# CSE537 Artificial Intelligence

## Assignment-5 Report

### Project ClickStream

Authors:

Vishal Nayak: 109892702: vnayak@cs.stonybrook.edu

Ashish Chaudhary: 109770154: ashchaudhary@cs.stonybrook.edu

# Contents

# 1. Intro and details

## 1.1. Compilation, machine details and commands

Compiler: ***python 2.7.6***

Machine: Windows 8 64-bit

Command: **python clickstream.py**

## 1.2. Goals and Approaches

***ClickStream mining with Decision Trees:***

Mining the click-stream data collected from Gazelle.com and predicting if the user visits another page on the site or will he leave. Approach to solve this problem is to construct an Iterative Dichotomiser decision tree and using it to predict the class attribute of the test data which represents user's action in the web site.

# 2. ClickStream

***Implementation:***

- Entropy of samples is calculated.
- Information gain of each attribute belonging to the data set at particular node is calculated.
- In order to calculate the information gain, split value for each continuous attribute is calculated.
- Heuristic for splitting continuous attribute:
- Finding and counting the attribute values which contains class attribute as positive.The value which has the maximum positives of all is used for splitting. This is not proven to be correct, but the result is not very bad.
- The attribute resulting in maximum information gain is computed and is used to create a model node.
- The samples are split using the chosen attribute and the resulting individual sets are used to recursively build the model.
- This model is an iterative dichotomiser decision tree.
- The length of the dichotomiser decision tree can grow exponentially.
- To avoid this, two steps are taken:
  - ➢ The samples at each node are split only into two halves.
  - ➢ Chi-Square criterion is used to prune the tree branches based on provided threshold values

