<u>HOMEWORK 4</u>

<u>EXERCISES AND EXPERIMENTS ON ASSOCIATION ANALYSIS</u>

Submitted by

Arun Rajagopalan

A20360689

For CS422 Data Mining

Illinois Institute of Technology

EXERCISES AND EXPERIMENTS ON ASSOCIATION ANALYSIS

Exercise 1 (Chapter 6 #2):

(i) Compute the support for itemsets {e}, {b, d}, and {b, d, e} by treating each transaction ID as a market basket considering the data set shown below.

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

In association analysis, support for a particular itemset denotes the number of instances where the itemset is present (σ{itemset}) with respect to the total count of instances.

Support, s{itemset} = σ{itemset} / Number of instances

Hence support for itemset {e} denoted by s{e} is given by **s{e} = 8/10 = 0.8**
Similarly the support for itemset {b,d} is given by **s{b,d} = 2/10 = 0.2**
and support for the set {b,d,e} is given by **s{b,d,e} = 2/10 = 0.2**

(ii) Use the results in part (a) to compute the confidence for the association rules {b, d}->{e} and {e}->{b, d}. Is confidence a symmetric measure?

Support and confidence are two significant metrics for quantifying the reliability of association rules. Association rule can be modelled as an expression that implies U tends to be associated with V, where U and V are itemsets with no common items.

Support for a rule indicates the number of instances where the association rule holds good and confidence is the measure of frequency of instances with U that also contain V. The mathematical definitions for the metrics follow.

Support for rule U -> V, s(U -> V) = σ(U ∪ V) / Number of instances
Confidence c(U -> V) = σ(U ∪ V) / σ(U)

Therefore, the confidence of the rule {b,d} tends to be associated with {e}, denoted by c(bd->e) is given by **c(bd->e) = (2/10)/(2/10) = 100 %**
Also the confidence of the rule {e} tends to be associated with {b,d} is given by
c(e->bd) = (2/10)/(8/10) = 25%

From the result, since c(bd->e) != c(e->bd), it is clear that **confidence is not a symmetric measure.**

Exercise 2 (Chapter 6 #6):

(i) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support), considering the market basket transactions shown below?

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

The number of rules that can be formulated using a dataset, depends on the count of items in the dataset, which can be mathematically written as follows.

Number of Rules = 3^d – 2^(d+1) + 1 where d is the count of items.

For the given dataset, since the number of items is 6, rules that can be derived totals to 3^6 – 2^(7) +1 = **602 rules**.

(ii) Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.

The maximum number of k-itemsets that can be achieved from a dataset of size d is expressed as (d k) which is nothing but the number of combinations of size k, that a d sized dataset can give rise to. Thus the expression can be alternately written as ${}^{d}C_{k.}$

For the given dataset, the **maximum number of 3-itemsets is (6 3) = (6X5X4)/(1X2X3) = 120/6 = 20**

(iii) Find an itemset (of size 2 or larger) that has the largest support.

Instead of a brute force approach, we can think through the problem to get the answer with minimal computations. The 1-itemset with largest support in the given dataset is diapers (7). The second largest are Milk, Bread and Butter with support of 5 each. First let us find the doublets between these items and find their support.

s{Diapers, Milk} – 4
s{Diapers, Bread} – 3
s{Diapers, Butter} – 3

s{Milk, Bread} - 3
s{Milk, Butter} - 3
s{Bread, Butter} - 5

Thus the {Bread, Butter} itemset has a maximum support of 5 among the others. No other 2-itemset in the given dataset can reach this support, since their individual support does not amount to 5 (s{Beer} = 4 and s{Cookies} = 4). So, we can ignore all other 2-itemsets.

In case of other k-itemsets where k is larger than 2, the only 3-itemsets and 4- itemsets that can possibly have a support equal to 5 are as follows with their actual support values.

s{Diapers, Milk, Bread} – 2
s{Diapers, Milk, Butter} – 2
s{Diapers, Milk, Bread, Butter} – 2
s{Bread, Butter, Milk} – 3
s{Bread, Butter, Diapers} – 3

Hence, we can conclude that the candidate with maximum support is {Bread, Butter}, where number of items is greater than 1.

(iv) Find a pair of items, a and b, such that the rules {a}->{b} and {b}->{a} have the same confidence.

The relationship c(a->b)=c(b->a) will hold good for any pair of items a and b, if their individual supports are equal. Equal support leads to equal denominators in confidence expression and thus yield same confidence, given that the numerator is always equal for c(a->b) and c(b->a).

Thus, the item pairs that will satisfy the given relationship are (Bread, Butter), (Milk, Butter), (Milk, Bread) and (Beer, Cookies) which have same individual supports pairwise.

Exercise 3 (Chapter 6 #8):

Suppose the Apriori algorithm is applied to the data set shown below with minsup = 30%, i.e., any itemset occurring in less than 3 transactions is considered to be infrequent.

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

(i) Draw an itemset lattice representing the data set given above. Label each node in the lattice with the following letter(s):
N: If the itemset is not considered to be a candidate itemset by the Apriori algorithm.
F: If the candidate itemset is found to be frequent by the Apriori algorithm.
I: If the candidate itemset is found to be infrequent after support counting.

Using the given dataset, enumerating the support all k-itemsets where k ranges from 1 to 4,
s{a}=5, s{b}=7, s{c}=5, s{d}=9, s{e}=6

All of the 1-itemsets satisfy the minimum support count, so we take all of them for next level calculations.
s{a,b}=3, s{a,c}=2, s{a,d}=4, s{a,e}=4
s{b,c}=3, s{b,d}=6, s{b,e}=4

s{c,d}=4, s{c,e}=2, s{d,e}=6
At this level, we can prune {a,c} and {c,e} as infrequent since they have a support less than minimum support. All of supersets containing {a,c} and {c,e} can be ignored as infrequent rightaway by Apriori principle which states that if an itemset is infrequent, then all of its supersets will also be infrequent.

Proceeding to next level with only the valid 2-itemsets,
s{a,b,d}=2, s{a,b,e}=2, s{a,d,e}=4
s{b,c,d}=2, s{b,d,e}=4
At this level, we can prune all itemsets except {a,d,e} and {b,d,e}.

