Assignment – 4

2) Considering the given $\{d(x_i,x_j)\}$ as Euclidian distances between pairs of n distinct objects as the distance proximity matrix, to find the pair of clusters out of ${}^{m}C_2$ we generally could have used the centroid to calculate the distance and find the minimum distance.

Since in this case the centroid is not given, we could follow the given algorithm below:

a) From the given group of m clusters, should select 2 test clusters to perform the algorithm. Now let us consider them to be m_a and m_b
b) From the set of all the points in the 2 test clusters, we can select 2 points, one from each cluster. Let us consider the 2 test points to be x and y from m_a and m_b respectively.
c) Now we should calculate the average distance between every possible object pair (x,y) in both the clusters.
d) Now using the distance formula $D(m_a, m_b)$ = Min (distance between (x,y))
e) For the required pair we can merge the clusters for the which the distance function $D(m_a, m_b)$ defined above is minimum. i.e when the distance between the (x,y) is minimum.

   As an alternate we could also use inter cluster similarity to find the required cluster. For the 2 test clusters and points defined above in the algorithm, for every combination of objects in the 2 clusters we could find the sum and divide the sum by total number of combinations of objects which will result in the inter cluster similarity. Again on repeating this algorithm for all the pairs of objects select the one with maximum inter cluster similarity. [source : few information taken from class slides]

3.

a) Implementation of F_1 and F_{k-1} & F_{k-1} and F_{k-1} :

Car Evaluation data set. Support Count: 17

Implementation Type	F_1 and F_{k-1}	F_{k-1} and F_{k-1} :
Candidate Item sets	22549	15194
Frequent Item sets	2318	2318

b) Implementation of the algorithm on the data set:

Data set1(Car Evaluation):

Minimum Support Threshold: 25

Implementation Type	F_1 and F_{k-1}	F_{k-1} and F_{k-1} :

Candidate Item sets	18718	11802
Frequent Item sets	1893	1893

Minimum Support Threshold: 50

Implementation Type	F_1 and F_{k-1}	F_{k-1} and F_{k-1} :
Candidate Item sets	4017	2117
Frequent Item sets	511	511

Minimum Support Threshold: 75

Implementation Type	F_1 and F_{k-1}	F_{k-1} and F_{k-1} :
Candidate Item sets	1911	690
Frequent Item sets	377	377

Data set2(Mushroom):

Note: Using high support count for this data set because in my way of pre processing it has been modified to 33 columns with the original 8900 rows. So please try to implement the same for this dataset.

Minimum Support Threshold: 2500

Implementation Type	F_1 and F_{k-1}	F_{k-1} and F_{k-1} :
Candidate Item sets	34989	11663
Frequent Item sets	2365	2365

Minimum Support Threshold: 3000

Implementation Type	F_1 and F_{k-1}	F_{k-1} and F_{k-1} :
Candidate Item sets	11297	3765
Frequent Item sets	931	931

Minimum Support Threshold: 3500

Implementation Type	F_1 and F_{k-1}	F_{k-1} and F_{k-1} :
Candidate Item sets	3429	1143
Frequent Item sets	369	369

Data set3 (Nursery):

Minimum Support Threshold: 700

Implementation Type	F_1 and F_{k-1}	F_{k-1} and F_{k-1} :
Candidate Item sets	4998	1666
Frequent Item sets	593	593

Minimum Support Threshold: 800

Implementation Type	F_1 and F_{k-1}	F_{k-1} and F_{k-1} :
Candidate Item sets	1687	562
Frequent Item sets	432	432

Minimum Support Threshold: 900

Implementation Type	F_1 and F_{k-1}	F_{k-1} and F_{k-1} :
Candidate Item sets	1229	409
Frequent Item sets	326	326

As we clearly see, there is reduction in the number of candidate item sets formed in the method of F(k-1) * F(k-1) rule generation. Even though the method of F1 * F(k-1) satisfies the completeness property of candidate generation, it generates a lot of unnecessary candidate item sets. Since both the method produces the same number of frequent item sets for equal minimum support threshold, the only difference is the number of candidate item sets.

c)Enumerate the number of closed and maximal frequent item sets

Data set1(Car Evaluation):

Minimum Support Threshold: 25

Type	Count
Maximal Frequent item set	478
Closed Frequent Item set	817
Frequent item set	1893

Minimum Support Threshold: 50

Type	Count
Maximal Frequent item set	123
Closed Frequent Item set	313

Frequent item set	511

Minimum Support Threshold: 75

Type	Count
Maximal Frequent item set	133
Closed Frequent Item set	215
Frequent item set	377

Data set2(Mushroom):

Note: Using high support count for this data set because in my way of pre processing it has been modified to 33 columns with the original 8900 rows. So please try to implement the same for this dataset.

Minimum Support Threshold: 2500

Type	Count
Maximal Frequent item set	16
Closed Frequent Item set	328
Frequent item set	2365

Minimum Support Threshold: 3000

Type	Count
Maximal Frequent item set	12
Closed Frequent Item set	196
Frequent item set	931

Minimum Support Threshold: 3500

Type	Count
Maximal Frequent item set	11
Closed Frequent Item set	94
Frequent item set	369

Data set3(Nursery):

Minimum Support Threshold: 700

Type	Count
Maximal Frequent item set	68
Closed Frequent Item set	258
Frequent item set	593

Minimum Support Threshold: 800

Type	Count
Maximal Frequent item set	55
Closed Frequent Item set	223
Frequent item set	432

Minimum Support Threshold: 900

Type	Count
Maximal Frequent item set	48
Closed Frequent Item set	167
Frequent item set	326

On comparing the number of closed and maximal frequent item sets with the number of frequent item sets, the maximal frequent item sets are the least. This is because all maximal frequent items are closed frequent item sets, for a frequent item set to be declared maximal none of its supersets has to be frequent item sets and hence they cannot have the same support count as their subsets. It is highly visible that both maximal and frequent are derived from frequent item sets, they are contained in the frequent item sets.

d) Association Rules:

Dataset1 (Car Evaluation): Support : 80

Confidence	Brute Force	Confidence Pruning
40	132	61
50	116	47
60	94	36

Dataset1 (Car Evaluation): Support : 90

Confidence	Brute Force	Confidence Pruning
40	112	57
50	102	44
60	90	34

Dataset1 (Car Evaluation): Support : 100

Confidence	Brute Force	Confidence Pruning
40	93	55
50	86	42
60	74	36

Dataset2 (Mushroom): Support : 3000

Confidence	Brute Force	Confidence Pruning
40	12066	6155
50	9943	5301
60	8600	3861

Dataset2 (Mushroom): Support : 3500

Confidence	Brute Force	Confidence Pruning
40	3466	1457
50	3024	1321
60	2172	1016

Dataset2 (Mushroom): Support : 4000

Confidence	Brute Force	Confidence Pruning
40	1181	456
50	1174	432
60	879	327

Dataset3 (Nursery): Support : 500

Confidence	Brute Force	Confidence Pruning
40	389	96
50	155	91
60	109	89

Dataset3 (Nursery): Support : 600

Confidence	Brute Force	Confidence Pruning
40	229	63
50	99	58
60	71	57

Dataset3(Nursery): Support : 700

Confidence	Brute Force	Confidence Pruning
40	196	57
50	73	47
60	61	46

The number of rules generated is shown above.