# 3. Results

## 3.1. ClickStream

***Threshold:*** *0.05*

***Mismatches:*** *6292.0*

***Accuracy:*** *74.832%*

***Tree Nodes Count:*** *353*

***Running Time:*** *63.0885041586 seconds.*

***Threshold:*** *0.01*

***Mismatches:*** *6292.0*

***Accuracy:*** *74.832%*

***Tree Nodes Count:*** *193*

***Running Time:*** *62.9320618995 seconds.*

***Threshold:*** *1.0*

***Mismatches:*** *6292.0*

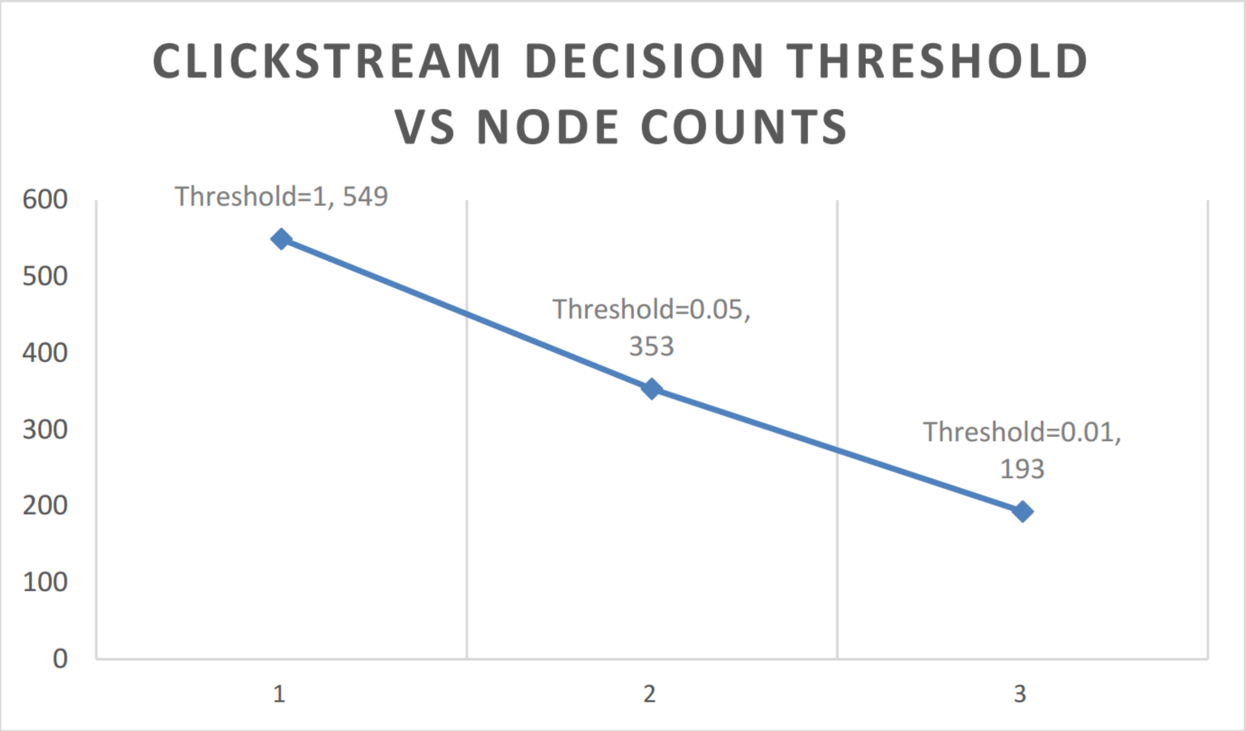
***Accuracy:*** *74.832%*

***Tree Nodes Count:*** *549*

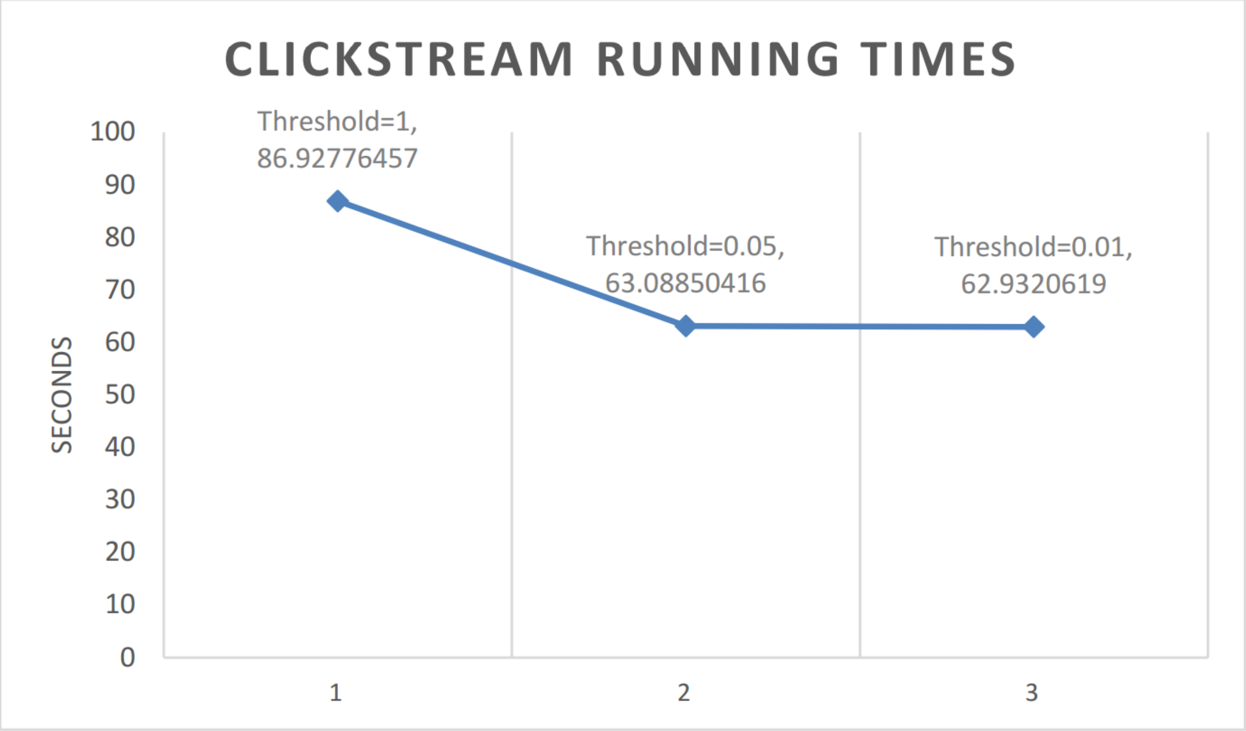
***Running Time:*** *86.9277645732 seconds.*