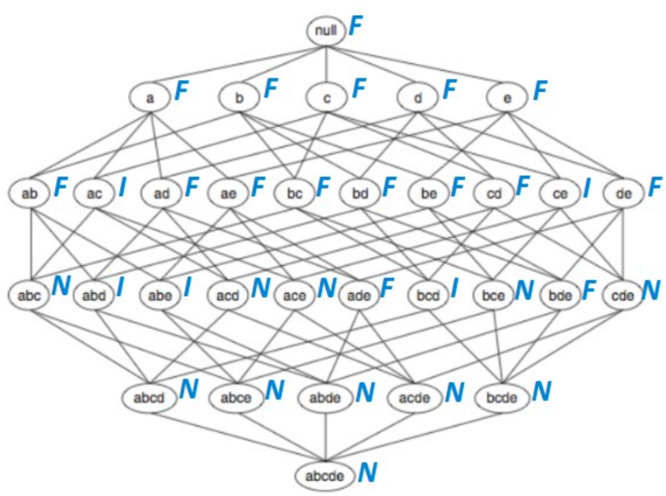
We need not calculate support for any more k-itemsets where k>3 since we know that all of them will be infrequent too by Apriori principle.

Thus, the frequent itemsets {null}, {a}, {b}, {c}, {d}, {e}, {a,b}, {a,d}, {a,e}, {b,c}, {b,d}, {b,e}, {c,d}, {d,e}, {a,d,e}, {b,d,e} can be assigned a label of F.

The infrequent itemsets {a,c}, {c,e}, {a,b,d}, {a,b,e}, {b,c,d} can be assigned the label of I.

All other itemsets {a,b,c}, {a,c,d}, {a,c,e}, {b,c,e}, {c,d,e}, {a,b,c,d}, {a,b,c,e}, {a,b,d,e}, {a,c,d,e}, {b,c,d,e}, {a,b,c,d,e} which were ignored as infrequent without calculating support can be assigned a label of N.

The resultant itemset lattice with the N, F and I labels added is as follows.



(ii) What is the percentage of frequent itemsets (with respect to all itemsets in the lattice)?

The percentage of frequent itemsets is given by the ratio of the same (number of F labels) and total number of itemsets.
Hence, **% of frequent sets = 16 / 32 = 50 %**

(iii) What is the pruning ratio of the Apriori algorithm on this data set?
(Pruning ratio is defined as the percentage of itemsets not considered to be a candidate because (1) they are not generated during candidate generation or (2) they are pruned during the candidate

pruning step.)

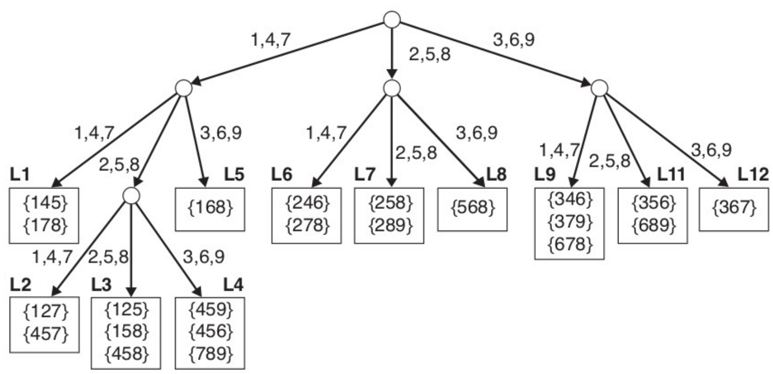
As per the definition, pruning ratio is the ratio between the number of itemsets that were ignored as infrequent without calculating support (number of N labels) and the total number of itemsets.
Thus, **pruning ratio = 11 / 32 = 34.375 %**

(iv) What is the false alarm rate (i.e, percentage of candidate itemsets that are found to be infrequent after performing support counting)?

The itemsets that were found to be infrequent after support counting are assigned a label of I. So, the false alarm rate can be found by the ratio of number of I labelled itemsets and the total itemset count.
False alarm rate = 5 / 32 = 15.625 %

Exercise 4 (Chapter 6 #9):

The Apriori algorithm uses a hash tree data structure to efficiently count the support of candidate itemsets. Consider the hash tree for candidate 3-itemsets shown below.



(i) Given a transaction that contains items {1, 3, 4, 5, 8}, which of the hash tree leaf nodes will be visited when finding the candidates of the transaction?

For the given transaction, subset enumeration can be done as follows.

Original Transaction				*[1 3 4 5 8]*		
Level 1		*1 [3 4 5 8]*		*3 [4 5 8]*		*4 [5 8]*
Level 2	*1 3 [4 5 8]*	*1 4 [5 8]*	*1 5 [8]*	*3 4[5 8]*	*3 5[8]*	***[4 5 8]***
Level 3	***[1 3 4]*** ***[1 3 5]*** ***[1 3 8]***	***[1 4 5]*** ***[1 4 8]***	***[1 5 8]***	***[3 4 5]*** ***[3 4 8]***	***[3 5 8]***	***[4 5 8]***

As per the given hash tree, the paths traversed by the final 3-itemsets that were achieved in Level 3 are as follows.

[1 3 4], [1 3 5], [1 3 8] – Branch left most at root, then branch right most at Level 1 node and reach L5.
[1 4 5], [1 4 8] – Branch left most at root, then branch left most at Level 1 node and reach L1.
[1 5 8], [4 5 8] – Branch left most at root, then branch middle at Level 1 node and branch middle again ending in L3.
[3 4 5], [3 4 8] – Branch right most at root, then branch left most at Level 1 node ending up at L9.
[3 5 8] – Branch right most at root, then branch middle in Level 1 node, reaching L11.

The leaf nodes visited were L1, L3, L5, L9 and L11.

(ii) Use the visited leaf nodes in part (b) to determine the candidate itemsets that are contained in the transaction {1, 3, 4, 5, 8}.

The candidate itemsets that were present in the transaction (The transaction subsets that had a match in the leaf nodes) **are {1,4,5}, {1,5,8} and {4,5,8}**.

Weka Experiments:

Association analysis is a means by which interesting and useful relations between several data facets that are not otherwise conspicuous could be revealed from massive datasets. To do this analysis now, we use the popular machine learning tool Weka with the supermarket dataset that is distributed along with Weka.

Supermarket dataset:
This dataset is a real world point of sale transaction data that was collected from a small New Zealand supermarket. There are 216 nominal attributes which represent the variety of departments that were available in the supermarket, apart from the class attribute 'total' which denotes whether the final sum price was higher or lower than 100$. The 4627 records represent the individual customer transactions. A value of 't' denotes the presence of item of that particular department in the market basket and the missing values indicate the items that the customer did not buy.

(i) Important parameters of Apriori algorithms.

