Assignment 5- Click Stream Project

Alam, Praveen ; Suresh, Muthukumar

May 12, 2015

Contents

1 Introduction

In this project we are building a decision tree based on click-stream data collected from Gazella.com. Target of the project is to given a set of page views of an user to gazella.com we need to decide if the user will continue to stay on the page and visit another page or leave the site.

We are using the given training data of page views and results to build a ID3 decision tree. While building decision tree, we are using concept of entropy and information gain to decide a preferred sequence of attributes. Resulting decision tree can be used to judge given set of page clicks, whether user will click another page or leave the website.

2 Algorithm Design

To build the decision tree, we parse give corpus of 40000 records to build a python data structure that resembles Conditional probability table. This CPT table has 40000 rows of training data and 274 columns of features or attributes. Last column represent whether user clicked another page or exited from the website. This CPT table is used find a node that gives best attribute(feature) that gives maximum information gain. After selecting an attribute to branch, we iterate over discrete values of the attribute and create temporary CPT tables for each value of the attribute. We recursively call decision tree build algorithm for each attribute value with the temporary CPT table.

Split Stopping criteria in building decision tree:

1. We stop branching in decision tree when the information gain in the CPT table is 0. That means all the attributes in the CPT table has same values
2. Number of attributes remaining in the CPT is one.
3. CPT table is empty

Chi-Square criteria for split stopping:

1. Instead of building and exploring the whole decision tree, which becomes difficult with growing number of attributes, we can use chi-square to decide whether an attribute is irrelevant in the decision tree. After choosing an attribute with maximum information gain in the cpt tables, we decide using the chi-square formula given in the project description to decide whether the attribute is relevant to our decision.

We use the equations to calculate this factor. let p, n denote the positive and negative examples and N is the total number of examples in the current set. The statistics of interest are given by S.

$$p_i^{'} = p\frac{T_i}{N}$$
$$n_i^{'} = n\frac{T_i}{N}$$
$$S = \sum_{i=1}^{m}(\frac{p_i^{'}-p_i}{p_i^{'}} + \frac{n_i^{'}-n_i}{n_i^{'}})$$

If the attribute is irrelevant, we replace the decision node with a leaf node with the decision if the user will click another page or exit. We are using chi-square threshold distribution table for a given P (probability) value from:
http://www.ndsu.edu/pubweb/ mcclean/plsc431/mendel/mendel4.htm

3 Analysis

We ran our algorithm for different values of " p values". As required in the project description, we have run our algorithm for p value 0.01, 0.05 and 1. P value '1' means constructing the whole tree.

Below table provides accuracy and space value for each value of p. We did not include leaf nodes (decision nodes are counted as part of the decision tree nodes)

S.No	P Value	Accuracy %	Space# Nodes
1	0.01	76.29	913
2	0.05	76.2	901
3	1	74.46	956

4 Conclusion

In this project, we can see the performance of the ID3 algorithm for the task of anayzing the user behavior on a particular site. At each stage we choose 1 attribute to branch on, the attribute chosen is based on the entropy and information gain we get by choosing that attribute. We use chi-Square for our split stopping criteria. We got decent results with an accuracy of around 76 %. By tweaking the p Value we can get slightly better or worse values but overall the performance of the algorithm remains consitent around the same value.