e) Rule Generation:

Data Set1(Car Evaluation):

Support : 25 , Minimum Confidence Level: 60

The top 10 rules are:

[['buying_high', 'doors_3', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_med', 'lug_boot_big', 'persons_2'], ['class_value_unacc'], 100.0]
[['doors_3', 'maint_med', 'persons_2'], ['class_value_unacc'], 100.0]
[['doors_4', 'maint_low', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_vhigh', 'lug_boot_big', 'maint_high'], ['class_value_unacc'], 100.0]
[['buying_med', 'persons_2'], ['class_value_unacc'], 100.0]
[['buying_med', 'lug_boot_med', 'persons_2'], ['class_value_unacc'], 100.0]
[['lug_boot_med', 'maint_high', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_vhigh', 'lug_boot_big', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_vhigh', 'persons_2', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_low', 'persons_2', 'safety_high'], ['class_value_unacc'], 100.0]

Support : 25 , Minimum Confidence Level: 75

The top 10 rules are:

[['buying_vhigh', 'doors_4', 'persons_2'], ['class_value_unacc'], 100.0]
[['lug_boot_small', 'persons_2', 'safety_med'], ['class_value_unacc'], 100.0]
[['buying_low', 'doors_4', 'safety_low'], ['class_value_unacc'], 100.0]
[['lug_boot_big', 'maint_low', 'persons_2'], ['class_value_unacc'], 100.0]
[['buying_high', 'maint_vhigh', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_vhigh', 'maint_vhigh', 'persons_2'], ['class_value_unacc'], 100.0]

[['buying_low', 'doors_2', 'safety_low'], ['class_value_unacc'], 100.0]
[['doors_3', 'persons_4', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_vhigh', 'doors_3', 'persons_2'], ['class_value_unacc'], 100.0]
[['maint_med', 'persons_2', 'safety_med'], ['class_value_unacc'], 100.0]
[['buying_med', 'maint_med', 'persons_2'], ['class_value_unacc'], 100.0]

Support : 25 , Minimum Confidence Level: 90

The top 10 rules are:
[['doors_3', 'lug_boot_big', 'persons_2'], ['class_value_unacc'], 100.0]
[['buying_med', 'lug_boot_small', 'safety_low'], ['class_value_unacc'], 100.0]
[['maint_high', 'persons_2', 'safety_high'], ['class_value_unacc'], 100.0]
[['lug_boot_med', 'maint_high', 'persons_2'], ['class_value_unacc'], 100.0]
[['lug_boot_big', 'maint_med', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_vhigh', 'lug_boot_big', 'maint_high'], ['class_value_unacc'], 100.0]
[['buying_vhigh', 'lug_boot_med', 'maint_vhigh'], ['class_value_unacc'], 100.0]
[['buying_vhigh', 'maint_med', 'persons_2'], ['class_value_unacc'], 100.0]
[['buying_low', 'lug_boot_small', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_high', 'maint_vhigh', 'persons_4'], ['class_value_unacc'], 100.0]
[['buying_high', 'lug_boot_small', 'persons_2'], ['class_value_unacc'], 100.0]

Support : 50 , Minimum Confidence Level: 60

[['doors_2', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_vhigh', 'maint_vhigh'], ['class_value_unacc'], 100.0]
[['doors_5more', 'persons_2'], ['class_value_unacc'], 100.0]
[['buying_med', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_high', 'persons_2'], ['class_value_unacc'], 100.0]
[['doors_2', 'persons_2'], ['class_value_unacc'], 100.0]
[['doors_3', 'safety_low'], ['class_value_unacc'], 100.0]
[['doors_3', 'persons_2'], ['class_value_unacc'], 100.0]
[['maint_low', 'safety_low'], ['class_value_unacc'], 100.0]
[['maint_low', 'persons_2'], ['class_value_unacc'], 100.0]
[['buying_med', 'persons_2'], ['class_value_unacc'], 100.0]

Support : 50 , Minimum Confidence Level: 75

The top 10 rules are:
[['lug_boot_big', 'persons_4', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_vhigh', 'persons_2'], ['class_value_unacc'], 100.0]
[['persons_2', 'safety_low'], ['class_value_unacc'], 100.0]
[['doors_3', 'persons_2'], ['class_value_unacc'], 100.0]
[['maint_med', 'persons_2'], ['class_value_unacc'], 100.0]

[['lug_boot_small', 'persons_2', 'safety_med'], ['class_value_unacc'], 100.0]
[['lug_boot_med', 'persons_2', 'safety_high'], ['class_value_unacc'], 100.0]
[['maint_vhigh', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_low', 'persons_2'], ['class_value_unacc'], 100.0]
[['buying_high', 'persons_2'], ['class_value_unacc'], 100.0]
[['maint_low', 'safety_low'], ['class_value_unacc'], 100.0]

Support: 50 , Minimum Confidence Level: 90

The top 10 rules are:
[['lug_boot_small', 'persons_2', 'safety_med'], ['class_value_unacc'], 100.0]
[['lug_boot_med', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_vhigh', 'maint_vhigh'], ['class_value_unacc'], 100.0]
[['buying_low', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_high', 'safety_low'], ['class_value_unacc'], 100.0]
[['lug_boot_med', 'persons_2', 'safety_high'], ['class_value_unacc'], 100.0]
[['persons_2', 'safety_high'], ['class_value_unacc'], 100.0]
[['lug_boot_med', 'persons_more', 'safety_low'], ['class_value_unacc'], 100.0]
[['maint_med', 'safety_low'], ['class_value_unacc'], 100.0]
[['lug_boot_med', 'persons_4', 'safety_low'], ['class_value_unacc'], 100.0]
[['lug_boot_small', 'persons_2', 'safety_high'], ['class_value_unacc'], 100.0]

Support: 75 , Minimum Confidence Level: 60

The top 10 rules are:

[['persons_4', 'safety_low'], ['class_value_unacc'], 100.0]
[['maint_vhigh', 'safety_low'], ['class_value_unacc'], 100.0]
[['maint_high', 'persons_2'], ['class_value_unacc'], 100.0]
[['lug_boot_small', 'safety_low'], ['class_value_unacc'], 100.0]
[['persons_more', 'safety_low'], ['class_value_unacc'], 100.0]
[['persons_2', 'safety_high'], ['class_value_unacc'], 100.0]
[['doors_2', 'safety_low'], ['class_value_unacc'], 100.0]
[['lug_boot_med', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_vhigh', 'safety_low'], ['class_value_unacc'], 100.0]
[['maint_high', 'safety_low'], ['class_value_unacc'], 100.0]
[['persons_2', 'safety_med'], ['class_value_unacc'], 100.0]

Support: 75 , Minimum Confidence Level: 75

The top 10 rules are:
[['buying_vhigh', 'maint_high'], ['class_value_unacc'], 100.0]
[['maint_low', 'safety_low'], ['class_value_unacc'], 100.0]
[['persons_4', 'safety_low'], ['class_value_unacc'], 100.0]