Delta: This parameter takes the value of the factor for decreasing minimum support each iteration.
LowerBoundMinSupport: This value specifies the lower bound for minimum support. Set to 0.1 (10%) by default.
Metric Type: This parameter allows the user to choose the metric that the minMetric will work on. By default, confidence is set as the type.
MinMetric: This parameter takes the minimum value for the selected metric type.
OutputItemSets: When set to true, this parameter prints the itemsets also, which will be useful for analysis and debugging.
RemoveAllMissingCols: When set to true, all the columns with missing values are removed, thus helping in generating rules for an ideal scenario.
UpperBoundMinSupport: This value specifies the upper bound for minimum support. Set to 1.0 (100%) by default.
NumRules: This parameter tells Weka how many rules to print to the console.

(ii) Top 4 association rules using default parameters.

Main parameters with default value: ***Lower minimum support: 0.1; minimum confidence: 0.9***

1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723 conf:(0.92)
2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696 conf:(0.92)
3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705 conf:(0.92)
4. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746 conf:(0.92)

Where number of cycles were 17 and count of large itemset levels generated was 6.

(iii) Table 1.1 Documenting and analysing the association rule results for modified support and count values.

Lower Min Support (LMS)	Min Confi -dence (MS)	Cycle count	Large itemset levels	**Top 10 association rules**
Section 1: Low (lower minimum support) with high, medium and low (minimum confidence)				
0.1	0.8	14	3	1. biscuits=t vegetables=t 1764 ==> bread and cake=t 1487 conf:(0.84) 2. total=high 1679 ==> bread and cake=t 1413 conf:(0.84) 3. biscuits=t milk-cream=t 1767 ==> bread and cake=t 1485 conf:(0.84) 4. biscuits=t fruit=t 1837 ==> bread and cake=t 1541 conf:(0.84) 5. biscuits=t frozen foods=t 1810 ==> bread and cake=t 1510 conf:(0.83) 6. frozen foods=t fruit=t 1861 ==> bread and cake=t 1548 conf:(0.83) 7. frozen foods=t milk-cream=t 1826 ==> bread and cake=t 1516 conf:(0.83) 8. baking needs=t milk-cream=t 1907 ==> bread and cake=t 1580 conf:(0.83) 9. milk-cream=t fruit=t 2038 ==> bread and cake=t 1684 conf:(0.83) 10. baking needs=t biscuits=t 1764 ==> bread and cake=t 1456 conf:(0.83)
0.1	0.5	11	2	1. biscuits=t 2605 ==> bread and cake=t 2083 conf:(0.8) 2. milk-cream=t 2939 ==> bread and cake=t 2337 conf:(0.8) 3. fruit=t 2962 ==> bread and cake=t 2325 conf:(0.78) 4. baking needs=t 2795 ==> bread and cake=t 2191 conf:(0.78) 5. frozen foods=t 2717 ==> bread and cake=t 2129 conf:(0.78) 6. vegetables=t 2961 ==> bread and cake=t 2298 conf:(0.78)

				7. vegetables=t 2961 ==> fruit=t 2207 conf:(0.75) 8. fruit=t 2962 ==> vegetables=t 2207 conf:(0.75) 9. bread and cake=t 3330 ==> milk-cream=t 2337 conf:(0.7) 10. bread and cake=t 3330 ==> fruit=t 2325 conf:(0.7)
0.1	0.2	11	2	1. biscuits=t 2605 ==> bread and cake=t 2083 conf:(0.8) 2. milk-cream=t 2939 ==> bread and cake=t 2337 conf:(0.8) 3. fruit=t 2962 ==> bread and cake=t 2325 conf:(0.78) 4. baking needs=t 2795 ==> bread and cake=t 2191 conf:(0.78) 5. frozen foods=t 2717 ==> bread and cake=t 2129 conf:(0.78) 6. vegetables=t 2961 ==> bread and cake=t 2298 conf:(0.78) 7. vegetables=t 2961 ==> fruit=t 2207 conf:(0.75) 8. fruit=t 2962 ==> vegetables=t 2207 conf:(0.75) 9. bread and cake=t 3330 ==> milk-cream=t 2337 conf:(0.7) 10. bread and cake=t 3330 ==> fruit=t 2325 conf:(0.7)
Section 2: Medium (lower minimum support) with high, medium and low (minimum confidence)				
0.3	0.8	14	3	1. biscuits=t vegetables=t 1764 ==> bread and cake=t 1487 conf:(0.84) 2. total=high 1679 ==> bread and cake=t 1413 conf:(0.84) 3. biscuits=t milk-cream=t 1767 ==> bread and cake=t 1485 conf:(0.84) 4. biscuits=t fruit=t 1837 ==> bread and cake=t 1541 conf:(0.84) 5. biscuits=t frozen foods=t 1810 ==> bread and cake=t 1510 conf:(0.83) 6. frozen foods=t fruit=t 1861 ==> bread and cake=t 1548 conf:(0.83) 7. frozen foods=t milk-cream=t 1826 ==> bread and cake=t 1516 conf:(0.83) 8. baking needs=t milk-cream=t 1907 ==> bread and cake=t 1580 conf:(0.83) 9. milk-cream=t fruit=t 2038 ==> bread and cake=t 1684 conf:(0.83) 10. baking needs=t biscuits=t 1764 ==> bread and cake=t 1456 conf:(0.83)
0.3	0.5	11	2	1. biscuits=t 2605 ==> bread and cake=t 2083 conf:(0.8) 2. milk-cream=t 2939 ==> bread and cake=t 2337 conf:(0.8) 3. fruit=t 2962 ==> bread and cake=t 2325 conf:(0.78) 4. baking needs=t 2795 ==> bread and cake=t 2191 conf:(0.78) 5. frozen foods=t 2717 ==> bread and cake=t 2129 conf:(0.78) 6. vegetables=t 2961 ==> bread and cake=t 2298 conf:(0.78) 7. vegetables=t 2961 ==> fruit=t 2207 conf:(0.75) 8. fruit=t 2962 ==> vegetables=t 2207 conf:(0.75) 9. bread and cake=t 3330 ==> milk-cream=t 2337 conf:(0.7) 10. bread and cake=t 3330 ==> fruit=t 2325 conf:(0.7)

