Prediction of Fastest Association Rules Mining Algorithm for any Dataset

Manish Sharma
2014A3PS181P

Deepankar Jain
2014A3PS225P

Prangav Singhal
2014A8PS332P

Ravindra Goyal
2014A3PS199P

Deepak Kumar Kar
2015A7PS129P

ABSTRACT

Frequent pattern mining is the most researched field in data mining. This paper provides comparative study of fundamental algorithms and performance analysis with respect to execution time. There are three widely-used algorithms of frequent pattern mining the algorithm, namely Apriori algorithm, FP-Growth algorithm and ECLAT algorithm. The Apriori based algorithm uses generate and test strategy approach to find frequent pattern by constructing candidate items and checking their counts and frequency from transactional databases. The FP-Growth algorithm uses a text only approach. There is no need to generate candidate item sets. FP-Growth algorithm extracts the frequent itemsets from FP-Tree directly. Most of the tree based structure allows efficient mining with single scan over the database.

Categories and Subject Descriptors

I.2 [Artificial Intelligence]: Miscellaneous; H.2.8 [Database Applications]: Data Mining

General Terms

Algorithms

Keywords

data mining, association rules, classification, apriori, FP-Tree, ECLAT

1. INTRODUCTION

Association Rule Mining (ARM) is one of the most important canonical tasks in data mining and probably one of the most studied techniques for pattern discovery. Association rules express associations between observations in transactional databases and the process of extracting them consists of initially finding frequent itemsets. Then Association Rules are extracted from these frequent itemsets.

1.1 BACKGROUND

There are three algorithms that are widely used for Association Rules Mining. They are - Apriori, FP-Growth and ECLAT. However, none of them have been proved to be better than the others. Frequent itemsets are all the sets that contain the item with minimum support The Apriori algorithm uses a “bottom up” approach, where frequent itemsets are extended one item at a time in a step known as candidate generation and groups of candidates are tested against the data. This requires multiple database scans. FP-Growth generates a data structure called the FP-Tree and it extracts the frequent itemsets from the FP-Tree directly. Unlike Apriori, it requires “only two” database scans. Both Apriori and FP-Growth use horizontal data format. Data can be represented as vertical format also, i.e., for each item we store the list of transaction

IDs that contain that item. Now to get the transactions ID list for a itemset, we intersect the transactions ID list for each item in the itemset. ECLAT then eliminates those itemsets that are not frequent. This is repeated until there are no candidate frequent itemsets. ECLAT is similar to Depth-First Search and it is faster than Apriori in certain conditions.

1.2 MOTIVATION

The three algorithms have their own benefits for different datasets. The speed of algorithm largely depends upon type of data. Our purpose is to find the best algorithm for a given dataset. This will reduce the processing time by a significant amount by preprocessing the best algorithm for given dataset. The motivation behind this idea is that data scientists may have to spend less time thinking about which algorithm to use for a datasets because if they choose the wrong algorithm then the processing time and resources will increase significantly.

1.3 OBJECTIVE

The main objective of this paper is to create a model that will predict the fastest algorithm for a given dataset using some metadata that will be extracted from the dataset. For each dataset in the training data, we will run all the three algorithms and store their running time. Then we will train our algorithm to predict the running time of an algorithm among given three algorithms for another dataset and it will output the fastest algorithm for that given dataset using its metadata.
The metadata of a dataset may include many features like number of transactions, average length of a transactions, total number of items etc.

2. RELATED WORK

In the field of Association Rule Mining(ARM), significant amount of work has already been done with respect to a number of these association rules algorithms implemented. For our work we are mainly concerned with FP-Growth[1], Apriori[2] and Éclat Algorithms[3]. As discussed above all algorithms are good enough for the rule creation job, but none of them is the best choice for a data miner keeping the time constraint in mind.

3. DATASETS USED

We have created and used 250 synthetic datasets with varying number of transaction lengths, number of unique items and total number of transactions, these all were generated randomly within some range.
We varied the support values for all these datasets to finally generate 2250 datasets for our work.

4. PROPOSED TECHNIQUES & ALGORITHMS

We have proposed the technique to find the best algorithm based on time constraint. We have extracted certain features like number of transactions, average length of transactions, number of unique items, largest and smallest size of transactions etc.
We ran the ARM algorithms on these datasets to find the algorithm which takes the least amount of time. For a particular dataset we assigned the best algorithm possible as the target class. This was done using various classification algorithms namely- C4.5, K-Nearest Neighbour[4], Ada-Boost ,Gaussian Process[5] and Support Vector Machines[6].
We applied Optimal Bayesian Technique to improve the accuracy in this work. Now based on the above features for any given dataset, we will be able to get the fastest algorithm out of the three.

5. EXPERIMENT AND RESULT

The created dataset was divided into training and testing data in 80:20 ratio. Then the testing data was used to train classifiers by the algorithms mentioned above. As the data used was synthetic, the accuracies achieved were not very great. To achieve higher accuracies, the optimal Bayesian technique was used. The results achieved are as follows:

C4.5	SVM	K-Nearest Neighbors	Gaussian Process	Adaboost	**Optimal Bayesian**
0.510	0.601	0.527	0.551	0.571	**0.651**

6. CONCLUSION AND FUTURE WORK

We have successfully presented a technique using a classifier which can predict the fastest ARM algorithm with a good degree of accuracy.

We have used mainly synthetic datasets so the accuracy judged is a bit low, but it can be increased by using real life datasets.

We can add more ARM algorithms which will make it more comprehensive. Also we can integrate this into a program which will automatically judge the fastest Association algorithm on its own without user approval.

8. REFERENCES

Main Paper-Metanat HooshSadat et al,"Fastest Association Rule Mining Algorithm Predictor (FARM-AP)" [1]Cristian Borgelt,"An Implementation of the FP-growth"

[2]Rakesh agarwal et al,"Fast Algorithms for Mining Association Rules"

[3]Xiaomei Yu, Hong Wang,"Improvement of Eclat Algorithm Based on Support in Frequent Itemset Mining"

[4]Min-Ling Zhang, "ML-KNN: A lazy learning approach to multi-label learning"

[5]C.K.I. Williams, D Barber,"Bayesian classification with Gaussian processes"

[6]Chih-Wei Hsu et al,"A Practical Guide to Support Vector Classification"

[7]PyFIM - http://www.borgelt.net/pyfim.html

[8]SciKit-Learn - http://scikit-learn.org/stable/