Comparative Analysis of Association Rule Mining and Decision Tree Algorithms

Group 3

Sahaj Agarwal S20160010077
Uday Raj Sawhney S20160010099

November 17, 2018

Overview

In the field of Data Mining, Association Rule Mining and Decision Tree Algorithms form the back bone of extracting knowledge from provided data in the form of Association Rules as in FP-Growth Algorithms and Classification of data into multiple classes using Decision Tree Algorithms.

Introduction

- Our research work presents a summarization and a comparative study of the available FP-growth and Decision Tree Algorithms.
- FP-Growth variations include AFOPT, NONORDFP and FP-GROWTH*.
- Decision Tree Algorithms include ID3, CART, C4.5 and C5.0.
- Finally we propose a approach where we can use Association Rules mined for Classification and compare results against Decision Tree Algorithms.

Insight to Decision Trees

This method uses tree structure to build the classification models. It divides a dataset into smaller subsets. Leaf node represents a decision. Based on feature values of instances, the decision trees classify the instances.

Each node in a decision tree represents a feature in an instance which is to be classified, and each branch represents a value. Classification of Instances starts from the root node and sorted based on their feature values. Categorical and numerical data can be handled by decision tress.

Decision Tree Algorithms

ID3 (or Iterative Dichotomizer)

The ID3 algorithm uses a greedy search. It selects a test using the information gain criterion, and then never explores the possibility of alternate choices. Only one attribute at a time is tested for making a decision.

C 4.5

Improved version on ID3, accepts both continuous and discrete features. It handles incomplete data points and solves over-fitting problem by bottom-up technique usually known as pruning. Moreover different weights can be applied to the features that comprise the training data.

C 5.0

C5 acknowledges noise and missing data. Problem of over fitting and error pruning is solved. C5.0 incorporates several new facilities such as variable misclassification costs and boosting.

Association Rule Mining

An association rule is defined as a relation between different itemsets. The process of extracting the association rules can be viewed as two-phases:

- The first phase is to mine all frequent patterns with each pattern happening at least as frequently as preset minimum support count.
- The second phase is to produce strong association rules from the frequent patterns complying with minimum support as well as minimum confidence.

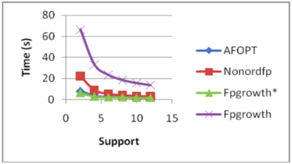
FP-Growth algorithm is an efficient method of mining all frequent itemsets without candidate generation using pattern growth approach. Variants of FP-Growth algorithm:

- AFOPT
- NONORDFP
- FPGROWTH*

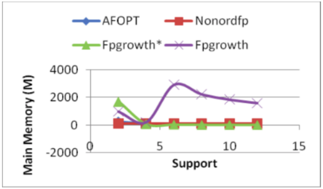
Comparison between FP-Growth Variants

T10I4D100k Dataset

- Items - 1000
- Avg. Length - 10
- Transactions - 100,000
- Type - Sparse



(a) Run-time



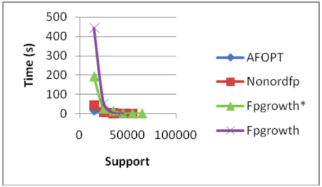
(b) Memory Consumption

Figure: Comparison between FP-Growth Variants

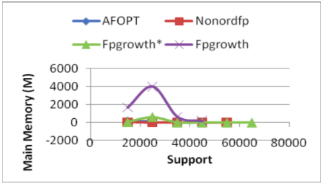
Comparison between FP-Growth Variants

Connect4 Dataset

- Items - 150
- Avg. Length - 43
- Transactions - 67,557
- Type - Dense



(a) Run-time



(b) Memory Consumption

Figure: Comparison between FP-Growth Variants

Comparison between FP-Growth and CART Decision Tree Algorithm

Dataset

We used the IRIS flower dataset for comparison between the two types of algorithms. The dataset contains 150 training examples divided into 3 equal partitions and 3 classes. We divide the dataset randomly into train and test splits in the ratio $9:1$.

Accuracy

On the basis of accuracy metric, we were able to achieve an accuracy of about 93.3% on the test-split using CART decision tree algorithm whereas about 93% using FP-Growth algorithm.

We can infer from the experiments that Decision Tree algorithms outperform FP-Growth based algorithms in the domain of classification. Although, the difference is marginal and can be altered depending on the optimization tricks and choice of hyper-parameters.

Thank You