MANIPAL INSTITUTE OF TECHNOLOGY
(A constituent Institute of Manipal University)
MANIPAL - 576 104, KARNATAKA, INDIA
MARCH 2017

Project Report

A Comparative Study Of Association Rule Mining Algorithms

By

Vanessa Singh
140911092

Ishita Bedi
140911104

IT B
6th Semester
Data Warehousing and Data Mining Lab
Department of Information and Communication Technology
Manipal Institute of Technology, Manipal

Under the Guidance of:

Ms. Anju R
Mr. Nirmal Kumar Nigam
Mrs. Girija A

INTRODUCTION

Data mining is an essential tool to gather information from large data sets. Association rule mining helps to determine the important and implicit information from data sets. A wide variety of algorithms exist to perform this task. Our study compares a few such algorithms, namely Apriori, FP-Growth and Eclat (Vertical Data Format).

PROBLEM DEFINITION

Association rule mining is the study of extracting interesting and essential association rules in a large data set. This is important so that data can be analysed effectively and knowledge can be formed for future prediction as well. Association rules are statements which help uncover relationships between seemingly unrelated data in a data set. Association rules analysis is a technique to uncover how items are associated to each other. Therefore, many algorithms exist to perform ARM.
In our study, implemented and compared a few such algorithms, namely Apriori, FP-Growth and Eclat based on variousparameters, such as dimensions (by increasing dimensionality of the data set), the tuples (increasing tuple count), time complexity, number of transactions, etc.
We performed these comparisons with with varying support and confidence for the data. We are determining the association rules on the basis of two input parameters - support and confidence. The rule $A \Rightarrow B$ holds in the transaction set D with support s, where s is the percentage of transactions in D that contain $A \cup B$ (i.e., both A and B). This is taken to be the probability, $P(A \cup B)$. The rule $A \Rightarrow B$ has confidence c in the transaction set D if c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, P(B|A).

OBJECTIVE

To implement and compare a few association rule mining algorithms and to analyse the advantages and disadvantages of each algorithm.

SCOPE/IMPORTANCE OF PROJECT

Association rule mining is itself a huge area of research. It also contributes greatly to other major research areas such as Machine Learning, which in turn is the backbone of AI. Association rule mining can be extensively used for decision support, financial forecasts, marketing policies, medical diagnosis and mining research. A comparison would hence help us obtain a clearer picture about the algorithms and their use in various scenarios, and also help us draw statistically verified conclusions about them.

METHODOLOGY

The basis of our implementation is a paper titled "Comparing the Performance of Frequent Pattern Mining Algorithms - Dr. Kanwal Garg, Deepak Kumar " and we planned on implementing and comparing the Apriori, FP growth and Eclat algorithms based on varying the dimensionality and tuple size of the same dataset and comparing statistically by plotting execution time vs support graphs.
Our implementation is in Python, and we're using a package called PyFIM, which is especially for

Frequent Itemset Mining and Frequent Association Rule Mining.

