Hyper clique Pattern Discovery

DMG ASSIGNMENT 3

Prashant Pathak
MT19051
DMG
22-Oct,2019

Table of Contents

Notation used in the report

- I = { i_1 , i_2 , i_3 ..... , i_n } be the number of the item present in the dataset.
- T = { t_1 , t_2, t_3, ... , t_n } be the set of all transaction.
- I is a subset of I and is called itemset.
- K-itemset means that it has k items.
- Transaction width is the number of item present in a transaction

Issues with "Support" as a measure

In Apriori algorithm we use the support as the measure of for finding and pruning the frequent itemset. Suppose we have to find support of x. Number of elements that has the query item (x), from the complete transaction is defined as the support of the transaction.

$$\text{Supp(x)} = |\{t_i | x \subseteq t_i, t_i \in T\}|$$

Support based pruning is not sufficient to prune highly skewed data because of the following reason:

1) If the minimum support threshold is low, we may extract too many spurious patterns involving item with substantial difference support level. These are called weakly- related cross support pattern. Ex {Caviar, Milk}
2) If the minimum support threshold is high, we may miss many interesting patterns occurring at low level of support. Ex {gold, earning} these are usually the expensive element.

Because of the above issue with support we have to some define some new type of measure. We cannot define the measure arbitrarily it must have some property which support has. Support was favorable to us because of its anti-monotone property i.e. subset of set has higher measure of the value than the set. Because of this property in Apriori algorithm we don't have to search the entire search space rather we search the a smaller search space which is sure to give us correct answer.

So we define a new measure called h-confidence.

h-confidence

Suppose P = { i_1 , i_2 , i_3 ..... , i_n } . Mathematically H -conf is defines as:

Hconf(p) = min {conf {i_1 -> i_2, i_3, ..., i_m}, conf {i_2 -> i_1, i_3, ..., i_m}, ..., conf {i_m -> i_1, i_2, ..., i_{m-1}}}

Here conf shows the confidence of an item in conf = $\frac{\text{support}(x \cup y)}{\text{support(x)}}$.

Hconf can be also calculated with help of support.

$$\text{Hconf(p)} = \frac{\text{support}(P)}{\max\{\text{support(i1)},\text{support(i2)},.....\text{support(i3)}\}}$$

Hconf is a good measure and a substitute for Support because it has the anti-monotone property.

According to Anti monotone if the h-confidence of an itemset P is greater than a user- specified threshold, so is every subset of P.

Drawback id Hconf: hyper clique mining framework may miss some interesting pattern too. {A, B, C} may have very low- confidence and yet may be interesting if one of its rules, satisfy AB - > C has very high confidence.

Besides the anti-monotone property of Hconf. Hconf has one more interesting pattern called Cross support property.

Cross-Support property

Given a threshold, t, a pattern P is a cross- support pattern with respect to t if P contains two item x and y such that $\frac{Supp(\{x\})}{Supp(\{y\})} < t$ where 0 < t < 1.

Cross support property ensures that an item set is not spurious i.e. support of two elements in an item set do not vary too much.

H-confidence has this property so it we can prune with help of it and can generate the correct pattern.