[['maint_high', 'safety_low'], ['class_value_unacc'], 100.0]
[['maint_vhigh', 'persons_2'], ['class_value_unacc'], 100.0]
[['maint_high', 'persons_2'], ['class_value_unacc'], 100.0]
[['buying_med', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_med', 'persons_2'], ['class_value_unacc'], 100.0]
[['lug_boot_big', 'safety_low'], ['class_value_unacc'], 100.0]
[['buying_high', 'persons_2'], ['class_value_unacc'], 100.0]
[['persons_2', 'safety_med'], ['class_value_unacc'], 100.0]

Support: 75 , Minimum Confidence Level: 90

The top 10 rules are:
[['doors_2', 'persons_2'], ['class_value_unacc'], 100.0]
[['maint_med', 'persons_2'], ['class_value_unacc'], 100.0]
[['buying_high', 'maint_vhigh'], ['class_value_unacc'], 100.0]
[['lug_boot_small', 'safety_low'], ['class_value_unacc'], 100.0]
[['persons_4', 'safety_low'], ['class_value_unacc'], 100.0]
[['doors_4', 'safety_low'], ['class_value_unacc'], 100.0]
[['doors_3', 'persons_2'], ['class_value_unacc'], 100.0]
[['maint_low', 'safety_low'], ['class_value_unacc'], 100.0]
[['maint_high', 'persons_2'], ['class_value_unacc'], 100.0]
[['doors_4', 'persons_2'], ['class_value_unacc'], 100.0]
[['persons_more', 'safety_low'], ['class_value_unacc'], 100.0]

For the first data set ‘Car evaluation’ which has 1728 transactions, as you can see in the listed rules the top 10 rules changes if we increase the minimum confidence level for the same level of support count. The quality of the rules is not below par I would say. For example consider the rule : [['buying_high', 'maint_vhigh'], ['class_value_unacc'] which means if the market price and maintenance of the car is very high the class of the car is unacceptable. We can judge many samples like that. As for as the peculiarity is concerned even though it is based on cars, we can use the pattern for any vehicles or even machines etc so it is not quite restricted.

Data set2(Mushroom):

Support: 2500, Minimum confidence level : 60

The top 10 rules are:

[['gillcolor_b', 'gillspacing_f', 'ringtype_o', 'sporeprintcolor_p'], ['gillsize_c', 'ringnumber_w', 'veilcolor_p'], 100.0]
[['gillcolor_b', 'ringnumber_w', 'ringtype_o', 'sporeprintcolor_p'], ['gillsize_c', 'gillspacing_f', 'veilcolor_p'], 100.0]

[['gillcolor_b', 'gillsize_c', 'gillspacing_f', 'ringtype_o', 'sporeprintcolor_p'], ['ringnumber_w', 'veilcolor_p'], 100.0]
[['gillcolor_b', 'gillsize_c', 'ringnumber_w', 'ringtype_o', 'sporeprintcolor_p'], ['gillspacing_f', 'veilcolor_p'], 100.0]
[['gillcolor_b', 'gillspacing_f', 'ringnumber_w', 'ringtype_o', 'sporeprintcolor_p'], ['gillsize_c', 'veilcolor_p'], 100.0]
[['gillcolor_b', 'gillspacing_f', 'ringtype_o', 'sporeprintcolor_p', 'veilcolor_p'], ['gillsize_c', 'ringnumber_w'], 100.0]
[['gillcolor_b', 'ringnumber_w', 'ringtype_o', 'sporeprintcolor_p', 'veilcolor_p'], ['gillsize_c', 'gillspacing_f'], 100.0]
[['gillcolor_b', 'gillsize_c', 'gillspacing_f', 'ringnumber_w', 'ringtype_o', 'sporeprintcolor_p'], ['veilcolor_p'], 100.0]
[['gillcolor_b', 'gillsize_c', 'gillspacing_f', 'ringtype_o', 'sporeprintcolor_p', 'veilcolor_p'], ['ringnumber_w'], 100.0]
[['gillcolor_b', 'gillsize_c', 'ringnumber_w', 'ringtype_o', 'sporeprintcolor_p', 'veilcolor_p'], ['gillspacing_f'], 100.0]
[['gillcolor_b', 'gillspacing_f', 'ringnumber_w', 'ringtype_o', 'sporeprintcolor_p', 'veilcolor_p'], ['gillsize_c'], 100.0]

Support: 2500, Minimum confidence level : 75

The top 10 rules are:

[['gillcolor_b', 'ringtype_o', 'stalksurfaceabovering_b'], ['gillsize_c', 'gillspacing_f', 'ringnumber_w', 'veilcolor_p'], 100.0]
[['gillcolor_b', 'gillsize_c', 'ringtype_o', 'stalksurfaceabovering_b'], ['gillspacing_f', 'ringnumber_w', 'veilcolor_p'], 100.0]
[['gillcolor_b', 'gillspacing_f', 'ringtype_o', 'stalksurfaceabovering_b'], ['gillsize_c', 'ringnumber_w', 'veilcolor_p'], 100.0]
[['gillcolor_b', 'ringnumber_w', 'ringtype_o', 'stalksurfaceabovering_b'], ['gillsize_c', 'gillspacing_f', 'veilcolor_p'], 100.0]
[['gillcolor_b', 'ringtype_o', 'stalksurfaceabovering_b', 'veilcolor_p'], ['gillsize_c', 'gillspacing_f', 'ringnumber_w'], 100.0]
[['gillcolor_b', 'gillsize_c', 'gillspacing_f', 'ringtype_o', 'stalksurfaceabovering_b'], ['ringnumber_w', 'veilcolor_p'], 100.0]
[['gillcolor_b', 'gillsize_c', 'ringnumber_w', 'ringtype_o', 'stalksurfaceabovering_b'], ['gillspacing_f', 'veilcolor_p'], 100.0]
[['gillcolor_b', 'gillsize_c', 'ringtype_o', 'stalksurfaceabovering_b', 'veilcolor_p'], ['gillspacing_f', 'ringnumber_w'], 100.0]
[['gillcolor_b', 'gillspacing_f', 'ringnumber_w', 'ringtype_o', 'stalksurfaceabovering_b'], ['gillsize_c', 'veilcolor_p'], 100.0]

[['gillcolor_b', 'gillspacing_f', 'ringtype_o', 'stalksurfaceabovering_b', 'veilcolor_p'], ['gillsize_c', 'ringnumber_w'], 100.0]
[['gillcolor_b', 'ringnumber_w', 'ringtype_o', 'stalksurfaceabovering_b', 'veilcolor_p'], ['gillsize_c', 'gillspacing_f'], 100.0]

Support: 2500, Minimum confidence level : 90

The top 10 rules are:

