CSE - 4255 Data Mining and Warehousing Lab

Comparison Between the Performance of Decision Tree and Naive Bayes Classifier in Classification

Saif Mahmud
Roll: SH - 54

M. Tanjid Hasan Tonmoy
Roll: SH - 09

Submitted To:

Dr. Chowdhury Farhan Ahmed
Professor

&

Abu Ahmed Ferdaus
Associate Professor

Department of Computer Science and Engineering
University of Dhaka

September 30, 2019

1 Problem Definition

In this experiment, we have implemented two different classification algorithms, namely Decision tree and Naive Bayes. The algorithms utilize discrete and continuous features to predict class labels. Comparative analysis of these two algorithms have been conducted using various evaluation metrics for both balanced and imbalanced datasets of varied sizes.

2 Theory

2.1 Decision Tree

2.2 Naive Bayes Classifier

3 Experimental Setup

3.1 Implementation

For the implementation of decision tree, two different attribute selection methods (entropy and Gini index) have been used for both discrete and continuous attributes and is available as option for training models. hen there is a large number of distinct values for a continuous attribute, the training time increases significantly due to the fact that all possible splitting points have to be considered. The tree is stored using a dictionary structure in python and built recursively. Prepruning of the tree based on a threshold given as input has been used to prevent over-fitting.

3.2 Datasets

4 Result

5 Discussion

Both of these algorithms produce reasonable performance when dealing with moderate sized datasets with close to balanced class distribution. However, in case of class imbalance, both of thee algorithms suffer.