0.3	0.2	11	2	1. biscuits=t 2605 ==> bread and cake=t 2083 conf:(0.8) 2. milk-cream=t 2939 ==> bread and cake=t 2337 conf:(0.8) 3. fruit=t 2962 ==> bread and cake=t 2325 conf:(0.78) 4. baking needs=t 2795 ==> bread and cake=t 2191 conf:(0.78) 5. frozen foods=t 2717 ==> bread and cake=t 2129 conf:(0.78) 6. vegetables=t 2961 ==> bread and cake=t 2298 conf:(0.78) 7. vegetables=t 2961 ==> fruit=t 2207 conf:(0.75) 8. fruit=t 2962 ==> vegetables=t 2207 conf:(0.75) 9. bread and cake=t 3330 ==> milk-cream=t 2337 conf:(0.7) 10. bread and cake=t 3330 ==> fruit=t 2325 conf:(0.7)
Section 3: High (lower minimum support) with high, medium and low (minimum confidence)				
0.5	0.8	10	2	No best rules were achieved
0.5	0.5	10	2	Only 4 best rules were achieved 1. milk-cream=t 2939 ==> bread and cake=t 2337 conf:(0.8) 2. fruit=t 2962 ==> bread and cake=t 2325 conf:(0.78) 3. bread and cake=t 3330 ==> milk-cream=t 2337 conf:(0.7) 4. bread and cake=t 3330 ==> fruit=t 2325 conf:(0.7)
0.5	0.2	10	2	Only 4 best rules were achieved 1. milk-cream=t 2939 ==> bread and cake=t 2337 conf:(0.8) 2. fruit=t 2962 ==> bread and cake=t 2325 conf:(0.78) 3. bread and cake=t 3330 ==> milk-cream=t 2337 conf:(0.7) 4. bread and cake=t 3330 ==> fruit=t 2325 conf:(0.7)
Section 4: Very high (lower minimum support) with high, medium and low (minimum confidence)				
0.7	0.8	0	0	No large itemsets / best rules were achieved
0.7	0.5	0	0	No large itemsets / best rules were achieved
0.7	0.2	0	0	No large itemsets / best rules were achieved

Inferences on association rule generation by modifying support and confidence values

- Consider the rules generated when lower minimum support (LMS) is low and minimum confidence (MS) is high e.g LMS = 0.1 and MS = 0.8. In these cases, we see a large number of rules generated. Though the confidence value for the generated rules is high, the rules are not interesting and may not be reliable since their support is comparatively low and that they generally map to the most frequent item which is bread and cake.

- In the first row of the table above, all the 10 rules have a support around 1800 which is not that great considering the size of the supermarket dataset. They have simply popped up because their confidence was high against bread and cake and the lower minimum support criteria was very low. In general, a lot of rules can get high confidence this way, but the associations are not always true.

- But as we relax confidence requirements, we begin to see rules with greater support at the top. Eg. LMS=0.1 and MS=0.5 or 0.2. This case is better than the previous one, not only because of the improved support but also because we begin to see some interesting unusual associations. Yet we can still do better.

- Now we increase the LMS and see rules with significantly greater support and acceptable confidence. Rules of this case seem more valuable than the types of rules addressed in the above points. It also has to be noted that if both LMS and MS are very high, rules don't get generated sometimes. This is expected because we cannot expect interesting associations to happen all the time. A certain optimum confidence on these associations would suffice.

- Also very high support values (eg. LMS=0.7) expectedly does not produce any rules at all, simple because there is no such itemset that is massively common among the transactions. Unless there is a clearance sale on any particular item, no item can satisfy this degree of support.

- The values of cycle count and large itemset levels are inversely relative to the LMS values at most cases. If LMS values increase, these values decrease.

(iv) Table 1.2 Removing some most frequent items and repeating the experiments.

Lower Min Support	Min Confi -dence	Cycle count	Large itemset levels	**Top 10 association rules**
Section 1: Removing the most frequent 'Bread and cake' department from original dataset				
0.1	0.9	18	6	Only 4 best rules were achieved 1. baking needs=t beef=t fruit=t total=high 527 ==> vegetables=t 485 conf:(0.92) 2. milk-cream=t beef=t fruit=t total=high 512 ==> vegetables=t 464 conf:(0.91) 3. biscuits=t beef=t fruit=t total=high 514 ==> vegetables=t 465 conf:(0.9) 4. frozen foods=t beef=t fruit=t total=high 543 ==> vegetables=t 491 conf:(0.9)
0.3	0.6	12	2	1. vegetables=t 2961 ==> fruit=t 2207 conf:(0.75) 2. fruit=t 2962 ==> vegetables=t 2207 conf:(0.75) 3. baking needs=t 2795 ==> vegetables=t 1949 conf:(0.7) 4. milk-cream=t 2939 ==> fruit=t 2038 conf:(0.69) 5. frozen foods=t 2717 ==> vegetables=t 1882 conf:(0.69) 6. milk-cream=t 2939 ==> vegetables=t 2025 conf:(0.69) 7. fruit=t 2962 ==> milk-cream=t 2038 conf:(0.69) 8. frozen foods=t 2717 ==> fruit=t 1861 conf:(0.68) 9. vegetables=t 2961 ==> milk-cream=t 2025 conf:(0.68) 10. baking needs=t 2795 ==> milk-cream=t 1907 conf:(0.68)
0.5	0.3	0	0	No large itemsets / best rules were achieved
Section 2: Removing the next most frequent 'baking needs' department from original dataset				

0.1	0.9	17	5	1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723 conf:(0.92) 2. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746 conf:(0.92) 3. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779 conf:(0.91) 4. biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t 725 conf:(0.91) 5. biscuits=t fruit=t total=high 954 ==> bread and cake=t 866 conf:(0.91) 6. frozen foods=t fruit=t vegetables=t total=high 834 ==> bread and cake=t 757 conf:(0.91) 7. frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877 conf:(0.91) 8. biscuits=t milk-cream=t total=high 907 ==> bread and cake=t 820 conf:(0.9) 9. biscuits=t vegetables=t total=high 950 ==> bread and cake=t 858 conf:(0.9) 10. tissues-paper prd=t fruit=t total=high 878 ==> bread and cake=t 792 conf:(0.9)
0.3	0.6	11	2	1. biscuits=t 2605 ==> bread and cake=t 2083 conf:(0.8) 2. milk-cream=t 2939 ==> bread and cake=t 2337 conf:(0.8) 3. fruit=t 2962 ==> bread and cake=t 2325 conf:(0.78) 4. frozen foods=t 2717 ==> bread and cake=t 2129 conf:(0.78) 5. vegetables=t 2961 ==> bread and cake=t 2298 conf:(0.78) 6. vegetables=t 2961 ==> fruit=t 2207 conf:(0.75) 7. fruit=t 2962 ==> vegetables=t 2207 conf:(0.75) 8. bread and cake=t 3330 ==> milk-cream=t 2337 conf:(0.7) 9. bread and cake=t 3330 ==> fruit=t 2325 conf:(0.7) 10. bread and cake=t 3330 ==> vegetables=t 2298 conf:(0.69)
0.5	0.3	10	2	Only 4 best rules were achieved 1. milk-cream=t 2939 ==> bread and cake=t 2337 conf:(0.8) 2. fruit=t 2962 ==> bread and cake=t 2325 conf:(0.78) 3. bread and cake=t 3330 ==> milk-cream=t 2337 conf:(0.7) 4. bread and cake=t 3330 ==> fruit=t 2325 conf:(0.7)
Section 3: Removing the 'Bread and cake', 'Baking needs' and 'Juice-sat-cord-ms' departments from original dataset				
0.1	0.9	18	6	Only 3 best rules were achieved 1. milk-cream=t beef=t fruit=t total=high 512 ==> vegetables=t 464 conf:(0.91) 2. biscuits=t beef=t fruit=t total=high 514 ==> vegetables=t 465 conf:(0.9) 3. frozen foods=t beef=t fruit=t total=high 543 ==> vegetables=t 491 conf:(0.9)
0.3	0.6	12	2	1. vegetables=t 2961 ==> fruit=t 2207 conf:(0.75)