[['gillsize_c', 'sporeprintcolor_p', 'stalkcolorabovering_s'], ['veilcolor_p'], 100.0]
[['gillcolor_b', 'ringtype_o', 'stalksurfaceabovering_b'], ['gillspacing_f', 'veilcolor_p'], 100.0]
[['gillcolor_b', 'gillspacing_f', 'ringtype_o', 'stalksurfaceabovering_b'], ['veilcolor_p'], 100.0]
[['gillcolor_b', 'ringtype_o', 'stalksurfaceabovering_b', 'veilcolor_p'], ['gillspacing_f'], 100.0]
[['capshape_e', 'gillspacing_f', 'stalksurfacebelowring_s'], ['ringnumber_w'], 100.0]
[['capshape_e', 'ringnumber_w', 'stalksurfacebelowring_s'], ['gillspacing_f'], 100.0]
[['capshape_p', 'habitat_v', 'odor_f'], ['gillspacing_f', 'ringnumber_w', 'ringtype_o'], 100.0]
[['capshape_p', 'gillspacing_f', 'habitat_v', 'odor_f'], ['ringnumber_w', 'ringtype_o'], 100.0]
[['capshape_p', 'habitat_v', 'odor_f', 'ringnumber_w'], ['gillspacing_f', 'ringtype_o'], 100.0]
[['capshape_p', 'habitat_v', 'odor_f', 'ringtype_o'], ['gillspacing_f', 'ringnumber_w'], 100.0]
[['capshape_p', 'gillspacing_f', 'habitat_v', 'odor_f', 'ringnumber_w'], ['ringtype_o'], 100.0]

Support: 3000, Minimum confidence level : 60

The top 10 rules are:
[['gillcolor_b', 'gillspacing_f', 'stalksurfacebelowring_s'], ['veilcolor_p'], 100.0]
[['gillcolor_b', 'stalksurfaceabovering_b'], ['gillsize_c', 'ringnumber_w', 'veilcolor_p'], 100.0]
[['gillcolor_b', 'gillsize_c', 'stalksurfaceabovering_b'], ['ringnumber_w', 'veilcolor_p'], 100.0]
[['gillcolor_b', 'ringnumber_w', 'stalksurfaceabovering_b'], ['gillsize_c', 'veilcolor_p'], 100.0]
[['gillcolor_b', 'stalksurfaceabovering_b', 'veilcolor_p'], ['gillsize_c', 'ringnumber_w'], 100.0]
[['gillcolor_b', 'gillsize_c', 'ringnumber_w', 'stalksurfaceabovering_b'], ['veilcolor_p'], 100.0]
[['gillcolor_b', 'gillsize_c', 'stalksurfaceabovering_b', 'veilcolor_p'], ['ringnumber_w'], 100.0]
[['gillcolor_b', 'ringnumber_w', 'stalksurfaceabovering_b', 'veilcolor_p'], ['gillsize_c'], 100.0]
[['stalkroot_t'], ['ringtype_o', 'veilcolor_p'], 100.0]
[['stalkroot_t', 'veilcolor_p'], ['ringtype_o'], 100.0]
[['class_value_d', 'gillspacing_f', 'ringtype_o'], ['veilcolor_p'], 100.0]

Support: 3000, Minimum confidence level : 75

The top 10 rules are:
[['capcolor_y', 'gillsize_c'], ['ringnumber_w'], 100.0]
[['gillspacing_f', 'sporeprintcolor_p'], ['veilcolor_p'], 100.0]

[['gillsize_c', 'stalkcolorbelowring_w'], ['gillspacing_f', 'ringnumber_w'], 100.0]
[['gillsize_c', 'gillspacing_f', 'stalkcolorbelowring_w'], ['ringnumber_w'], 100.0]
[['gillsize_c', 'ringnumber_w', 'stalkcolorbelowring_w'], ['gillspacing_f'], 100.0]
[['capcolor_y', 'ringnumber_w', 'ringtype_o'], ['veilcolor_p'], 100.0]
[['capsurface_x', 'gillspacing_f', 'ringtype_o'], ['ringnumber_w'], 100.0]
[['capsurface_x', 'ringnumber_w', 'ringtype_o'], ['gillspacing_f'], 100.0]
[['capshape_e', 'gillcolor_b', 'gillspacing_f'], ['ringnumber_w', 'veilcolor_p'], 100.0]
[['capshape_e', 'gillcolor_b', 'ringnumber_w'], ['gillspacing_f', 'veilcolor_p'], 100.0]
[['capshape_e', 'gillcolor_b', 'gillspacing_f', 'ringnumber_w'], ['veilcolor_p'], 100.0]

Support: 3000, Minimum confidence level : 90

The top 10 rules are:
[['gillspacing_f', 'ringtype_o', 'stalksurfaceabovering_b'], ['veilcolor_p'], 100.0]
[['ringtype_o', 'stalksurfaceabovering_b', 'veilcolor_p'], ['gillspacing_f'], 100.0]
[['capshape_p', 'ringnumber_w', 'ringtype_o'], ['veilcolor_p'], 100.0]
[['gillspacing_f', 'sporeprintcolor_p', 'stalksurfacebelowring_s'], ['veilcolor_p'], 100.0]
[['capshape_e', 'gillspacing_f', 'ringtype_o'], ['ringnumber_w'], 100.0]
[['capshape_e', 'ringnumber_w', 'ringtype_o'], ['gillspacing_f'], 100.0]
[['veiltype_w'], ['ringnumber_w', 'veilcolor_p'], 100.0]
[['class_value_d'], ['ringnumber_w', 'veilcolor_p'], 100.0]
[['capshape_e', 'ringnumber_w', 'stalkcolorabovering_s'], ['veilcolor_p'], 100.0]
[['gillsize_c', 'ringtype_o', 'stalksurfaceabovering_b'], ['gillspacing_f', 'ringnumber_w'], 100.0]
[['gillsize_c', 'gillspacing_f', 'ringtype_o', 'stalksurfaceabovering_b'], ['ringnumber_w'], 100.0]
Support: 3500, Minimum confidence level : 60

The top 10 rules are:
[['gillspacing_f', 'ringtype_o', 'veiltype_w'], ['ringnumber_w', 'veilcolor_p'], 100.0]
[['ringnumber_w', 'ringtype_o', 'veiltype_w'], ['gillspacing_f', 'veilcolor_p'], 100.0]
[['ringtype_o', 'veilcolor_p', 'veiltype_w'], ['gillspacing_f', 'ringnumber_w'], 100.0]
[['gillspacing_f', 'ringnumber_w', 'ringtype_o', 'veiltype_w'], ['veilcolor_p'], 100.0]
[['gillspacing_f', 'ringtype_o', 'veilcolor_p', 'veiltype_w'], ['ringnumber_w'], 100.0]
[['ringnumber_w', 'ringtype_o', 'veilcolor_p', 'veiltype_w'], ['gillspacing_f'], 100.0]
[['gillsize_c', 'stalkroot_t'], ['gillspacing_f', 'ringnumber_w', 'ringtype_o'], 100.0]
[['gillsize_c', 'gillspacing_f', 'stalkroot_t'], ['ringnumber_w', 'ringtype_o'], 100.0]
[['gillsize_c', 'ringnumber_w', 'stalkroot_t'], ['gillspacing_f', 'ringtype_o'], 100.0]
[['gillsize_c', 'ringtype_o', 'stalkroot_t'], ['gillspacing_f', 'ringnumber_w'], 100.0]
[['gillsize_c', 'gillspacing_f', 'ringnumber_w', 'stalkroot_t'], ['ringtype_o'], 100.0]

Support: 3500, Minimum confidence level : 75

The top 10 rules are:
[['gillsize_c', 'habitat_v', 'ringtype_o'], ['veilcolor_p'], 100.0]

