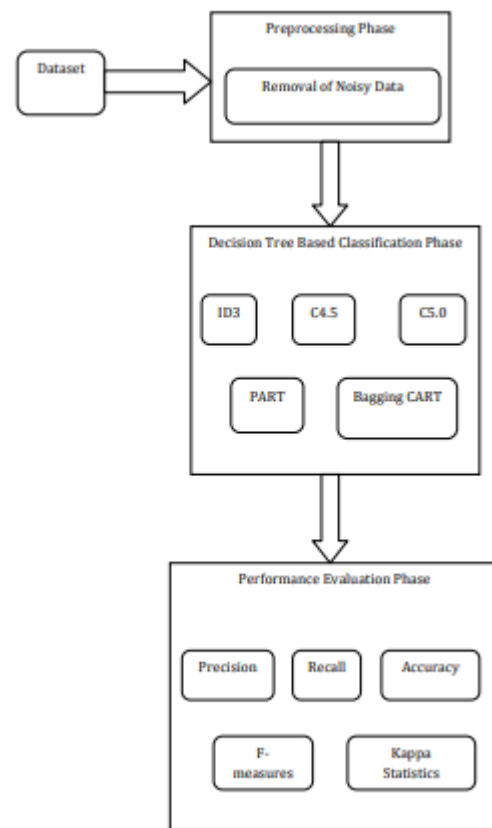


fig:Decision Tree based Classification Algorithms

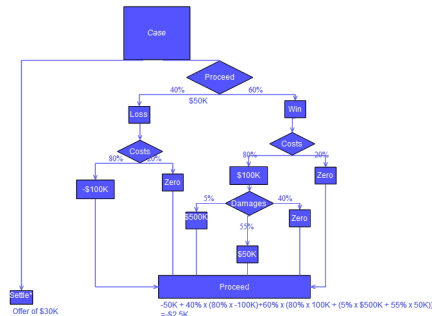
IV. DECISION RULES

The decision tree can be linearized into decision rules,[2] where the outcome is the contents of the leaf node, and the conditions along the path form a conjunction in the if clause. In general, the rules have the form:

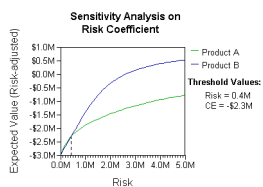
if condition1 and condition2 and condition3 then outcome.

Decision rules can be generated by constructing association rules with the target variable on the right. They can also denote temporal or causal relations.

V. DECISION TREE USING FLOWCHART SYMBOLS



VI. ANALYSIS EXAMPLE



The basic interpretation in this situation is that the company prefers B's risk and payoffs under realistic risk preference commonly used in operations research courses, is the distribution of lifeguards on beaches. The example describes two beaches with lifeguards to be distributed on each beach. There is maximum budget B that can be distributed among the two beaches (in total), and using a marginal returns table, analysts can decide how many lifeguards to allocate to each beach.

VII. ADVANTAGES AND DISADVANTAGES

Among decision support tools, decision trees (and influence diagrams) have several advantages. Decision trees:

Are simple to understand and interpret. People are able to understand decision tree models after a brief explanation. Have value even with little hard data. Important insights can be generated based on experts describing a situation (its alternatives, probabilities, and costs) and their preferences for outcomes. Help determine worst, best and expected values for different scenarios. Use a white box model. If a given result is provided by a model. Can be combined with other decision techniques.

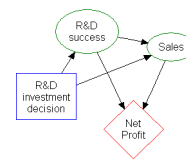
Disadvantages of decision trees:

They are unstable, meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree. They are often relatively inaccurate. Many other

predictors perform better with similar data. This can be remedied by replacing a single decision tree with a random forest of decision trees, but a random forest is not as easy to interpret as a single decision tree. For data including categorical variables with different number of levels, information gain in decision trees is biased in favor of those attributes with more levels.[7] Calculations can get very complex, particularly if many values are uncertain and/or if many outcomes are linked

VIII. INFLUENCE DIAGRAM

Much of the information in a decision tree can be represented more compactly as an influence diagram, focusing attention on the issues and relationships between events.



IX. CONCLUSION

Data mining is used to extract useful knowledge from large data repositories. Recently data mining techniques have enclosed every field in our life. Data mining have numerous algorithms to use for different purpose. In this paper discussed about the classification techniques. From this, the decision tree based classification algorithms namely ID3, C4.5, C5.0, PART and Bagging CART are used to perform classification process. The four data sets Iris, Balance scale, Contact lenses, Pima Indian Diabetes have been applied and performance is validated based on Accuracy (CA), Precision, Recall, F-Measure and Kappa Statistics.