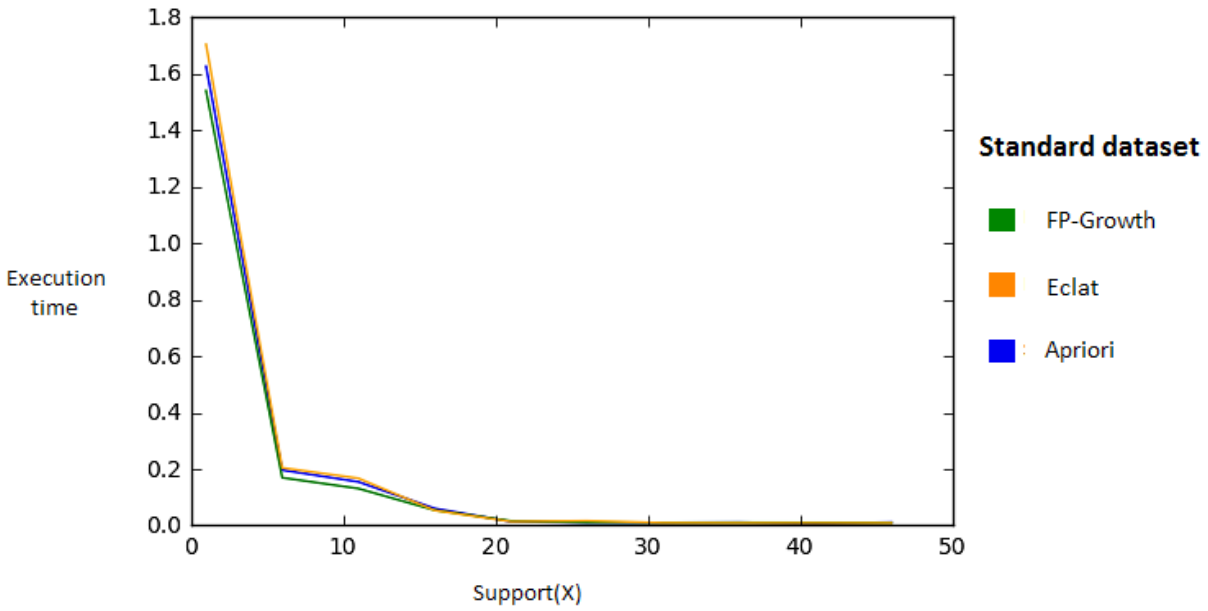
Our comparative study is performed as follows:

Our dataset, for which we shall be showing graphs, is a grocery dataset which has been preprocessed for mining. Four variations, which are - the standard dataset, dataset with twice the number of tuples, dataset with twice the number of attributes and dataset with twice the number of attributes and tuples, are used for comparison. For every such comparison, we implement all the algorithms on each of the datasets for varying supports and for each support, calculating the execution time. This is plotted on graphs to analyse which algorithm performs more efficiently for which scenario. In the end an analysis is also done to analyse the performance of individual algorithms for the four scenarios.