[['capsurface_x', 'ringnumber_w'], ['veilcolor_p'], 100.0]
[['gillspacing_f', 'stalkcolorabovering_s', 'stalksurfacebelowring_s'], ['ringnumber_w'], 100.0]
[['ringnumber_w', 'stalkcolorabovering_s', 'stalksurfacebelowring_s'], ['gillspacing_f'], 100.0]
[['capshape_e', 'gillspacing_f'], ['veilcolor_p'], 100.0]
[['capshape_p', 'ringnumber_w', 'ringtype_o'], ['gillspacing_f'], 100.0]
[['gillspacing_f', 'ringnumber_w'], ['veilcolor_p'], 100.0]
[['capshape_p', 'gillsize_c', 'gillspacing_f'], ['ringnumber_w'], 100.0]
[['gillsize_c', 'gillspacing_f', 'stalksurfacebelowring_s'], ['ringnumber_w', 'veilcolor_p'], 100.0]
[['gillsize_c', 'ringnumber_w', 'stalksurfacebelowring_s'], ['gillspacing_f', 'veilcolor_p'], 100.0]
[['gillsize_c', 'gillspacing_f', 'ringnumber_w', 'stalksurfacebelowring_s'], ['veilcolor_p'], 100.0]

Support: 3500, Minimum confidence level : 90

The top 10 rules are:
[['capshape_e', 'gillcolor_b', 'gillspacing_f'], ['veilcolor_p'], 100.0]
[['gillspacing_f', 'stalksurfacebelowring_s'], ['veilcolor_p'], 100.0]
[['gillspacing_f', 'ringtype_o', 'stalkcolorbelowring_w'], ['ringnumber_w'], 100.0]
[['ringnumber_w', 'ringtype_o', 'stalkcolorbelowring_w'], ['gillspacing_f'], 100.0]
[['gillspacing_f', 'stalkcolorabovering_s'], ['veilcolor_p'], 100.0]
[['gillcolor_b', 'gillspacing_f', 'ringtype_o'], ['ringnumber_w'], 100.0]
[['gillcolor_b', 'ringnumber_w', 'ringtype_o'], ['gillspacing_f'], 100.0]
[['gillspacing_f', 'ringtype_o', 'veiltype_w'], ['ringnumber_w', 'veilcolor_p'], 100.0]
[['ringnumber_w', 'ringtype_o', 'veiltype_w'], ['gillspacing_f', 'veilcolor_p'], 100.0]
[['ringtype_o', 'veilcolor_p', 'veiltype_w'], ['gillspacing_f', 'ringnumber_w'], 100.0]
[['gillspacing_f', 'ringnumber_w', 'ringtype_o', 'veiltype_w'], ['veilcolor_p'], 100.0]

For the second data set ‘Mushroom which has 8500 rows as in the case of 1st dataset the top 10 rules are changing for the same value of support count. In this the quality of the rules are not classified as in the case of first data set. But these pattern is not general so the peculiarity is high for these rules.

Dataset 3(Nursery):

Support = 800, Minimum confidence level = 60

The top 10 rules are:

[['form_incomplete', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_2', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_2', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['has_nurs_critical', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['has_nurs_less_proper', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['finance_inconv', 'health_not_recom'], ['class_value_not_recom'], 100.0]

[['has_nurs_very_crit', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_more', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_more', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['finance_convenient', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_1', 'class_value_not_recom'], ['health_not_recom'], 100.0]

Support = 800, Minimum confidence level = 75

The top 10 rules are:
[['has_nurs_very_crit', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_2', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_2', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['has_nurs_proper', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_more', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_more', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_foster', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_incomplete', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['finance_convenient', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_completed', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_1', 'class_value_not_recom'], ['health_not_recom'], 100.0]

Support = 800, Minimum confidence level = 90

The top 10 rules are:
[['form_complete', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_completed', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_1', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_1', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['has_nurs_critical', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['has_nurs_improper', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['has_nurs_very_crit', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_incomplete', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_2', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_2', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['finance_convenient', 'health_not_recom'], ['class_value_not_recom'], 100.0]

Support = 900, Minimum confidence level = 60

The top 10 rules are:
[['finance_convenient', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_incomplete', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_complete', 'health_not_recom'], ['class_value_not_recom'], 100.0]

[['form_completed', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['finance_inconv', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_3', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_3', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_1', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_1', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_foster', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_more', 'class_value_not_recom'], ['health_not_recom'], 100.0]

Support = 900, Minimum confidence level = 75

The top 10 rules are:
[['children_more', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_more', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_2', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_2', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_complete', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_foster', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_1', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_1', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['finance_convenient', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_completed', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_3', 'class_value_not_recom'], ['health_not_recom'], 100.0]

Support = 900, Minimum confidence level = 90

[['finance_convenient', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_more', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_more', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_incomplete', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_completed', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_complete', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_foster', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_1', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_1', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_2', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_2', 'health_not_recom'], ['class_value_not_recom'], 100.0]

Support = 1000, Minimum confidence level = 60

The top 10 rules are:
[['children_3', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_3', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['finance_inconv', 'health_not_recom'], ['class_value_not_recom'], 100.0]

[['form_complete', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['finance_convenient', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_incomplete', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_1', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_1', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_more', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_more', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_2', 'class_value_not_recom'], ['health_not_recom'], 100.0]

Support = 1000, Minimum confidence level = 75

The top 10 rules are:
[['children_1', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_1', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_2', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_2', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_complete', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['finance_inconv', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_completed', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['finance_convenient', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_incomplete', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_3', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_3', 'health_not_recom'], ['class_value_not_recom'], 100.0]

Support = 1000, Minimum confidence level = 90

The top 10 rules are:
[['form_foster', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_2', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_2', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_3', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_3', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['children_more', 'class_value_not_recom'], ['health_not_recom'], 100.0]
[['children_more', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['finance_inconv', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_incomplete', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['form_complete', 'health_not_recom'], ['class_value_not_recom'], 100.0]
[['finance_convenient', 'health_not_recom'], ['class_value_not_recom'], 100.0]

For the third data set (nursery) with 12 200 rows the property of top 10 rules changing for the same support count value remains the same. But the quality and peculiarity Is more comparable with the first data set more than the second. It is more general.

3f) Lift measure of interestingness:

Dataset 1(Car evaluation):

Support = 25

The top 10 rules with lift as measure are:
[['buying_low', 'safety_high'], ['class_value_vgood'], 7.2000000000000002]
[['lug_boot_big', 'safety_high'], ['class_value_vgood'], 5.5384615384615392]
[['maint_low', 'safety_med'], ['class_value_good'], 4.5217391304347823]
[['buying_low', 'safety_med'], ['class_value_good'], 4.5217391304347823]
[['maint_med', 'safety_high'], ['class_value_vgood'], 4.7999999999999998]
[['persons_more', 'safety_high'], ['class_value_vgood'], 4.8461538461538458]
[['buying_med', 'safety_high'], ['class_value_vgood'], 4.7999999999999998]
[['persons_4', 'safety_high'], ['class_value_vgood'], 4.1538461538461542]
[['maint_low', 'safety_high'], ['class_value_vgood'], 4.7999999999999998]
[['safety_high'], ['class_value_vgood'], 3.0]
[['class_value_vgood'], ['safety_high'], 3.0]

