CS4445 B10 Homework 4 Part I Solution

Yutao Wang

Consider the zoo.arff dataset converted to arff from the Zoo Data Set available at Univ. of California Irvine KDD Data Repository.

1. Load this dataset onto Weka. Remove the 1st attribute (animal_name) which is a string. Go to "Associate" and run Apriori with "numRules = 30", "outputItemSets = True", "verbose = True", and default values for the remaining parameters.

2. Now run Apriori with "numRules = 30", "outputItemSets = True", "verbose = True", treatZeroAsMissing = True", and default values for the remaining parameters.

1. [5 points] What difference do you see between the rules obtained in Parts 1 and 2 above? Explain.
Part 2 doesn't generate rules with value 0 while part 1 does. Since 0 is a more common (frequent) value than 1, then part 1 generates a much larger number of rules (if Weka didn't limit the output to the first 30 rules generated), and these rules will tend to have many more occurrences of 0 values than of 1 values.

2. [5 points] From now on, consider just the second set of rules (that is, when "treatZeroAsMissing = True"). Find an association rule you find interesting and explain it. Include the confidence and support values in your explanation of the rule.
For rule: X->Y
Confidence: c(X->Y) = σ(XUY)/ σ (X) = P(X and Y)/P(X)
Support: s(X->Y) = σ(XUY)/N = P(X)
Take rule #15 as an example:
hair=1 backbone=1 39==> milk=1 39 <conf:(1)> lift:(2.46) lev:(0.23) [23] conv:(23.17)
The support value is 39/110, and confidence value is 39/39 = 1. This means that all animals in the dataset who have hair and a backbone, produce/drink milk.

3. [10 points] What are "lift", "leverage", and "conviction"? Provide an explicit formula for each one of them (look at the Weka code to find those formulas). Use the values of these metrics for the association rule you chose in the previous part to judge how interesting/useful this rule is.
Lift (X->Y)= c(X->Y)/s(Y) = P(X and Y)/P(X)*P(Y) = P(Y|X)/P(Y).
Leverage (X->Y) = P(X and Y)-P(X)*P(Y)
Conviction(X->Y) = P(X)*P(~Y)/P(X and ~Y)
In the previous example, lift = 2.46>1 indicates that having hair and backbone increases the probability of producing/drinking milk by a factor of 2.46. Leverage = 0.23 >0. According to Weka "… Leverage is the proportion of additional examples covered by both the premise and consequence above those expected if the premise and consequence were independent of each other. The total number of examples that this represents is presented in brackets following the leverage. Conviction is another measure of departure from independence." Hence, in this example, the antecedent and the consequent of the rule cover 23 instances more than expected if they were independent. This provides additional evidence that the antecedent and the consequent of this rule are not independent from each other. The high value of Conviction (23.17) reinforces this point as well.

4. Look at the itemsets generated. Let's consider in particular the generation of 5-itemsets from 4-itemsets:

Minimum support: 0.35 (35 instances)
...
Size of set of large itemsets L(4): 8
Large Itemsets L(4):
hair=1 milk=1 toothed=1 backbone=1 38
hair=1 milk=1 toothed=1 breathes=1 38
hair=1 milk=1 backbone=1 breathes=1 39
hair=1 toothed=1 backbone=1 breathes=1 38
milk=1 toothed=1 backbone=1 breathes=1 40
milk=1 backbone=1 breathes=1 tail=1 35
toothed=1 backbone=1 breathes=1 legs=4 35
toothed=1 backbone=1 breathes=1 tail=1 38
Size of set of large itemsets L(5): 1
Large Itemsets L(5):
hair=1 milk=1 toothed=1 backbone=1 breathes=1 38

1. [5 points] State what the "join" condition is (called "merge" in the Fk-1xFk-1 method in your textbook p. 341). Show how the "join" condition was used to generate 5-itemsets from 4-itemsets. (Warning: not all candidate 5-itemsets are shown above.)
Join condition is: Merge two (k-1)-itemsets into a k-itemset if their first (k-2) items are identical.
Here, merge
(1)
hair=1 milk=1 toothed=1 backbone=1 38
hair=1 milk=1 toothed=1 breathes=1 38
into
hair=1 milk=1 toothed=1 backbone=1 breathes=1
(2)
toothed=1 backbone=1 breathes=1 legs=4 35
toothed=1 backbone=1 breathes=1 tail=1 38
into
toothed=1 backbone=1 breathes=1 legs=4 tail=1

No other pair of frequent 4-itemsets satisfies the merge condition, so no more candidate 5-itemsets are generated.

