Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 49 (2015) 104 – 110

Procedia Computer Science

ICAC3'15

Comparative Study of Techniques to improve Efficiency of Association Rule Mining

Meera Narvekar,Apurva Kulkarni, S.S.Santha

D J Sanghvi college of Engineering,Vile Parle(w),Mumbai 400056,India
D J Sanghvi college of Engineering,Vile Parle(w),Mumbai 400056,India

Abstract

In the field of data mining, deriving set of rules from large dataset is primary objective. Association rule mining plays very important role to discover interesting rules according to domain knowledge specified as a minimum support threshold. The efficiency of association rule algorithm can be increased by focusing on reducing number of passes, sampling, parallel execution and constraint based association rule mining. In this paper various ways to enhance performance of association rule mining are discussed.

Peer-review under responsibility of organizing committee of the 4th International Conference on Advances in Computing, Communication and Control (ICAC3'15)

Keywords: Association Rule Mining,FP Tree.

1. Introduction

Association rules Mining is one of the core method used for mining frequent and interested patterns from large database. Database is collection of transaction and transaction describes interested items and their occurrences. Mining is used to discover rules from large dataset. Rule defines dependency between different items. For example, consider a database which describes properties of different animals and then categorizing them into animal type like mammals, amphibian, fish, bird etc. From database we can derive "If animal contains property of producing milk then it belongs to class mammal". Above statement can be structurally adjourn into the 'If' part and 'Then' part, where 'if' part denotes the condition and 'then' part states the result. Association Rule Mining is methodology to describe such facts or rules from database.

Peer-review under responsibility of organizing committee of the 4th International Conference on Advances in Computing, Communication and Control (ICAC3'15)
doi:10.1016/j.procs.2015.04.233

2.Background

Let I={ $I_1, I_2, I_3 \ldots I_n$ } be item set.T={$T_1, T_2, T_3 \ldots T_n$ } be collection of transaction i.e. dataset .Association rule mining generates the rule in the form of A → B, where first part of the rule is antecedent and later part is consequent. Here A can be set of items from item set and B can be one or more resultant items. Here A∩B=Ø that is A and B should be unique items. There are some interesting concepts like support and confidence are applicable to generate rules. The performance of association rule mining can be improved in four ways [9], Fig 2:

- By reducing the number of passes over the database
- By sampling the database
- By adding extra constraints on the structure of patterns
- Through parallelization.

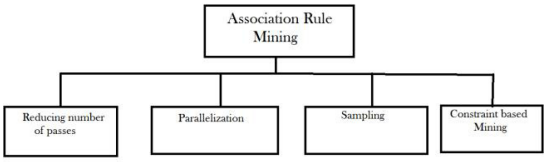

Fig 1: Performance improvisation of association rule mining

3. Different Approaches to improve Efficiency of Association Rule Mining Algorithm

3.1Reducing Number of passes :

Mining is a process to discover interesting rules from large dataset. To reduce repeated scanning of dataset, we can use data structure to store dataset which can provide efficient access to dataset. Following data structure can be used to build dataset.

Red Black Tree:

The Red Black Tree is based on the basic concept of binary search tree with some additional constraints. The binary search tree takes O(h) time to insert, delete or search the element, here h is nothing but height of the tree. As binary search tree is not balanced search tree it has maximum height as n which forms skewed binary tree. Hence the worst case complexity of binary search tree is O(n).

Red Black Tree[3] is a self balancing binary search tree. Apart from the constraints imposed on binary search tree, red black tree shares some more constraints. Every node is either red or black. Root node is always black. All the leaf nodes are black. If parent node is of red colour then both children will share black colour but inverse is not mandatory. For all nodes, all paths from the node to decedent leaves contain the same number of black nodes. The important characteristic of red black tree is that the longest path from the root to any leaf will not be greater than twice the shortest path in that tree. The above constraints make red black tree almost binary tree, which leads to minimum height to be log(n+1) and maximum height will be 2log(n+1). Hence the worst case complexity for insertion, deletion and searching will be (2log(n+1)) i.e. O(log(n)).