Support = 50

[['safety_high'], ['class_value_vgood'], 3.0]
[['class_value_vgood'], ['safety_high'], 3.0]
[['class_value_unacc', 'lug_boot_big', 'persons_more'], ['safety_low'], 2.1818181818181821]
[['class_value_unacc', 'lug_boot_big', 'safety_high'], ['persons_2'], 2.1818181818181821]
[['persons_more', 'safety_high'], ['class_value_acc'], 2.25]
[['class_value_unacc', 'lug_boot_med', 'safety_high'], ['persons_2'], 2.1818181818181821]
[['buying_high', 'safety_high'], ['class_value_acc'], 2.15625]
[['persons_more', 'safety_med'], ['class_value_acc'], 2.109375]
[['persons_4', 'safety_high'], ['class_value_acc'], 2.53125]
[['persons_4', 'safety_med'], ['class_value_acc'], 2.109375]
[['class_value_unacc', 'lug_boot_big', 'persons_4'], ['safety_low'], 2.1818181818181821]

Support 75:

[['persons_4', 'safety_med'], ['class_value_acc'], 2.109375]
[['persons_more', 'safety_med'], ['class_value_acc'], 2.109375]
[['persons_4', 'safety_high'], ['class_value_acc'], 2.53125]
[['persons_more', 'safety_high'], ['class_value_acc'], 2.25]
[['maint_high'], ['buying_vhigh', 'class_value_unacc'], 1.0]
[['buying_vhigh', 'class_value_unacc'], ['maint_high'], 1.2]
[['buying_vhigh', 'maint_high'], ['class_value_unacc'], 1.428099173553719]
[['buying_high'], ['persons_4'], 1.0]
[['persons_4'], ['buying_high'], 1.0]

[['buying_vhigh'], ['doors_4'], 1.0]
[['doors_4'], ['buying_vhigh'], 1.0]

Dataset 2(Mushroom):

Support: 2500

[['odor_t', 'ringnumber_w'], ['gillcolor_b', 'stalkcolorabovering_s', 'stalksurfacebelowring_s', 'veilcolor_p'], 2.04826574932251]
[['odor_t', 'stalkcolorabovering_s'], ['gillcolor_b', 'ringnumber_w', 'stalksurfacebelowring_s', 'veilcolor_p'], 2.0498637770897834]
[['odor_t', 'stalksurfacebelowring_s'], ['gillcolor_b', 'ringnumber_w', 'stalkcolorabovering_s', 'veilcolor_p'], 2.0434865805782576]
[['odor_t', 'veilcolor_p'], ['gillcolor_b', 'ringnumber_w', 'stalkcolorabovering_s', 'stalksurfacebelowring_s'], 2.1851971643190007]
[['gillcolor_b', 'ringnumber_w', 'stalkcolorabovering_s'], ['odor_t', 'stalksurfacebelowring_s', 'veilcolor_p'], 2.0434865805782576]
[['gillcolor_b', 'ringnumber_w', 'stalksurfacebelowring_s'], ['odor_t', 'stalkcolorabovering_s', 'veilcolor_p'], 2.0498637770897834]
[['gillcolor_b', 'stalkcolorabovering_s', 'stalksurfacebelowring_s'], ['odor_t', 'ringnumber_w', 'veilcolor_p'], 2.0482657493225105]
[['odor_t', 'ringnumber_w', 'veilcolor_p'], ['gillcolor_b', 'stalkcolorabovering_s', 'stalksurfacebelowring_s'], 2.04826574932251]
[['odor_t', 'stalkcolorabovering_s', 'veilcolor_p'], ['gillcolor_b', 'ringnumber_w', 'stalksurfacebelowring_s'], 2.0498637770897834]
[['odor_t', 'stalksurfacebelowring_s', 'veilcolor_p'], ['gillcolor_b', 'ringnumber_w', 'stalkcolorabovering_s'], 2.0434865805782576]
[['gillcolor_b', 'ringnumber_w', 'stalkcolorabovering_s', 'stalksurfacebelowring_s'], ['odor_t', 'veilcolor_p'], 2.1851971643190007]

Support: 3000

The top 10 rules with lift as measure are:
[['gillspacing_f', 'sporeprintcolor_p'], ['odor_t', 'ringnumber_w', 'ringtype_o', 'veilcolor_p'], 2.1514830508474576]
[['odor_t', 'ringnumber_w'], ['gillspacing_f', 'ringtype_o', 'sporeprintcolor_p', 'veilcolor_p'], 2.2006253447557724]
[['odor_t', 'ringtype_o'], ['gillspacing_f', 'ringnumber_w', 'sporeprintcolor_p', 'veilcolor_p'], 2.1514830508474576]
[['odor_t', 'veilcolor_p'], ['gillspacing_f', 'ringnumber_w', 'ringtype_o', 'sporeprintcolor_p'], 2.2006253447557724]
[['ringnumber_w', 'sporeprintcolor_p'], ['gillspacing_f', 'odor_t', 'ringtype_o', 'veilcolor_p'], 2.1514830508474576]

[['ringtype_o', 'sporeprintcolor_p'], ['gillspacing_f', 'odor_t', 'ringnumber_w', 'veilcolor_p'], 2.0819399329037758]
[['gillspacing_f', 'odor_t', 'ringnumber_w'], ['ringtype_o', 'sporeprintcolor_p', 'veilcolor_p'], 2.0819399329037758]
[['gillspacing_f', 'odor_t', 'ringtype_o'], ['ringnumber_w', 'sporeprintcolor_p', 'veilcolor_p'], 2.1514830508474576]
[['gillspacing_f', 'odor_t', 'veilcolor_p'], ['ringnumber_w', 'ringtype_o', 'sporeprintcolor_p'], 2.2006253447557724]
[['gillspacing_f', 'ringnumber_w', 'sporeprintcolor_p'], ['odor_t', 'ringtype_o', 'veilcolor_p'], 2.1514830508474576]
[['gillspacing_f', 'ringtype_o', 'sporeprintcolor_p'], ['odor_t', 'ringnumber_w', 'veilcolor_p'], 2.2006253447557724]

Support: 3500

The top 10 rules with lift as measure are:
[['gillcolor_b', 'ringnumber_w', 'ringtype_o'], ['veilcolor_p'], 1.0]
[['veiltype_w'], ['veilcolor_p'], 1.0]
[['gillspacing_f', 'stalkcolorabovering_s'], ['veilcolor_p'], 1.0]
[['capsurface_x'], ['gillspacing_f', 'veilcolor_p'], 1.011373081191739]
[['gillspacing_f'], ['capsurface_x', 'veilcolor_p'], 1.011373081191739]
[['capsurface_x', 'gillspacing_f'], ['veilcolor_p'], 1.0]
[['capsurface_x', 'veilcolor_p'], ['gillspacing_f'], 1.011373081191739]
[['gillspacing_f', 'veilcolor_p'], ['capsurface_x'], 1.011373081191739]
[['gillcolor_b', 'ringnumber_w'], ['veilcolor_p'], 1.0]
[['ringnumber_w', 'ringtype_o'], ['stalksurfaceabovering_b'], 1.0792840331913152]
[['gillsize_c'], ['habitat_v', 'ringnumber_w', 'ringtype_o'], 1.0816067161619274]

Dataset 3 (Nursery):

