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.

Dataset	Dataset Size	k-Fold Cross Validation (k = 5) Accuracy (%)	Avg. Accuracy (%)
Adult	32561	82.9879	83.35124
		83.2463	
		83.4459	
		83.6302	
		83.4459	
Breast Cancer	286	86.2069	73.38174
		82.4561	
		64.9123	
		64.9123	
		68.4211	
Census-Income	199523	85.6939	85.57612
		85.255	
		85.5654	
		85.8235	
		85.5428	
Chess	3196	89.8438	87.63982
		86.25	
		88.1064	
		88.7324	
		85.2665	
Chess - II	28056	36.2989	36.2274
		36.545	
		36.1255	
		36.3361	
		35.8315	
Connect-4	67557	72.1137	72.15832
		72.7353	
		72.2617	
		71.7786	
		71.9023	
Credit Card Default	30000	67.2055	65.12984
		70.4167	
		53.0167	
		73.2167	
		61.7936	
Iris	150	93.3333	95.33334
		90	
		96.6667	
		96.6667	
		100	

Dataset	**Dataset Size**	**k-Fold Cross Validation (k = 5) Accuracy (%)**	**Avg. Accuracy (%)**
Mushroom	8124	95.326	95.37192
		95.0154	
		94.4	
		95.6281	
		96.4901	
Pendigits	7494	78.7333	78.0755
		78.6	
		77.7333	
		77.5183	
		77.7926	
Poker	1000000	50.119	50.1187
		50.114	
		50.12	
		50.1205	
		50.12	
Breast Cancer - Wisconsin	699	65.035	62.78208
		63.8298	
		62.4113	
		60.8696	
		61.7647	