## 4. Statistics

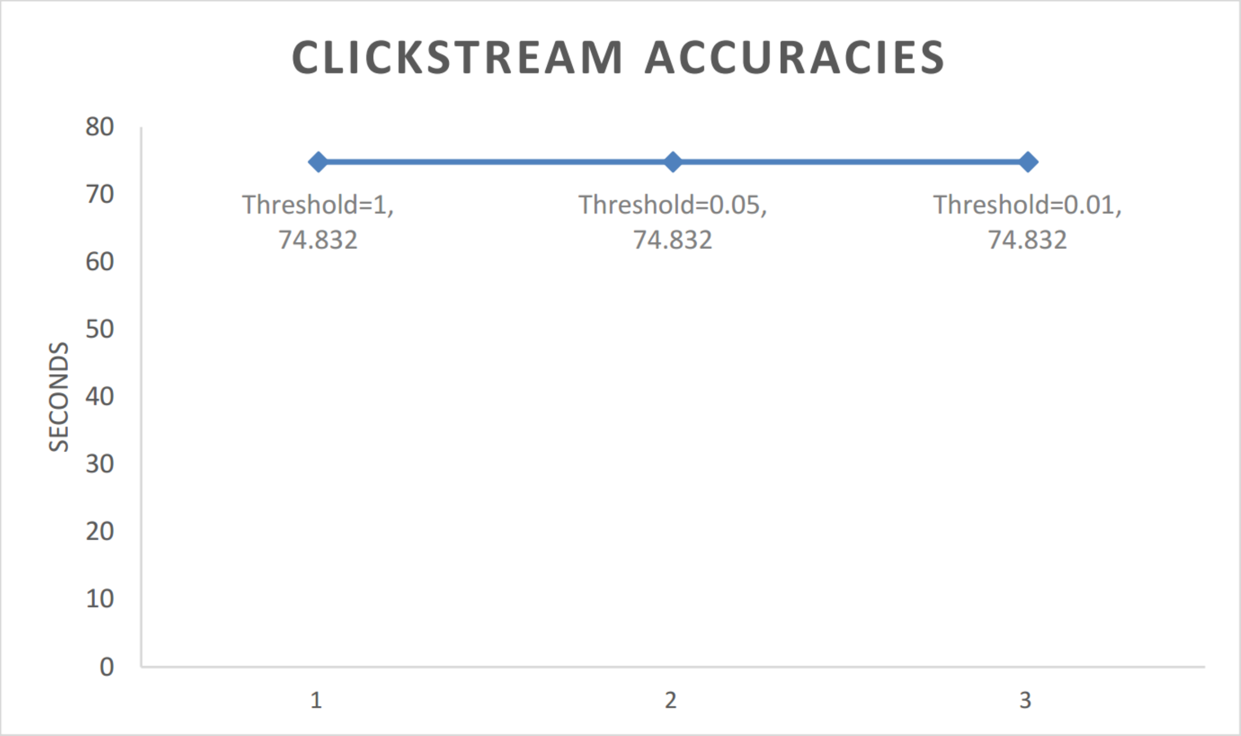
### 4.1. Node Counts



### 4.2. Running Times



## 4.3. Accuracies


CLICKSTREAM ACCURACIES
SECONDS
Threshold=1, 74.832
Threshold=0.05, 74.832
Threshold=0.01, 74.832