Support: 800

The top 10 rules with lift as measure are:
[['form_completed', 'health_not_recom'], ['class_value_not_recom'], 3.0]
[['has_nurs_very_crit', 'health_priority'], ['class_value_spec_prior'], 3.1713649851632053]
[['has_nurs_critical', 'health_not_recom'], ['class_value_not_recom'], 3.0]
[['has_nurs_very_crit', 'health_not_recom'], ['class_value_not_recom'], 3.0]
[['has_nurs_proper', 'health_not_recom'], ['class_value_not_recom'], 3.0]
[['children_2', 'class_value_not_recom'], ['health_not_recom'], 3.0]
[['children_2', 'health_not_recom'], ['class_value_not_recom'], 3.0]
[['form_incomplete', 'health_not_recom'], ['class_value_not_recom'], 3.0]
[['children_3', 'class_value_not_recom'], ['health_not_recom'], 3.0]
[['children_3', 'health_not_recom'], ['class_value_not_recom'], 3.0]

[['has_nurs_improper', 'health_not_recom'], ['class_value_not_recom'], 3.0]

Support : 900

The top 10 rules with lift as measure are:
[['form_incomplete', 'health_not_recom'], ['class_value_not_recom'], 3.0]
[['children_more', 'class_value_not_recom'], ['health_not_recom'], 3.0]
[['children_more', 'health_not_recom'], ['class_value_not_recom'], 3.0]
[['health_not_recom'], ['class_value_not_recom'], 3.0]
[['class_value_not_recom'], ['health_not_recom'], 3.0]
[['finance_inconv', 'health_not_recom'], ['class_value_not_recom'], 3.0]
[['children_1', 'class_value_not_recom'], ['health_not_recom'], 3.0]
[['children_1', 'health_not_recom'], ['class_value_not_recom'], 3.0]
[['finance_convenient', 'health_not_recom'], ['class_value_not_recom'], 3.0]
[['children_2', 'class_value_not_recom'], ['health_not_recom'], 3.0]
[['children_2', 'health_not_recom'], ['class_value_not_recom'], 3.0]

Support : 1000

The top 10 rules with lift as measure are:
[['children_3', 'class_value_not_recom'], ['health_not_recom'], 3.0]
[['children_3', 'health_not_recom'], ['class_value_not_recom'], 3.0]
[['children_2', 'class_value_not_recom'], ['health_not_recom'], 3.0]
[['children_2', 'health_not_recom'], ['class_value_not_recom'], 3.0]
[['health_not_recom'], ['class_value_not_recom'], 3.0]
[['class_value_not_recom'], ['health_not_recom'], 3.0]
[['children_more', 'class_value_not_recom'], ['health_not_recom'], 3.0]
[['children_more', 'health_not_recom'], ['class_value_not_recom'], 3.0]
[['finance_convenient', 'health_not_recom'], ['class_value_not_recom'], 3.0]
[['form_incomplete', 'health_not_recom'], ['class_value_not_recom'], 3.0]
[['form_completed', 'health_not_recom'], ['class_value_not_recom'], 3.0]

On using ‘lift’ in place of confidence the majority of top 10 rules generated go changed. Theoretically confidence has a disadvantage, because it ignores the support of the itemset in the rule consequent. This is where using lift has the advantage because it also counts the support of the item set in the rule consequent. Lift is the ratio of confidence of the rule and the support of the item set in the rule consequent. Hence using lift has a better advantage over using confidence.

1. The minimum and the maximum number of thresholds for the required condition would be 3 and ‘n’ respectively. Since for the threshold to be maximum the true positive and false positive should not be at separate sides of the graph.

Summary: 4a followed by 4b.

Summary
- An impossibility Theorem for Clustering

Introduction to the author:

The author of this paper, Dr. Jon Klienberg is the chair person at the Information Science department and Professor Computer Science department at Cornell University. His major research areas focus on issues relating to the interface of network and information with an emphasis on the social and information networks.

Abstract of the paper:

The core of this publication is based on clustering, application of the clustering function on the data set and its properties. Let us briefly understand the concept of clustering.
The general principle of 'clustering' is, where a set of heterogeneous data points are grouped into clusters based on the distance function applied on the data set. The distance function which is being applied varies based on the clustering function where the most common distance function being, Euclidean distance. The basic idea of clustering is that the point in the same clusters are closer than the distance between the points in any other cluster.

The crux of this paper is the 'Impossibility theorem', which is based on 3 properties of the clustering function applied to the data set. Namely,

1. Consistency
2. Richness
3. Scale Invariance

The argument of the author is that, there is no clustering function which can satisfy all the three properties. Now, let us analyze the theorem and its argument in detail.

The 3 properties:

All the 3 properties revolve around the clustering function applied to the data set.
Now, let us consider the clustering function (C), data set (D) with 'n' number of data points, and a distance function 'd' which calculates the pairwise distance between the data points using any distance measure.

Scalar Invariance:

To understand this property, let us consider a clustering function which applies 2 different distance functions on the data set. Let the different distance functions be d_a and d_b. According to the rule, on applying the 2 different distance functions on the same data set, the result of clustering should be the same if $d_a = \alpha d_b$ where $\alpha > 0$. The α in scalar invariance property is used

to negate the effect of different measurement scales used in 2 distance functions applied on the data set.

Richness:

As the name suggests, this property expresses that the output of the clustering function applied to the data set is equal to the number of partitions that can be generated from the data set. With this property even if the distance between the points are unknown, the clustering function applied to the data set could generate all possible clusters from the data set. This could also suggest that the clustering function was able to predict the number and proportion of clusters from the data set[1]. This can be compared to the brute force method used in classification or association rule mining where every possible outcome is determined and analyzed.

Consistency:

This property states that even if the distance between the points in the same clusters decreases, or the distance between the points in different clusters increases, the output of the clustering function applied to the dataset should remain the same. Mathematically it can be explained as, if 2 different distance functions d_a and d_b is applied to the same data set and for every pair of points in the same cluster if the distance decreases i.e. $d_a >= d_b$, and for every pair of points belonging to different clusters the distance increases $d_a <= d_b$ the clustering should provide the same output, $C(d_a) = C(d_b)$ [1].

Now that we have discussed the 3 main properties that defines the theorem, we will look at the 'Impossibility Theorem'

According to the theorem, for any pair of points where n >= 2 there is no clustering function that satisfies all the 3 properties explained above.

To understand the theorem, consider the single linkage clustering. In single linkage clustering, the proximity of 2 clusters is defined as the minimum of the distance between any 2 points in different clusters. For the process of creating clusters each and every point is considered as singleton clusters and the shortest links between the points are added one at a time. These links combine the points to form the clusters [2]. The clustering function can be formed by choosing a stoppage condition for the single linkage clustering.

The author discuses 3 types of stopping condition:

1. K cluster stopping condition
2. Distance – r stopping condition
3. Scale - α stopping condition.

, and to choose a stopping condition that satisfies 2 out of the 3 properties.

Let us briefly understand the stopping conditions used to form the clustering function.

K- cluster stopping condition:

This type of stopping condition stops the clustering mechanism, when the required number of clusters(k) is reached. It is similar to k- means clustering algorithm. This stopping condition violates the richness property but satisfies others. Since the process of clustering is stopped when k value is reached, the function producing all possible clusters of the data set may not be possible.