				2. fruit=t 2962 ==> vegetables=t 2207 conf:(0.75) 3. milk-cream=t 2939 ==> fruit=t 2038 conf:(0.69) 4. frozen foods=t 2717 ==> vegetables=t 1882 conf:(0.69) 5. milk-cream=t 2939 ==> vegetables=t 2025 conf:(0.69) 6. fruit=t 2962 ==> milk-cream=t 2038 conf:(0.69) 7. frozen foods=t 2717 ==> fruit=t 1861 conf:(0.68) 8. vegetables=t 2961 ==> milk-cream=t 2025 conf:(0.68) 9. vegetables=t 2961 ==> frozen foods=t 1882 conf:(0.64) 10. fruit=t 2962 ==> frozen foods=t 1861 conf:(0.63)
0.5	0.3	0	0	No large itemsets / best rules were achieved
Section 4: Removing the 'biscuits', 'canned fish-meat', 'canned fruit', 'canned vegetables' and 'breakfast food' departments from original dataset				
0.1	0.9	18	7	1. baking needs=t cheese=t fruit=t vegetables=t total=high 519 ==> bread and cake=t 483 conf:(0.93) 2. frozen foods=t party snack foods=t tissues-paper prd=t fruit=t total=high 518 ==> bread and cake=t 482 conf:(0.93) 3. baking needs=t frozen foods=t party snack foods=t fruit=t total=high 558 ==> bread and cake=t 518 conf:(0.93) 4. baking needs=t cheese=t fruit=t total=high 584 ==> bread and cake=t 542 conf:(0.93) 5. baking needs=t frozen foods=t tissues-paper prd=t fruit=t vegetables=t total=high 513 ==> bread and cake=t 476 conf:(0.93) 6. party snack foods=t cheese=t fruit=t total=high 535 ==> bread and cake=t 496 conf:(0.93) 7. party snack foods=t tissues-paper prd=t fruit=t vegetables=t total=high 530 ==> bread and cake=t 491 conf:(0.93) 8. frozen foods=t party snack foods=t milk-cream=t fruit=t total=high 528 ==> bread and cake=t 489 conf:(0.93) 9. baking needs=t frozen foods=t tissues-paper prd=t fruit=t total=high 581 ==> bread and cake=t 538 conf:(0.93) 10. baking needs=t frozen foods=t margarine=t fruit=t total=high 553 ==> bread and cake=t 512 conf:(0.93)
0.3	0.6	11	2	1. milk-cream=t 2939 ==> bread and cake=t 2337 conf:(0.8) 2. fruit=t 2962 ==> bread and cake=t 2325 conf:(0.78) 3. baking needs=t 2795 ==> bread and cake=t 2191 conf:(0.78) 4. frozen foods=t 2717 ==> bread and cake=t 2129 conf:(0.78) 5. vegetables=t 2961 ==> bread and cake=t 2298 conf:(0.78) 6. vegetables=t 2961 ==> fruit=t 2207 conf:(0.75) 7. fruit=t 2962 ==> vegetables=t 2207 conf:(0.75) 8. bread and cake=t 3330 ==> milk-cream=t 2337 conf:(0.7) 9. bread and cake=t 3330 ==> fruit=t 2325 conf:(0.7) 10. bread and cake=t 3330 ==> vegetables=t 2298 conf:(0.69)
0.5	0.3	10	2	Only 4 best rules were achieved

				1. milk-cream=t 2939 ==> bread and cake=t 2337 conf:(0.8) 2. fruit=t 2962 ==> bread and cake=t 2325 conf:(0.78) 3. bread and cake=t 3330 ==> milk-cream=t 2337 conf:(0.7) 4. bread and cake=t 3330 ==> fruit=t 2325 conf:(0.7)
Section 5: Removing all the above 8 departments from original dataset				
0.1	0.9	18	6	Only 2 best rules were achieved 1. milk-cream=t beef=t fruit=t total=high 512 ==> vegetables=t 464 conf:(0.91) 2. frozen foods=t beef=t fruit=t total=high 543 ==> vegetables=t 491 conf:(0.9)
0.3	0.6	12	2	1. vegetables=t 2961 ==> fruit=t 2207 conf:(0.75) 2. fruit=t 2962 ==> vegetables=t 2207 conf:(0.75) 3. milk-cream=t 2939 ==> fruit=t 2038 conf:(0.69) 4. frozen foods=t 2717 ==> vegetables=t 1882 conf:(0.69) 5. milk-cream=t 2939 ==> vegetables=t 2025 conf:(0.69) 6. fruit=t 2962 ==> milk-cream=t 2038 conf:(0.69) 7. frozen foods=t 2717 ==> fruit=t 1861 conf:(0.68) 8. vegetables=t 2961 ==> milk-cream=t 2025 conf:(0.68) 9. vegetables=t 2961 ==> frozen foods=t 1882 conf:(0.64) 10. fruit=t 2962 ==> frozen foods=t 1861 conf:(0.63)
0.5	0.3	0	0	No large itemsets / best rules were achieved

Inferences on association rule generation by modifying support and confidence values

- Several itemsets tending to associate with most frequent items is predictable. Those rules are neither interesting nor profitable because there are good number of chances that the frequent items will be unaffected if the itemsets that are associated with them are removed.

- Consider the first row in Table 1.1 again. 20% of top rules say that 'baking needs' is associated with bread and butter. But in section 2 of Table 1.2, when 'baking needs' is removed, all items still map to 'bread and cake' when LMS=0.1 and MS=0.9. Thus the frequent itemsets might have a clout on the association rules and if we remove them and run association analysis, we could, if our assumption is right, get interesting rules.