CONCLUSION

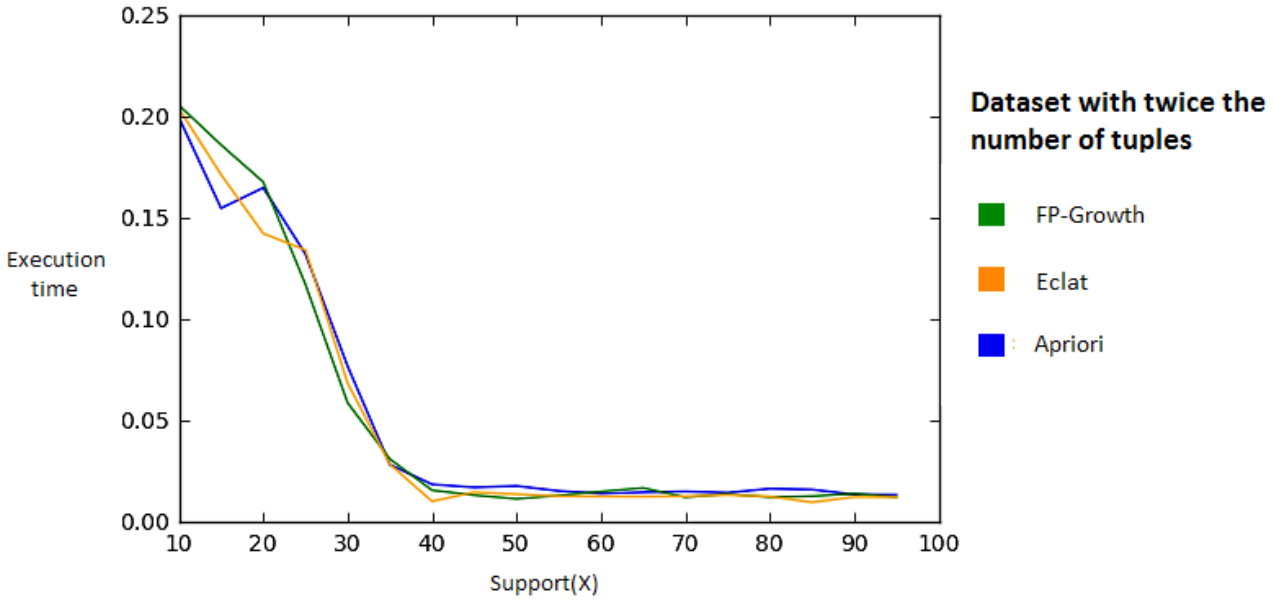
- **Standard Grocery Dataset**

  FPGrowth performed the best for 10% confidence and support count varying from 1 to 50 at a step size of 5. The average execution time was 0.2101, 0.1963, 0.2195 for apriori, fpgrowth and eclat respectively. FPGrowth's performance was the best followed by apiori and eclat.



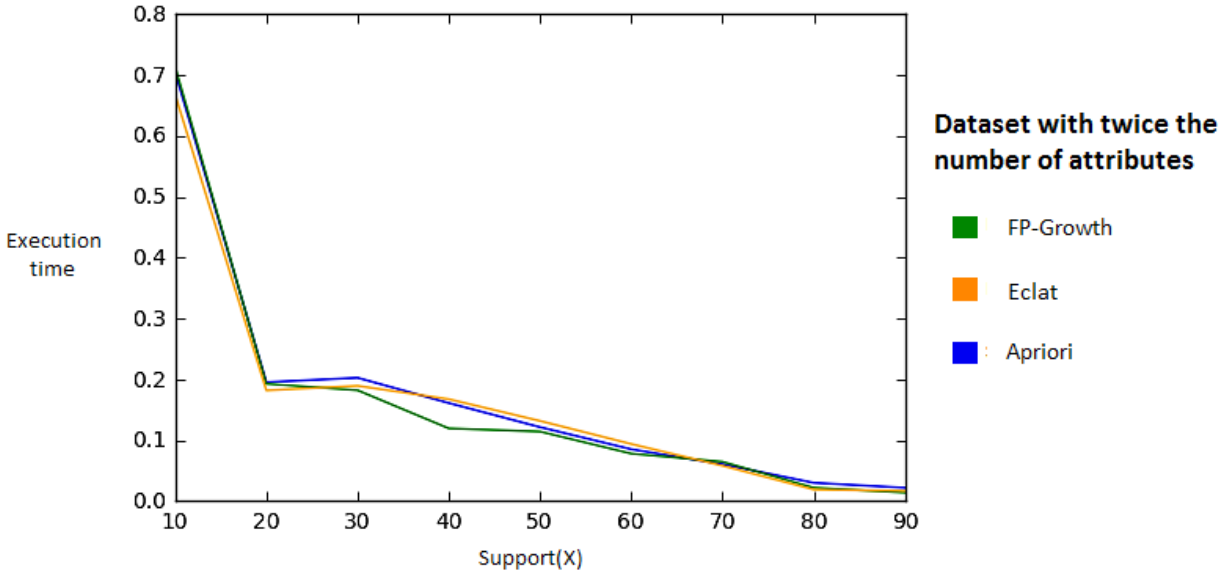
- **Dataset with double the number of tuples**

  Eclat achieved the best average for 10% confidence and support count varying from 10 to 100 at a step size of 5. The average execution time was 0.05265, 0.05188, 0.05016 for apriori, fpgrowth and eclat respectively. Though eclat received the best average the graph clearly shows thta FPGrowth's overall permormance was better than eclat's. Apriori performed the poorest.