Plots of Hconf with Anti-monotone and Cross-Support

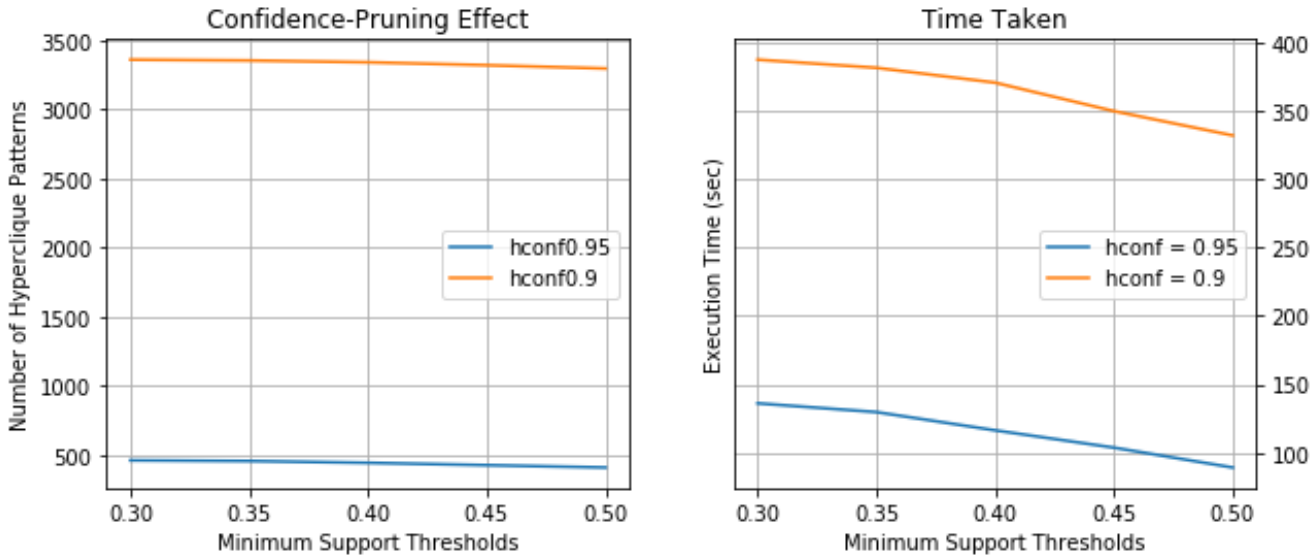


Figure 1 plot showing the result in pumsb data set

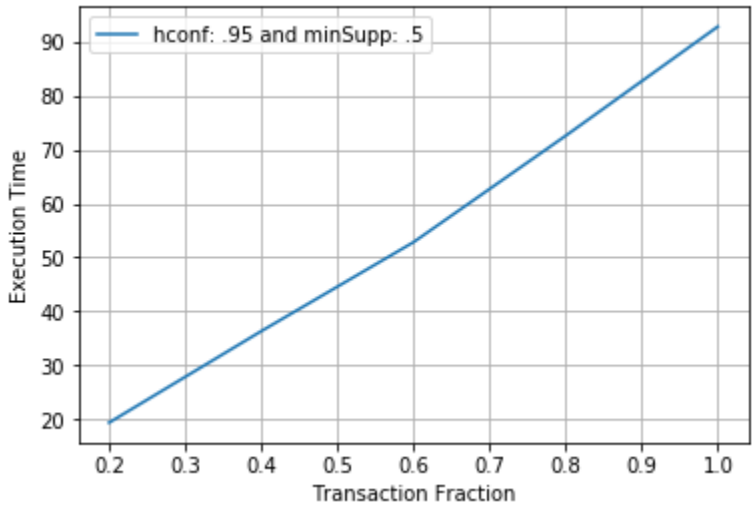
```
For Hconficence value  of  0.95
No of Pattern    Execution time
464              136.1045846939087
459              129.70216536521912
446              116.37369203567505
429              103.73913192749023
412              89.22113084793091


For Hconficence value  of  0.9
No of Pattern    Execution time
3360             387.0953860282898
3354             381.1484923362732
3341             370.2861473560333
3320             349.5009272098541
3296             331.74887132644653
```

Figure 2 Actual value of the output

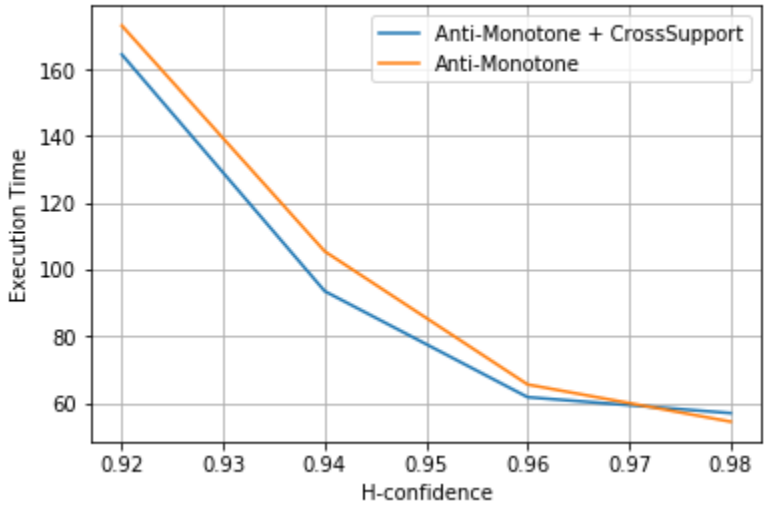
The above graph shows that we get less hyper clique pattern at lower h-confidence value. Then execution time of the code also increased as we decrease the h-confidence value. This is because we get more pattern when we reduce the h-confidence value. But Compared to the Apriori algorithm this take less time.

Plot Between No of Transaction and Time taken

As the number of transactions increase the time taken to execute it will automatically increase. This was clear intuitively and same was shown by the graph.



Plot with Cross-Support and Without Cross-Support



The time with Cross support is less compared to without CrossSupport.

References

1) https://adataanalyst.com/machine-learning/apriori-algorithm-python-3-0/ : This link was used to take idea of Apriori algorithm.