- We just do that to understand the effect of frequent items on association rules. First when the most frequent 'bread and cake' is removed, LMS=0.1 and MS=0.9 case does not perform well. Same phenomenon is observed in section 3. Only 3-4 rules are generated which means that this case depends heavily on the most frequent dataset and hence we can conclude that our inference for Table 1.1 for this LMS, MS combination is correct.

- As the most frequent itemsets are removed and when LMS has an optimum value with acceptable confidence, we see some interesting rules. One such is when LMS=0.3 and MS=0.6 in section 1 and 3, we see a rule which associates frozen foods with vegetables. This is a bit interesting since people buy frozen foods to save the time spent cooking fresh edibles like vegetables. But the rule could be true, since people may tend to buy frozen foods to eat immediately after returning home from a tiresome shopping trip and buy vegetables to cook for later.

(v) Summary of inferences that were developed using the experiments

- To summarize the inferences, we can say that when lower minimum support is too high or too low with high confidence, the association analysis is not that helpful. Former does not generate any rules while the latter generates unreliable, uninteresting rules.

- With decrease in confidence for high LMS, we get some better rules. Also when support is medium with medium confidence, we get even better, interesting rules.

- One such rule is frozen food associated with fresh vegetables and that may actually be true, people could buy frozen foods for a quick dinner when they reach home and load some fresh vegetables for cooking later.

Vote dataset:
The dataset, published by the Congressional Quarterly Almanac, gathers the votes of 435 U.S House of Representatives Congressmen on 16 categories/issues. Class attribute defines whether the representative is a Democrat or Republican. Thus there are 435 records and 17 attributes in total.

(i) Parameters of Apriori algorithm.

The significant parameters are same as that of the previous dataset.

(ii) Top 4 association rules using default parameters.

Main parameters with default value: ***Lower minimum support: 0.1; minimum confidence: 0.9***

1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1)
2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> Class=democrat 198 conf:(1)
3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210 conf:(1)
4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 conf:(1)

Where number of cycles were 11 and count of large itemset levels generated was 4.

(iii) Documenting and analysing the association rule results for modified support and count values.

Lower Min Support	Min Confi -dence	Cycle count	Large itemset levels	**Top 10 association rules**
Low (lower minimum support) with high, medium and low (minimum confidence)				
0.1	0.9	11	4	1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1) 2. adoption-of-the-budget-resolution=y physician-fee-freeze=n

				aid-to-nicaraguan-contras=y 198 ==> Class=democrat 198 conf:(1) 3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210 conf:(1) 4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 conf:(1) 5. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99) 6. el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197 conf:(0.99) 7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204 conf:(0.98) 8. adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y Class=democrat 203 ==> physician-fee-freeze=n 198 conf:(0.98) 9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197 conf:(0.97) 10. aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210 conf:(0.96)
0.1	0.5	10	3	1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1) 2. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99) 3. adoption-of-the-budget-resolution=y Class=democrat 231 ==> physician-fee-freeze=n 219 conf:(0.95) 4. Class=democrat 267 ==> physician-fee-freeze=n 245 conf:(0.92) 5. adoption-of-the-budget-resolution=y 253 ==> Class=democrat 231 conf:(0.91) 6. aid-to-nicaraguan-contras=y 242 ==> Class=democrat 218 conf:(0.9) 7. physician-fee-freeze=n Class=democrat 245 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 8. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 9. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y Class=democrat 219 conf:(0.89) 10. adoption-of-the-budget-resolution=y 253 ==> physician-fee-freeze=n 219 conf:(0.87)
0.1	0.2	10	3	1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1) 2. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99) 3. adoption-of-the-budget-resolution=y Class=democrat 231 ==> physician-fee-freeze=n 219 conf:(0.95)

				4. Class=democrat 267 ==> physician-fee-freeze=n 245 conf:(0.92) 5. adoption-of-the-budget-resolution=y 253 ==> Class=democrat 231 conf:(0.91) 6. aid-to-nicaraguan-contras=y 242 ==> Class=democrat 218 conf:(0.9) 7. physician-fee-freeze=n Class=democrat 245 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 8. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 9. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y Class=democrat 219 conf:(0.89) 10. adoption-of-the-budget-resolution=y 253 ==> physician-fee-freeze=n 219 conf:(0.87)
Medium (lower minimum support) with high, medium and low (minimum confidence)				
0.3	0.9	11	4	1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1) 2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> Class=democrat 198 conf:(1) 3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210 conf:(1) 4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 conf:(1) 5. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99) 6. el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197 conf:(0.99) 7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204 conf:(0.98) 8. adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y Class=democrat 203 ==> physician-fee-freeze=n 198 conf:(0.98) 9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197 conf:(0.97) 10. aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210 conf:(0.96)
0.3	0.5	10	3	1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1) 2. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99) 3. adoption-of-the-budget-resolution=y Class=democrat 231 ==> physician-fee-freeze=n 219 conf:(0.95) 4. Class=democrat 267 ==> physician-fee-freeze=n 245 conf:(0.92)

				5. adoption-of-the-budget-resolution=y 253 ==> Class=democrat 231 conf:(0.91) 6. aid-to-nicaraguan-contras=y 242 ==> Class=democrat 218 conf:(0.9) 7. physician-fee-freeze=n Class=democrat 245 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 8. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 9. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y Class=democrat 219 conf:(0.89) 10. adoption-of-the-budget-resolution=y 253 ==> physician-fee-freeze=n 219 conf:(0.87)
0.3	0.2	10	3	1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1) 2. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99) 3. adoption-of-the-budget-resolution=y Class=democrat 231 ==> physician-fee-freeze=n 219 conf:(0.95) 4. Class=democrat 267 ==> physician-fee-freeze=n 245 conf:(0.92) 5. adoption-of-the-budget-resolution=y 253 ==> Class=democrat 231 conf:(0.91) 6. aid-to-nicaraguan-contras=y 242 ==> Class=democrat 218 conf:(0.9) 7. physician-fee-freeze=n Class=democrat 245 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 8. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 9. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y Class=democrat 219 conf:(0.89) 10. adoption-of-the-budget-resolution=y 253 ==> physician-fee-freeze=n 219 conf:(0.87)
High (lower minimum support) with high, medium and low (minimum confidence)				
0.5	0.9	10	3	Only 6 best rules were achieved 1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1) 2. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99) 3. adoption-of-the-budget-resolution=y Class=democrat 231 ==> physician-fee-freeze=n 219 conf:(0.95) 4. Class=democrat 267 ==> physician-fee-freeze=n 245 conf:(0.92) 5. adoption-of-the-budget-resolution=y 253 ==> Class=democrat 231 conf:(0.91) 6. aid-to-nicaraguan-contras=y 242 ==> Class=democrat 218