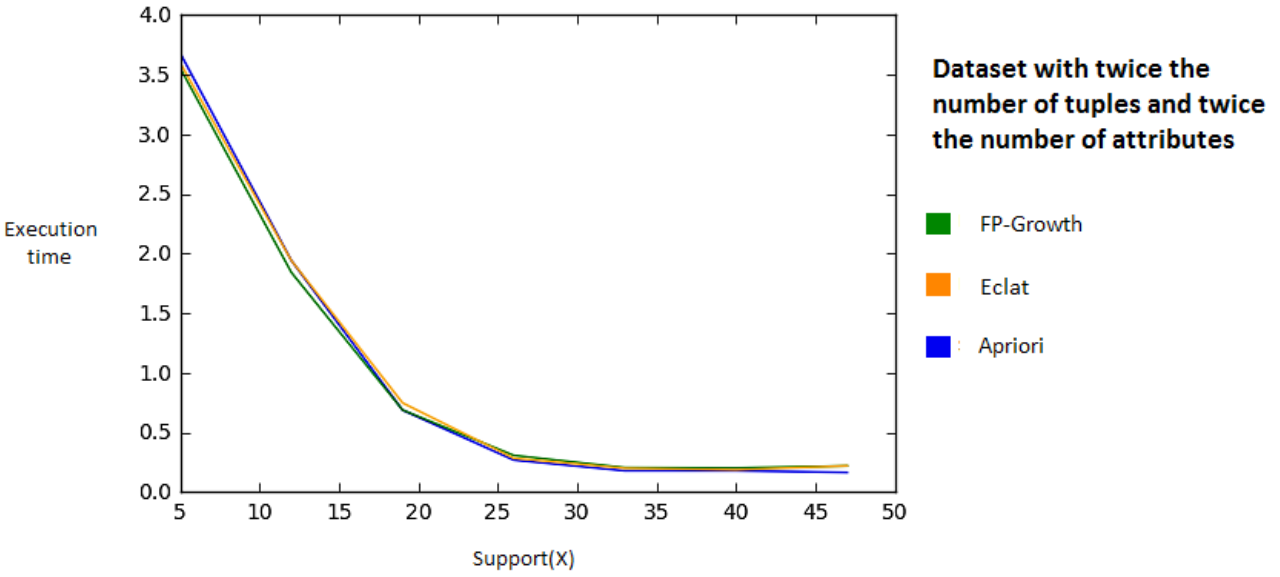
- **Dataset with double the number of attributes**

  FPGrowth performed the best for 10% confidence and support count varying from 10 to 100 at a step size of 10. The average execution time was 0.17611, 0.16698, 0.16978 for apriori, fpgrowth and eclat respectively. FPGrowth's performance was the best followed by apiori and eclat.