2. [5 points] State what the "subset" condition is (called "candidate pruning" in the Fk-1xFk-1 method in your textbook p. 341). Show how the "subset" condition was used to eliminate candidate 5-itemsets from consideration before unnecessarily counting their support.
Subset condition: check all the subsets of the resulting k-itemset that contain (k-1) items to see if they all are frequent (that is, have enough support). If at least one of these subset is not frequent, then the k-itemset cannot be frequent (due to that apriori principle).

For these two itemsets:
hair=1 milk=1 toothed=1 backbone=1 breathes=1

all subsets of 4 items in this itemset are frequent

hair=1 milk=1 toothed=1 backbone=1 38

hair=1 milk=1 toothed=1 breathes=1 38

hair=1 milk=1 backbone=1 breathes=1 39

hair=1 toothed=1 backbone=1 breathes=1 38

so this itemset is a candidate frequent itemset whose support will need to be calculated by scanning the dataset to determine whether or not if it is indeed frequent.

toothed=1 backbone=1 breathes=1 legs tail=1

Note that "backbone=1 breathes=1 legs=4 tail=1" doesn't appear on the list of frequent 4-itemsets, and therefore it is not frequent. Hence the 5-itemset cannot be frequent.

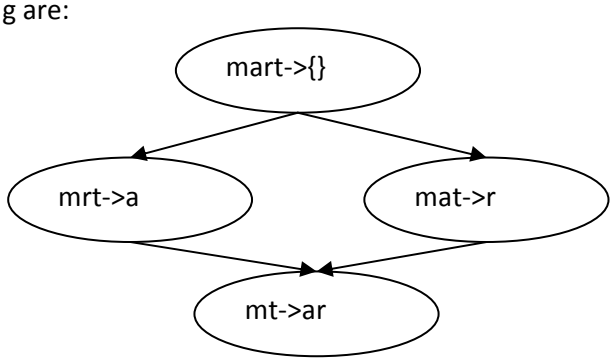
5. [10 points] Consider the following frequent 4-itemset:

milk=1 backbone=1 breathes=1 tail=1

Use Algorithms 6.2 and 6.3 (pp. 351-352), which are based on Theorem 6.2, to construct all rules with Confidence = 100% from this 4-itemset. Show your work by neatly constructing a lattice similar to the one depicted in Figure 6.15 (but you don't need to expand/include pruned rules).

Let's use the following notation: m=milk, a=backbone, r=breathes, t=tail:

conf(art->m) = 35/60 = 0.5833<1, so prune all rules containing item m in their consequent.
conf(mrt->a) = 35/35 = 1
conf(mat->r) = 35/35 = 1
conf(mar->t) = 35/41 = 0.8537 <1, so prune all rules containing item t in their consequent.
conf(rt->ma): pruned
conf(at->mr) : pruned
conf(ar->mt) : pruned
conf(mt->ar) = 35/35 = 1
conf(mr->at) : pruned
conf(ma->rt) : pruned
conf(t->mar): pruned
conf(r->mat): pruned
conf(a->mrt): pruned
conf(m->art): pruned
The final rules after pruning are:



3. [5 points] Explain how the processs of mining association rules in Weka's Apriori is performed in terms of the following parameters: lowerBoundMinSupport, upperBoundMinSupport, delta, metricType, minMetric, numRules.