				conf:(0.9)
0.5	0.5	10	3	1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1) 2. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99) 3. adoption-of-the-budget-resolution=y Class=democrat 231 ==> physician-fee-freeze=n 219 conf:(0.95) 4. Class=democrat 267 ==> physician-fee-freeze=n 245 conf:(0.92) 5. adoption-of-the-budget-resolution=y 253 ==> Class=democrat 231 conf:(0.91) 6. aid-to-nicaraguan-contras=y 242 ==> Class=democrat 218 conf:(0.9) 7. physician-fee-freeze=n Class=democrat 245 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 8. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 9. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y Class=democrat 219 conf:(0.89) 10. adoption-of-the-budget-resolution=y 253 ==> physician-fee-freeze=n 219 conf:(0.87)
0.5	0.2	10	3	1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1) 2. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99) 3. adoption-of-the-budget-resolution=y Class=democrat 231 ==> physician-fee-freeze=n 219 conf:(0.95) 4. Class=democrat 267 ==> physician-fee-freeze=n 245 conf:(0.92) 5. adoption-of-the-budget-resolution=y 253 ==> Class=democrat 231 conf:(0.91) 6. aid-to-nicaraguan-contras=y 242 ==> Class=democrat 218 conf:(0.9) 7. physician-fee-freeze=n Class=democrat 245 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 8. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 9. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y Class=democrat 219 conf:(0.89) 10. adoption-of-the-budget-resolution=y 253 ==> physician-fee-freeze=n 219 conf:(0.87)
Very high (lower minimum support) with high, medium and low (minimum confidence)				
0.7	0.9	0	0	No large itemsets / best rules were achieved
0.7	0.5	0	0	No large itemsets / best rules were achieved

0.7	0.2	0	0	No large itemsets / best rules were achieved

Inferences on association rule generation by modifying support and confidence values

- The inferences that were made for the previous dataset, holds good for vote dataset to a certain extent. Very low LMS and high confidence yields uninteresting results.

- As we increase confidence for very high LMS, we see some interesting results, and even more when we increase LMS but keep confidence between medium and high. Very high LMS does not give any results as it did for the previous dataset.

- One key thing to notice in the results is that we see the class democrats often with which we can have an opinion that the democrats are predictable and they seem to think in unison. Also the repetition of some key issues over and over again like physician fee freeze and adoption of budget resolution indicate that their propaganda will be centred over those issues.

- The other frequent rules linking aid to nigaraguan contras and El salvador denote the close association between the two issues which is true in real world.

(iv) Removing some most frequent items and repeating the experiments.

Lower Min Support	Min Confi -dence	Cycle count	Large itemset levels	**Top 10 association rules**
Removing the nearly uniformly distributed 'water-project-cost-sharing' from original dataset				
0.1	0.9	11	4	The other frequent rules linking aid to nigaraguan contras and el salvador denote the close association between the two issues which is true in real world. 1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1) 2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> Class=democrat 198 conf:(1) 3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210 conf:(1) 4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 conf:(1) 5. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99) 6. el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197 conf:(0.99) 7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204 conf:(0.98) 8. adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y Class=democrat 203 ==> physician-fee-freeze=n 198

				conf:(0.98) 9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197 conf:(0.97) 10. aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210 conf:(0.96)
0.3	0.6	10	3	1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1) 2. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99) 3. adoption-of-the-budget-resolution=y Class=democrat 231 ==> physician-fee-freeze=n 219 conf:(0.95) 4. Class=democrat 267 ==> physician-fee-freeze=n 245 conf:(0.92) 5. adoption-of-the-budget-resolution=y 253 ==> Class=democrat 231 conf:(0.91) 6. aid-to-nicaraguan-contras=y 242 ==> Class=democrat 218 conf:(0.9) 7. physician-fee-freeze=n Class=democrat 245 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 8. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 9. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y Class=democrat 219 conf:(0.89) 10. adoption-of-the-budget-resolution=y 253 ==> physician-fee-freeze=n 219 conf:(0.87)
0.5	0.3	10	3	1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1) 2. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99) 3. adoption-of-the-budget-resolution=y Class=democrat 231 ==> physician-fee-freeze=n 219 conf:(0.95) 4. Class=democrat 267 ==> physician-fee-freeze=n 245 conf:(0.92) 5. adoption-of-the-budget-resolution=y 253 ==> Class=democrat 231 conf:(0.91) 6. aid-to-nicaraguan-contras=y 242 ==> Class=democrat 218 conf:(0.9) 7. physician-fee-freeze=n Class=democrat 245 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 8. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 9. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y Class=democrat 219 conf:(0.89) 10. adoption-of-the-budget-resolution=y 253 ==> physician-fee-freeze=n 219 conf:(0.87)

Removing the nearly uniformly distributed 'immigration' from original dataset				
0.1	0.9	11	4	1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1) 2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> Class=democrat 198 conf:(1) 3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210 conf:(1) 4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 conf:(1) 5. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99) 6. el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197 conf:(0.99) 7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204 conf:(0.98) 8. adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y Class=democrat 203 ==> physician-fee-freeze=n 198 conf:(0.98) 9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197 conf:(0.97) 10. aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210 conf:(0.96)
0.3	0.6	10	3	1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1) 2. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99) 3. adoption-of-the-budget-resolution=y Class=democrat 231 ==> physician-fee-freeze=n 219 conf:(0.95) 4. Class=democrat 267 ==> physician-fee-freeze=n 245 conf:(0.92) 5. adoption-of-the-budget-resolution=y 253 ==> Class=democrat 231 conf:(0.91) 6. aid-to-nicaraguan-contras=y 242 ==> Class=democrat 218 conf:(0.9) 7. physician-fee-freeze=n Class=democrat 245 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 8. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 9. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y Class=democrat 219 conf:(0.89) 10. adoption-of-the-budget-resolution=y 253 ==> physician-fee-freeze=n 219 conf:(0.87)
0.5	0.3	10	3	1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1)