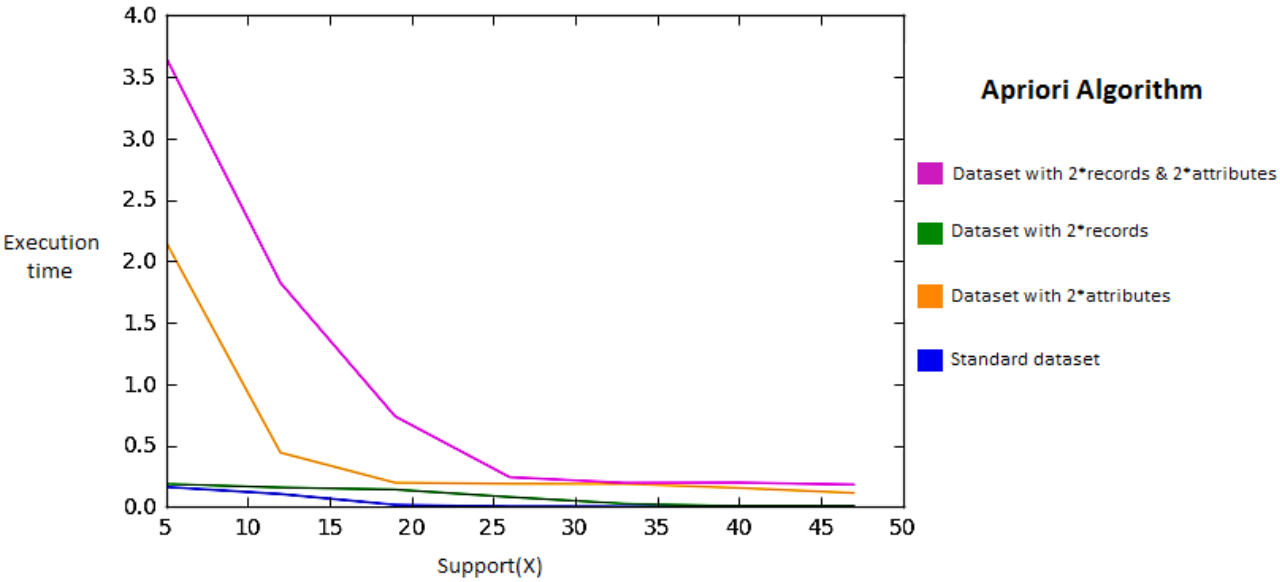
- **Dataset with double tuples and double attributes**

  FPGrowth performed the best for 10% confidence and support count varying from 5 to 50 at a step size of 7. The average execution time was 1.0150, 1.0036, 1.02625 for apriori, fpgrowth and eclat respectively. FPGrowth's performance was the best followed by eclat and apriori.
  As we can observe this dataset took the longest time to execute.

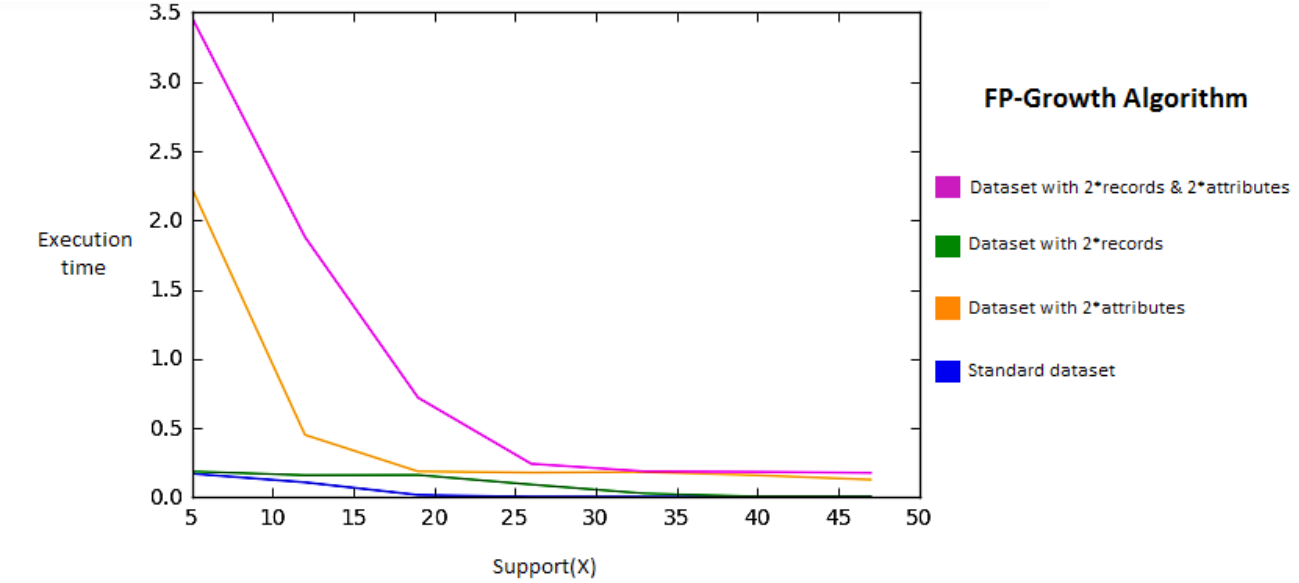


In the following graphs we can observe that execution time decreases in the order of dataset with double tuples and double attributes, dataset with double attributes, dataset with double rows and standard dataset. This analysis is done for all the three algorithms.