Since it might be difficult for a user to figure out a good value for minimum support so that sufficient association rules are produced (not too many or too few), Weka provides the ability to state the number of association rules desired instead of a min. support value. Then, it follows a process like this:

min. support = upperBoundMinSupport
repeat
    mine association rules with this min. support and the minMetric (say confidence) threshold values.
    When/if at least numRules are generated, then return them and exit
    If not enough rules were generated,
        then decrease the min. support: min.support = min. support - delta
until min. support == lowerBoundMinSupport

4. [10 points] Exercise 16, p. 411 of the textbook.

(a) Range: (-∞, 1]. Here is why:

Note that M = $\frac{P(B|A)-P(B)}{1-P(B)} = \frac{P(B|A)}{1-P(B)} - \frac{P(B)}{1-P(B)} \leq \frac{1}{1-P(B)} - \frac{P(B)}{1-P(B)} = 1$ since $P(B|A) \leq 1$.

M will be equal to 1 when P(B|A) = 1.

On the other hand, note that when P(B|A) = 0, then M = $\frac{-P(B)}{1-P(B)}$. In the limit when P(B) goes to 1, M goes to -∞.

(b) $M = \frac{P(B|A)-P(B)}{1-P(B)} = \frac{\frac{P(AB)}{P(A)}-P(B)}{1-P(B)}$. So when P(AB) is increased while P(A) and P(B) remain unchanged, M will increase.

(c) As shown in the previous formula, when P(A) is increased while P(A,B) and P(B) remain unchanged, M will decrease.

(d) $M = \frac{P(B|A)-P(B)}{1-P(B)} = \frac{1-P(B)+P(B|A)-1}{1-P(B)} = 1 + \frac{\frac{P(AB)}{P(A)}-1}{1-P(B)}$. So when P(B) is increased while P(A,B) and P(A) remain unchanged, M will increase.

(e) $M(A \rightarrow B) = \frac{P(B|A)-P(B)}{1-P(B)} = \frac{P(AB)-P(A)P(B)}{P(A)-P(A)P(B)}$, $M(B \rightarrow A) = \frac{P(AB)-P(A)P(B)}{P(B)-P(A)P(B)}$, so M is not symmetric.

(f) When A and B are independent then $M = \frac{P(B|A)-P(B)}{1-P(B)} = \frac{P(B)-P(B)}{1-P(B)} = 0$.

(g) No, it is not null-invariant.

$M = 1 + \frac{\frac{P(AB)}{P(A)}-1}{1-P(B)} = 1 + \frac{\frac{f_{11}}{f_{1+}}-1}{1-\frac{f_{+1}}{N}}$, when f_{00} increases then N will increase while all the other parts in the right hand side of the equation will remain the same, so M will increase.

(h) No, M is not invariant under row/column scaling.

$M(A' \rightarrow B') = \frac{P(A'B')-P(A')P(B')}{P(A')-P(A')P(B')} = \frac{N*f'_{11}-f'_{1+}*f'_{+1}}{N*f'_{1+}-f'_{1+}*f'_{+1}} =$

$\frac{N*k_1*k_3*f_{11}-(k_1*k_3*f_{11}+k_2*k_3*f_{10})*(k_1*k_4*f_{01}+k_1*k_3*f_{11})}{N*(k_1*k_3*f_{11}+k_2*k_3*f_{10})-(k_1*k_3*f_{11}+k_2*k_3*f_{10})*(k_1*k_4*f_{01}+k_1*k_3*f_{11})} \neq \frac{N*f_{11}-f_{1+}*f_{+1}}{N*f_{1+}-f_{1+}*f_{+1}}$.

(i) Under inversion, A becomes ~A, B become ~B (where ~A is the complement of A), so:

$$M(A' \to B') = \frac{P(\sim B|\sim A)-P(\sim B)}{1-P(\sim B)} = \frac{1-P(B|\sim A)-1+P(B)}{1-P(\sim B)} = \frac{P(B)-\frac{P(\sim A|B)}{P(\sim A)}P(B)}{P(B)} = 1 - \frac{P(\sim A|B)}{P(\sim A)} = \frac{1-P(A)-1+P(A|B)}{1-P(A)} = \frac{P(A|B)-P(A)}{1-P(A)}$$

So after inversion the measure M(A'->B') equals M(B->A). So M is not invariant under the inversion operation.