				2. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99) 3. adoption-of-the-budget-resolution=y Class=democrat 231 ==> physician-fee-freeze=n 219 conf:(0.95) 4. Class=democrat 267 ==> physician-fee-freeze=n 245 conf:(0.92) 5. adoption-of-the-budget-resolution=y 253 ==> Class=democrat 231 conf:(0.91) 6. aid-to-nicaraguan-contras=y 242 ==> Class=democrat 218 conf:(0.9) 7. physician-fee-freeze=n Class=democrat 245 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 8. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y 219 conf:(0.89) 9. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y Class=democrat 219 conf:(0.89) 10. adoption-of-the-budget-resolution=y 253 ==> physician-fee-freeze=n 219 conf:(0.87)
Removing some key democrat supports from original dataset				
0.1	0.9	13	3	1. crime=n 170 ==> Class=democrat 167 conf:(0.98) 2. mx-missile=n duty-free-exports=n 162 ==> religious-groups-in-schools=y 153 conf:(0.94) 3. mx-missile=n crime=y 176 ==> religious-groups-in-schools=y 166 conf:(0.94) 4. Class=republican 168 ==> crime=y 158 conf:(0.94) 5. superfund-right-to-sue=y crime=y 177 ==> religious-groups-in-schools=y 166 conf:(0.94) 6. superfund-right-to-sue=y duty-free-exports=n 165 ==> religious-groups-in-schools=y 154 conf:(0.93) 7. anti-satellite-test-ban=n 182 ==> religious-groups-in-schools=y 169 conf:(0.93) 8. anti-satellite-test-ban=y superfund-right-to-sue=n 170 ==> Class=democrat 157 conf:(0.92) 9. superfund-right-to-sue=y duty-free-exports=n 165 ==> crime=y 152 conf:(0.92) 10. duty-free-exports=y 174 ==> Class=democrat 160 conf:(0.92)
0.3	0.6	12	2	1. mx-missile=n 206 ==> religious-groups-in-schools=y 188 conf:(0.91) 2. mx-missile=y 207 ==> Class=democrat 188 conf:(0.91) 3. superfund-right-to-sue=n 201 ==> Class=democrat 179 conf:(0.89) 4. superfund-right-to-sue=y 209 ==> religious-groups-in-schools=y 186 conf:(0.89) 5. mx-missile=y 207 ==> anti-satellite-test-ban=y 182 conf:

				(0.88) 6. crime=y 248 ==> religious-groups-in-schools=y 214 conf:(0.86) 7. mx-missile=n 206 ==> crime=y 176 conf:(0.85) 8. superfund-right-to-sue=y 209 ==> crime=y 177 conf:(0.85) 9. anti-satellite-test-ban=y 239 ==> Class=democrat 200 conf:(0.84) 10. duty-free-exports=n 233 ==> religious-groups-in-schools=y 193 conf:(0.83)
0.5	0.3	0	0	No large itemsets / rules were achieved
Removing most of the key democrat supports from the original dataset				
0.1	0.9	14	3	1. crime=n 170 ==> Class=democrat 167 conf:(0.98) 2. duty-free-exports=n Class=republican 142 ==> crime=y 136 conf:(0.96) 3. synfuels-corporation-cutback=n Class=republican 138 ==> crime=y 132 conf:(0.96) 4. Class=republican 168 ==> crime=y 158 conf:(0.94) 5. religious-groups-in-schools=y Class=republican 149 ==> crime=y 140 conf:(0.94) 6. handicapped-infants=n duty-free-exports=n 146 ==> religious-groups-in-schools=y 137 conf:(0.94) 7. duty-free-exports=y 174 ==> Class=democrat 160 conf:(0.92) 8. duty-free-exports=n Class=republican 142 ==> religious-groups-in-schools=y 130 conf:(0.92) 9. handicapped-infants=n duty-free-exports=n 146 ==> crime=y 133 conf:(0.91) 10. crime=y duty-free-exports=n 188 ==> religious-groups-in-schools=y 170 conf:(0.9)
0.3	0.6	13	3	1. crime=n 170 ==> Class=democrat 167 conf:(0.98) 2. Class=republican 168 ==> crime=y 158 conf:(0.94) 3. duty-free-exports=y 174 ==> Class=democrat 160 conf:(0.92) 4. crime=y duty-free-exports=n 188 ==> religious-groups-in-schools=y 170 conf:(0.9) 5. handicapped-infants=n crime=y 173 ==> religious-groups-in-schools=y 155 conf:(0.9) 6. religious-groups-in-schools=y duty-free-exports=n 193 ==> crime=y 170 conf:(0.88) 7. crime=y 248 ==> religious-groups-in-schools=y 214 conf:(0.86) 8. handicapped-infants=y 187 ==> Class=democrat 156 conf:(0.83) 9. duty-free-exports=n 233 ==> religious-groups-in-schools=y 193 conf:(0.83) 10. handicapped-infants=n religious-groups-in-schools=y 190 ==>

				crime=y 155 conf:(0.82)
0.5	0.3	0	0	No large itemsets / rules were achieved

Inferences on association rule generation by removing some attributes:

- Unlike the previous dataset, here the removal of some attributes gives information on the degree of support of democrats for other issues. When their main focus issues like physician fee freeze, aid to El Salvador and nicaragua are removed, other issues they support are identified like superfund right to use, mx-missiles and in some cases, the handicapped infants.

- When all the concerns of democrats are removed, we begin to see republicans and their concerns appearing in the rules. Some of the main concerns of republicans seem to be crime, synfuels-corporation-cutback and so on.

- Also it is to be noted that the removal of 'water-cost-project-sharing' and 'immigration' has minimal effect on the rules leading to the conclusion that those issues are do not have a high regard among both parties.

- Also the strong presence of democrats even after removing those not so important issues again indicate the predictive behaviour of democrats and that the republicans do not have that much unity towards the issues they support.

(v) Summary of inferences that were developed using the experiments

- The performance of Apriori algorithm when the support and confidence values change was comparable to that of the previous dataset.

- General analysis of rules indicate the main concerns of each parties. Removal of those concerns show the additional concerns that are important to the party and also the concerns of the other party.

- Thus attribute removal method is useful in knowing which of the concerns are most supported by a party and what are all the concerns that the party targets as a whole. The mindsets of the politicians and their goals can be read using the association rules.

(vi) Comparing association rules to SimpleCart decision tree

Both Simple Cart decision tree algorithm and association analysis work in a similar fashion where they predict whether a representative will be a democrat or republican based on the vote patterns. In case of association analysis, it seems a bit more reliable since it takes into usually takes into factor a number of issues, usually two, to predict whether the candidate will be a republican or democrat. Simple Cart branches on a single attribute or issue and hence it may give comparatively unreliable results. While Simple Cart is not guaranteed to perform less than association analysis in all cases, by the outlook, it seems like association analysis would be a better way for prediction.

One more advantage to association analysis is that it does not necessarily predict the class alone, where Simple cart can predict only the class. Association analysis can tell us that if a person votes for aid to nicaragua, he will most likely support El Salvador aid, no matter whether he is a democrat or republican. Thus we can know the relation between several data attributes also in association analysis.