Stream Project

High Performance Computing and Data Engineering

Semester V
2016-2017

Mentor: Dr. Ravindranath Chowdary C
Name: Deepak Yadav
Roll No: 14075020
B-Tech Part III
Computer Science and Engineering

Association rule mining with mostly associated sequential patterns

Objective

Mining of structured data to find potentially useful patterns by association rule mining.

Abstract

Different than the traditional find-all-then-prune approach, a heuristic method is implemented to extract mostly associated patterns. This approach utilizes a maximally-association constraint to generate patterns without searching the entire lattice of item combinations. This approach does not require a pruning process. The proposed approach requires less computational resources in terms of time and memory requirements while generating a long sequence of patterns that have the highest co-occurrence. Furthermore, k-item patterns can be obtained thanks to the sub-lattice property of the MASPs. In addition, the algorithm produces a tree of the detected patterns; this tree can assist decision makers for visual analysis of data. The algorithm is tested on three datasets – Pumsb, Connect, Blog.

What does ARM do?

Discover hidden rules among enormous pattern combination based on individual and conditional frequencies.

Traditional Association Rule Mining

- Generate all possible patterns from data while pruning out non-frequent ones.
- Produce rules from frequent patterns.
- Apply some interesting measures to obtain interesting rules that can be used in decision making.

Basics of Association Rule Mining

In association rule mining, data to be mined is stored in the form of transactions. A transaction t is composed of some items in the form of (attribute = value) pairs as in t = {Age = Young, intoxicated = Yes, Day = Friday}. The goal of the ARM is to discover non-trivial patterns hidden in the transactional data set. An association rule A => C has two sets of items (itemsets), namely the antecedent A and the consequent C. An itemset of A U C is considered to be a rule if its frequency (Support(itemset) = P(itemset)) satisfies a minimum support threshold and the conditional probability P(C|A) satisfies a minimum confidence threshold. The support can be considered as a global measure of being interesting, and the confidence is used as a localization measure.

Mostly Associated Sequential Patterns (MASP)

In case of MASPs no need to search entire lattice because one more constraint other than the default has been imposed. MASPs imposes interestingness constraint on patterns to detect highest co-occurence without searching all possible combinations. During search process, MASP tree will be formed. MASP has sub-lattice property that reveals k-item rules from MASP.

MASP+

T-ARM can be conducted within small dataset of each MASP. The combination of rules obtained from T-ARM and MASP give rise to MASP+ approach.

Terminology

Transactional Data : A set of attribute-value pairs. e.g. t = {Age:Young, Day:Sunday}

A rule in the form **A -> B** is an association rule if its support and confidence crossed the thresholds.

Support : Probability of occurence of AUB together in the complete data set.

Confidence : The conditional probability P(B|A) i.e. probability of getting B when A is given.

Format of input data

A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
V11	V21	V31	V41	V51	V61	V71	V81	V91
V11	V22	V32	V41	V51	V62	V71	V81	V92

Where **Ai** are columns names and **Vij** is the value of column **Ai**

How to generate MASP tree ?

Notation

Let X and Y are two sets then
X (Y implies X is a subset of Y
X\Y implies set of elements of X which are not in Y
|X| = number of entries in set X
If D is a data table:

One can say that it is composed of cells C_{ri} = (A_i = V_{ij}) where r = 1, 2, ….., |D| and i = 1,2,......, |A|. J = 1, 2, ,3,......., |unique entries in the ith column|

An item I is unique attribute-value pair I = V_{ij}

Condition for MASP

A set M = {I_1, I_2, I_3, ……………, I_k, ……………., I_K} will be MASP iff for all k belongs to {1, 2, 3, ……….., K}

1. P(I_1 , I_2, I_3, ……., I_k) >= (threshold value of support)
2. P(I_k | I_1 , I_2, I_3, ……., I_{k-1}) >= (threshold value of confidence)
3. P(I_k | I_1 , I_2, I_3, ……., I_{k-1}) must be maximum

Block (M) and Counter Block (M)

M = {V_{11}, V_{23}, V_{42}}
Attr(D) = A_1, A_2, A_3, A_4, A_5, A_6
B(M) = {select A3, A5, A6; FROM D; where (A1 = V11 && A2 = V23 && A4 = V42)}
CB(M) = {select A3, A4, A5, A6; FROM D; where (A1 = V11 && A2 = V23 && A4 V42)}

How to generate rules using MASP Tree ?

If MASP M = {I_1, I_2, I_3, …. , I_K} then rules will be
(I_1, I_2, I_3, …. ,I_{K-1}) -> I_K
(I_1, I_2, I_3, …. ,I_{K-2}) -> I_{K-1}
(I_1, I_2, I_3, …. ,I_{K-3}) -> I_{K-2}
………………………..
………………………..
………………………..
I_1 -> I_2

Additional rule

(I_1, I_2, ……, I_{k-1}) -> (I_k, I_{k+1}, ……, I_K) satisfies the minimum support condition. If it will satisfy the minimum confidence then it will be included in the rule set.

MASP+ patterns

The combination of MASP and the rules obtained from its block is also a rule.

MASP M = {V11, V41, V31}
Rules from its block {V21, V61 -> V51; V22 -> V53}

New rules {((V11, V41, V31), V21, V61) -> V51, ((V11, V41, V31), V22) -> V53}

Results on different datasets - Expected vs Observed longest rule size

Connect

Threshold support	Threshold confidence	Expected	Observed
0.1	0.8	26	27
0.1	0.5	28	29
0.1	0.25	28	29
0.001	0.8	26	27
0.001	0.5	36	37
0.001	0.25	36	39

Pumsb

Threshold support	Threshold confidence	Expected	Observed
0.1	0.8	19	20
0.1	0.5	29	30
0.1	0.25	29	30
0.001	0.8	19	20
0.001	0.5	43	44
0.001	0.25	45	49

Blog

Threshold support	Threshold confidence	Expected	Observed
0.1	0.8	16	17
0.1	0.5	16	34
0.1	0.25	33	34
0.001	0.8	16	17
0.001	0.5	50	50

0.001	0.25	50	59

Thank You