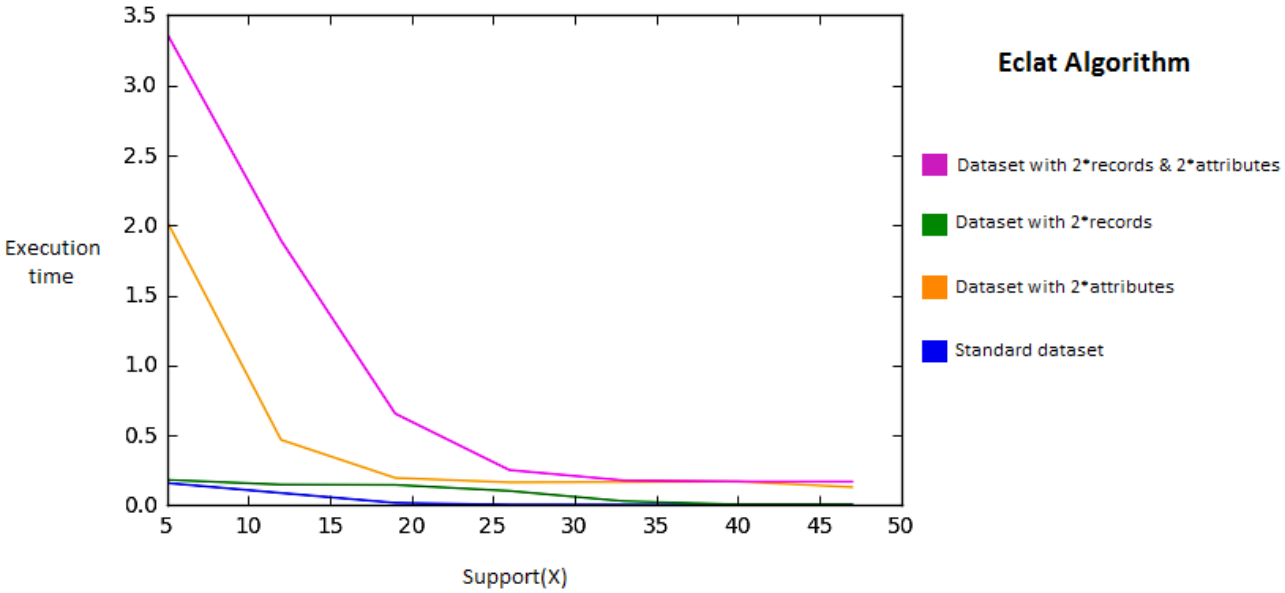
- **Apriori**



- **FPGrowth**



- **ECLAT**



REFERENCES

[1] Dr. Kanwal Garg, Deepak Kumar - "Comparing the Performance of Frequent Pattern Mining Algorithms", International Journal of Computer Applications (0975 – 8887), Volume 69– No.25, May 2013.

[2] Cornelia Győrödi, Robert Győrödi, Prof. Dr. Ing. Stefan Holban - "A Comparative Study of Association Rules Mining Algorithms".

[3] "Association Analysis: Basic Concepts and Algorithms"

[4] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg - "Top 10 algorithms in data mining" [5] Data Mining - Concepts and Techniques by Jiawei Han and Micheline Kamber, 2nd edition

[6] Data Mining Techniques by Arun K. Pujari

[7] Christian Borgelt's library PyFIM

[8] Orange - A Python package

[9] Github, Kaggle, etc for datasets

[10] Google

[11] Wikipedia

[12] Quora and a few other sources on the net