Distance – r stopping condition:

This condition stops the clustering process when the nearest 2 clusters are farther than a predefined distance r [1]. It satisfies richness by forming all the clusters of the dataset. It violates Scale Invariance property.

Scale - α stopping condition:

This condition is similar to the Distance – r stopping condition, except that the distance between 2 clusters is farther than the fraction of maximum distance between 2 points.
It satisfies scale invariance and richness.

Antichains and Clustering:

In a partially ordered set, if 2 points are incomparable where operations such as x >= y, x<= y are not possible between the points they are referred as antichains [3].

The author states 2 theorems for the proof:

According to the first theorem if the clustering function C, satisfies the properties of Scale invariance and consistency, then all the partitions produced by the clustering function becomes antichain.

According to the second theorem, if a partition has antichains, then there is a clustering function where all the clusters generated by the function is equal to the partition and it satisfies the properties of Scale invariance and consistency. This can be considered as the inverse of the first theorem.

Centroid based clustering:

In this section the author describes the general principle of centroid based clustering where k centroids are chosen prior to the clustering, and the clusters are formed based on the distance between the data points and the centroid.

The author argues that in centroid based clustering, where the number of clusters is >= 2 applied on a relatively larger data set, the clustering function does not satisfy the Consistency property.

In the last section the author tries to suggest and formulate few compromises to the consistency property so that the clustering function can be formed which satisfies the other 2 properties and a compromised variant of Consistency property.

Conclusion:

The perspective of the author to prove the negativity of the clustering function and to define axioms to show no clustering function can satisfy all three sounded bit a negative approach. Still in the early stages of understanding clustering so found the overwhelming information gained as a positive of the publication.

References:

[1] http://alexhwilliams.info/itsneuronalblog/2015/10/01/clustering2/
[2] https://en.wikipedia.org/wiki/Antichain
[GENERAL] https://www.cs.cornell.edu/home/kleinber/nips15.pdf

Summary
Measures of Clustering Quality: A Working Set of Axioms for Clustering

Introduction to the authors:

The author of this publication, Dr. Margareta Ackerman, is an Assistant Professor at the Florida State University, Computer Science department. Area of research revolves around Machine Learning, but major focus is on 'Clustering' and the gap between the theoretical explanation and application of clustering.

The other author Dr. Shai Ben David is currently a professor at University of Waterloo, school of Information science.

Both the speakers focus on application and the theoretical explanation of Machine learning and Data Mining paradigms.

Abstract:

The main essence of the paper is to explain that the 'Impossibility Theorem' by Dr. Jon Klienberg is not an innate or natural characteristic of clustering rather, it occurs as result of particular claims in his publication. They also propose a clustering-quality measure (CQM). When a data set is passed to this function, it not only forms the clusters also calculates a quality measure of the clustering which is a real number. They stress the point that instead of framing axioms and properties and proving them to be impossible, it is better to analyze the quality of data clustering at hands.

Introduction to Clustering-quality measure(CQM):

Let us understand the CQM measure in detail as it forms the base of opposing the axioms of Impossibility theorem.

A CQM measure is a function, which forms clusters of the input data set and match the pair of data and clusters to a real number to analyze the quality of the clustering. The use of CQM measure is beyond opposing the impossibility theorem, it also helps to rate the clustering mechanism. Since the prime goal of clustering, like any other data mining paradigm is to analyze hidden patterns, hence analyzing the quality of the clustering can alter the judging scale of whether or not to rely on the analysis. It can also be used to compare a variety of clustering

process when applied on the same data set which can be used to increase the efficiency of the process.

Even in this publication few axioms are proposed but to analyze the quality of clustering.

Brief Recap of Klienberg's Axiom:

3 axioms are proposed by Dr.Klienberg:

1. **Scale Invariance** is the property by which applying 2 different distance functions on the same data sets, it provides same results provided if da = αdb where α>0 (as explained in the previous summary)
2. **Consistency** stating that the output of the clustering function remains the same, even if the distance between the data points in the same cluster or different cluster decreases or increases respectively.
3. **Richness** stating that the output of the clustering function equals the total number of partitions that can be generated from the data set.

And he argues that there is no clustering function that can satisfy all the 3 properties listed above.

Axioms of Clustering Quality Measures:

The 3 axioms of Dr. Klienberg were rephrased for the quality control measure.

Scale Invariance: Consider a cluster formed from the input data set, if the cluster is scaled by λ, where λ is strictly positive, it is stated that the Clustering-quality measure of the cluster will be greater than the original cluster.

Consistency: Consider 2 distance functions d_a and d_b where d_b is the transformed function of d_a, the transformation of the distance function is done for the same cluster, it is plausible that the Clustering-quality measure will be higher than the original cluster.

Richness:

Apart from the three axioms, a new ratio of distance between a center and the closest point to the distance from the same center to the second closest point is defined as 'relative margin'.
The lower the value of relative margin higher the CQM measure of the clustering.
The relative margin is stated to be satisfying the 3 rephrased axioms.

Representation of Axioms:

To express the trueness and representation of the axioms we can discuss on two properties of the axioms namely, Soundness and Completeness. The soundness is defined to express the consistency of the axioms, by showing that each element in the class satisfies all the three axioms. The completeness means the range of representation of the axioms i.e. the complete set of properties exhibited by the class variables should be expressed by the axioms.

But as simple as it sounds, it is quite intricate to prove the properties of the axioms described above, as there is prescribed definition for clustering. In that case we settle for 'relaxed Clustering'[1] and 'relaxed completeness'[1] by considering the functions which may not be directly considered as clustering functions.

A solution to the complexness described above is isomorphic invariance, according to which 2 clusters are termed isomorphic, if two points from different function can share the same cluster when the functions of 2 different points share the same cluster. This property is now included in the consistent set of axioms.

Now, let us understand a few examples of Clustering Quality measures.

Weakest Link:

This is a concept of linkage based clustering, where cluster contains a pair of points and they are connecting through a tight chain of other points in the cluster. The longest link of this chain is known as the weakest link. There are various mathematical representations of the weakest links.

Computational Complexity:

It is useful to apply the Clustering Quality Measures to analyze the quality of the clustering but it is complex. For example, using relative margin the complexity is $O(n^{k+1})$ [1].

Limitations:

We have been discussing how to use the CQM to analyze the quality of clustering with independent on the number of clusters. There is also a other case where the quality of clustering depends on the loss and the number of clusters. The consequence of this loss function is it fails to satisfy scale invariance and richness. Below are few techniques explained to overcome this: [1]

1. L – Normalization
2. Refinement and Coarsening
3. Refinement / Coarsening

Conclusion:

As a budding data scientist, I personally felt this publication had clearer and representative point of view on clustering. Instead of framing axioms to prove the clustering function will not be able to satisfy that, the rephrased axioms were able to make me understand the concept better. It was more direct and proofs were consistent with the axioms.

But felt it had more theoretical explanations which is more common in publications, more real life examples would have been better is my humble opinion.

Refrences:

[1] https://pdfs.semanticscholar.org/01c3/b7cc80be991ef4741de83dc113d91856f2a5.pdf
[2] http://slideplayer.com/slide/4462606/