TDT4300 Datavarehus og datagruvedrift
Assignment 2: Apriori Rule Generation

Magnus L Kirø

February 28, 2013

BruteForce

The brute force way is basically three steps. 1: First we split the data input sets to get all the singletons. All possibilities.
2: Then we calculate the confidence and support for all the possibilities.
3: And lastly we remove all the elements that does not surpase the minimum supported limit.
(4: We recombine the sets to create rules.)

In the code this is done mainly with two methods. getSingles and removeUnsupported. getSingles creates all the possible sets. While removeUnsupported removes singeltons that does not exceed the minimum supported level. The minimum confidence level is also checked.

The minimum confidence level is set to 0.8. After some undecisive testing this seemed to be quite a good number.

In the smal dataset the brute force approach gets 3 more levels than the other two methods.

While in the big dataset we get 122 candidates, possible rule creating sets, on the first level. And 7381 candidates on the second. Later levels were skipped du to computational time and resource limitations.

The number of possible candidates will increase exponentially without pruning the data tree.

FKMinus1F1Apriori

This methot checks if two sets has any common elements. If there are common elements, an intersection, the combination is disregarded, prunde away.

If the two sets has nothing in common the union of the two sets are added to as a new set and the thresholds for the new combination is checked.

This gives us all possible permutations in a more efficient way.

FkMinus1FKMinus1

We find the difference set of the two given sets. This is the elements that does not occure in both the initial sets. Then we check for all elements in the difference set, to see in the first initial set + the vurrent element creates a new set that has a size that's equal to the size of the first initial set + 1.

If this is the case we have a new set and we add that to the collection of combinations. Also calculating the thresholds.

This is an even more efficient way of finding all the different combinations. As the iterative way of trying "currentSet + oneElement" to create all combinations it's efficient. And there are fewer levels created this way.

generateRulesBase

For the rule creation we use a simple way of creating rules. The association rule consists of the set of frequent items, the consequent item set and the two calculated thresholds.

The combination of support threshold and confidence threshold will give quite a unique rule. The support threshold is calculated by providing the union of the frequent set and the consequent set to the supportCach.get() method, which returns the threshold for a given set.

Pruning Table

(algorithm, levels, levels)

Algorithm	dataset 1	dataset 2
BruteForce	6	-
FKMinus1F1Apriori	3	2
FkMinus1FKMinus1	3	2