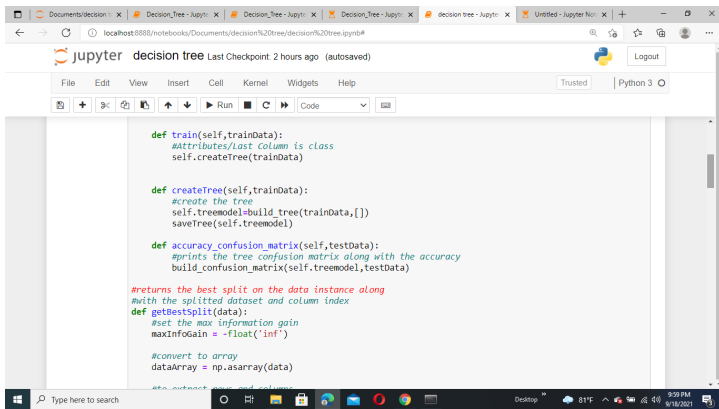
X. CODE

```
In [2]: import numpy as np
from itertools import groupby
import math
import collections
from copy import deepcopy
import pickle

class TreeNode:
    def __init__(self, split_col_index):
        self.col_idx = col_index
        self.split_value = split
        self.parent = None
        self.left = None
        self.right = None

class Tree():
    def __init__(self):
        self.treemodel = None

    def train(self, traindata):
        #MATH10062/infocv2014/4_v0.015
```



```
def train(self, trainData):
    #Attributes/last column is class
    self.createTree(trainData)

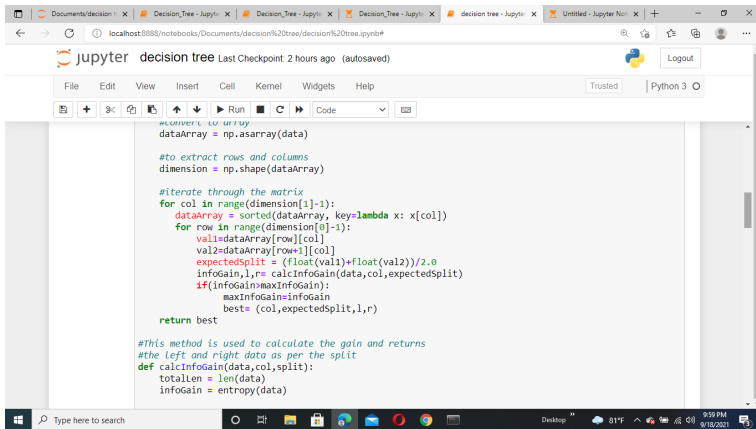
def createTree(self, trainData):
    #create the tree
    self.treemodel=build_tree(trainData,[])
    saveTree(self.treemodel)

def accuracy_confusion_matrix(self, testData):
    #prints the tree confusion matrix along with the accuracy
    build_confusion_matrix(self.treemodel, testData)

#returns the best split on the data instance along
#with the splitted dataset and column index
def getBestSplit(data):
    #set the max information gain
    maxInfoGain = -float('inf')

    #convert to array
    dataArray = np.asarray(data)

    #to extract rows and columns
    dimension = np.shape(dataArray)
```



```
dataArray = np.asarray(data)

#to extract rows and columns
dimension = np.shape(dataArray)

#iterate through the matrix
for col in range(dimension[1]-1):
    dataArray = sorted(dataArray, key=lambda x: x[col])
    for row in range(dimension[0]-1):
        val1=dataArray[row][col]
        val2=dataArray[row+1][col]
        expectedSplit = ((float(val1)+float(val2))/2.0)
        infoGain,l,r= calcInfoGain(data,col,expectedSplit)
        if(infoGain>maxInfoGain):
            maxInfoGain=infoGain
            best= (col,expectedSplit,l,r)

    return best

#This method is used to calculate the gain and returns
#the left and right data as per the split
def calcInfoGain(data,col,split):
    totalLen = len(data)
    infoGain = entropy(data)
```

ACKNOWLEDGMENT

I would like to thank my honourable **Khan Md. Hasib Sir** for his time, generosity and critical insights into this project.

REFERENCES

- [1] Osmar R.; Zaine. (1999): Introduction to DataMining, CMPUT690 Principles of Knowledge Discovery in Databases, University of Alberta, Department of Computing Science. .
- [2] Han Wu.; Shangqi Yang.; Zhangqin Hung.; Jian He.; Xiaoyi Wang. (2018): Type 2 diabetes mellitus prediction model based on data mining, Informatics in Medicine Unlocked
- [3] K. Karimi and H.J. Hamilton (2011), "Generation and Interpretation of Temporal Decision Rules", International Journal of Computer Information Systems and Industrial Management Applications, Volume 3
- [4] Wagner, Harvey M. (1 September 1975). Principles of Operations Research: With Applications to Managerial Decisions (2nd ed.). Englewood Cliffs, NJ: Prentice Hall. ISBN 9780137095926.
- [5] Geetha Kashyap.; Ekta chauhan. (2016): Parametric comparisons of classification techniques in Data mining application, International journal of Engineering Development Research 4(2), pp. 1117-1123.
- [6] Utgoff, P. E. (1989). Incremental induction of decision trees. Machine learning, 4(2), 161–186 .
- [7] Deng,H.; Runger, G.; Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions. Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN).