Assignment 4
B565 – Data Mining

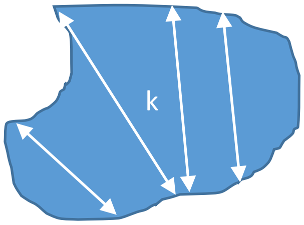
(rutakulk)

Question 2

Let {d(xi , xj)}ij be a set of n 2 Euclidean distances between pairs of n distinct objects from some space X . Suppose the data set of objects is partitioned into m groups, where {cj} m j=1 is a set of m centroids for each (non-overlapping) subset of objects j; that is, for each subgroup of objects j ∈ {1, 2 . . . m}, cj is their centroid. Now, if two of the m groups of objects are to be merged to minimize the sum of squared errors between each data point and its respective centroid, show how you will compute which two clusters to merge out of m 2 pairs in total. Note that you are only allowed to work with distances between these points because you are not given the original objects xi . You are not allowed to use any embedding techniques such as multidimensional scaling, to map these points into a Euclidean space. Prove any complex statements in your algorithm.

➔ Let there be k clusters, having n points in all. We have a mapping of which point belong to which cluster i.e point → cluster mapping. Assuming there are 'n' points in total.
In order to find out which clusters could be readily merged, we will find out boundary of the cluster. Boundary point of the cluster are assumed to the points belonging to the same cluster having maximum k distances. The idea is to generate 'k' diagonals what would be sufficient to generate the boundary of the cluster.
Higher the value of k, more definite will be the boundary.

Boundary point diagonal is calculated by taking distance of a single point with every other point in the cluster.
When the highest distance is found, we will get 2 points in the boundary.



Suppose we have generated the boundary using boundary points, we can find out the centroid. Centroid can be found out by taking the average of all the boundaries.
This centroid may not be the actual centroid of the cluster, but it will be somewhere in or near to the central region of the cluster.

We will find out such centroids for all 'k' clusters. Once we get the centroid or central region point for all clusters, we can easily combine two clusters having minimum distance between the centroids.

Question 3

Apriori algorithm. Implement the Apriori algorithm by first determining frequent itemsets and then proceeding to identify association rules. Consider that the input to your program is a sparse matrix where the rows are transactions and columns are items. Each value in your matrix is a binary variable from {0, 1} that indicates presence of an item in the transaction

➔ In order to convert any dataset into a sparse matrix, every unique value for every column is converted into a column of Sparse Matrix. Thus the entry dataMatrix[i][j] =1 only if the item is present in the transaction. Thus transactions can be easily represented.

a) *Implement both Fk−1 × F1 and Fk−1 × Fk−1 methods. Allow in your code to track the number of generated candidate itemsets as well as the total number of frequent itemsets.*

➔ While implementing both the methods of generating frequent itemsets, it was observed that the number of itemsets generated by both the methods are different whereas both the methods generate same frequent itemsets.

➔ In Fk-1 x F1 method, we combine the itemsets generated in previous iteration with 1-itemsets to get a new set of itemsets of one length greater.

➔ In Fk-1 x Fk-1 method, we combine the itemsets generated in the previous iteration with itemsets generated in the iteration.

b) *Use three data sets from the UCI Machine learning repository to test your algorithms. The data sets should contain at least 1000 examples, and at least one data set should contain 10,000 examples or more. You can convert any classification or regression data set into a set of transactions and you are allowed to discretize all numerical features into two or more categorical features. Compare these two candidate generation methods on each of the three data sets for three different meaningful levels of the minimum support threshold (the thresholds should allow you to properly compare different methods and make useful conclusions). Provide the numbers of candidate itemsets considered in a table and discuss the observed savings that one of these methods achieves.*

➔ I have used the following datasets from the UCI Machine learning repository-

1. Car Dataset – (1728)
2. Contraceptive Dataset – (1473)
3. Nursery Dataset – (12960)

Car Evaluation:

Threshold	**Properties**	**F_{k-1} X F_1**	**F_{k-1} X F_{k-1}**
1%	Total Itemsets	9544	6981
	Frequent Itemsets	2291	2291
5%	Total Itemsets	2471	2342
	Frequent Itemsets	502	571
10%	Total Itemsets	1297	945
	Frequent Itemsets	227	227

Contraceptive

Threshold	Properties	F_{k-1} X F_1	F_{k-1} X F_{k-1}
5%	Total Itemsets	8467	6473
	Frequent Itemsets	1445	1445
10%	Total Itemsets	3085	
	Frequent Itemsets	436	
15%	Total Itemsets	1234	633
	Frequent Itemsets	233	233

Nursery:

Threshold	Properties	F_{k-1} X F_1	F_{k-1} X F_{k-1}
7%	Total Itemsets	2275	1938
	Frequent Itemsets	299	299
5%	Total Itemsets	3973	4319
	Frequent Itemsets	559	601
10%	Total Itemsets	1156	904
	Frequent Itemsets	159	159

3c)

- Car Evaluation:

Support Threshold	Frequent Itemsets	Frequent closed itemsets	Maximal frequent itemsets
1%	2291	1951	1099
5%	350	310	221
10%	87	74	53

- Contraceptive Method Choice:

Support Threshold	Frequent Itemsets	Frequent closed itemsets	Maximal frequent itemsets
5%	1445	1392	804
8%	640	628	354
10%	436	432	244

- Nursery:

Support Threshold	Frequent Itemsets	Closed Frequent Itemsets	Maximal Frequent itemsets
5%	559	517	438
7%	299	280	243
10%	159	148	114

3d)

- Car Evaluation:

Minimum Support Threshold	Minimum Confidence Threshold	Number of generated confidence rules
1%	0.3	4824
5%	0.5	212
10%	0.7	22

- Contraceptive Method Choice:

Minimum Support Threshold	Minimum Confidence Threshold	Number of generated confidence rules
5%	0.7	3178
8%	0.5	1160
10%	0.5	700

- Nursery:

Minimum Support Threshold	Minimum Confidence Threshold	Number of generated confidence rules
5%	0.5	426
7%	0.5	54
10%	0.5	22

3e) Car Evaluation

Minimum Support Threshold	Minimum Confidence Threshold	Top 10 Association Rules generated
1%	0.3	i. ['buying - vhigh', 'maint- vhigh'] --> ['class-unacc'] ii. ['buying - vhigh', 'persons - 2'] --> ['class - unacc'] iii. ['buying - vhigh', 'safety - low'] --> ['class- unacc'] iv ['buying - high', 'maint - vhigh'] --> ['class-unacc'] v. ['maint - high', 'persons - 2'] --> ['class - unacc'] vi. ['maint - low', 'persons - 2'] --> ['class - unacc'] vii. ['doors', '4 - safety - low'] --> ['class - unacc'] viii. ['lug_boot - big', 'safety - low'] --> ['class -unacc'] ix. ['persons - more', 'safety - low'] --> ['class - unacc'] x. ['lug_boot-med', 'safety-low'] --> ['class-unacc']
	0.5	
	0.7	
5%	0.3	i. ['buying - vhigh', 'maint- vhigh'] --> ['class-unacc'] ii. ['buying - vhigh', 'persons - 2'] --> ['class - unacc'] iii. ['buying - vhigh', 'safety - low'] --> ['class- unacc'] iv ['buying - high', 'maint - vhigh'] --> ['class-unacc'] v. ['maint - high', 'persons - 2'] --> ['class - unacc'] vi. ['maint - low', 'persons - 2'] --> ['class - unacc'] vii. ['doors', '4 - safety - low'] --> ['class - unacc'] viii. ['lug_boot - big', 'safety - low'] --> ['class -unacc'] ix. ['persons - more', 'safety - low'] --> ['class - unacc'] x. ['lug_boot-med', 'safety-low'] --> ['class-unacc']
	0.5	
	0.7	
10%	0.7	i. ['persons - '2', 'lug_boot - small'] --> ['class-unacc'] ii. ['persons - 2', 'lug_boot - med'] --> ['class-unacc'] iii. ['persons - 2', 'lug_boot - big'] --> ['class- unacc'] iv. ['persons-2', 'safety-low'] --> ['class-unacc'] v. ['persons'-4', 'safety-low'] --> ['class-unacc'] vi. ['persons-2', 'safety-med'] --> ['class-unacc'] vii. ['persons-2', 'safety-high'] --> ['class-unacc'] viii. ['persons-4', 'safety-low'] --> ['class-unacc'] ix. ['persons- more', 'safety-low'] --> ['class-unacc'] x. ['lug_boot-small', 'safety-low'] --> ['class-unacc']
	0.5	
	0.3	

Comment: The rule - ['buying - vhigh', 'maint- vhigh'] --> ['class-unacc'] states that if the Buying expenditure is very high and maintenence is very high then, the car is unacceptable. This is a valid rule which related the 2 basic components of expenciture and maintainence related to cars to the acceptance rate.

Contraceptive Method Choice

Minimum Support Threshold	Minimum Confidence Threshold	Top 10 Association Rules generated
5%	0.3	['4', '1'] --> ['1', '0', '1'] ['4', '1', '1'] --> ['0', '1'] ['4', '1', '1', '0'] --> ['1'] ['4'] --> ['1', '1', '4', '0'] ['4', '1'] --> ['1', '4', '0'] ['4', '1', '1'] --> ['4', '0'] ['4', '1', '1', '4'] --> ['0'] ['4'] --> ['1', '1', '4', '0'] ['4', '1'] --> ['1', '4', '0'] ['4', '1', '1'] --> ['4', '0']
	0.5	
	0.7	
8%	0.3	['4', '1'] --> ['1', '0', '1'] ['4', '1', '1'] --> ['0', '1'] ['4', '1', '1', '0'] --> ['1'] ['4'] --> ['1', '1', '4', '0'] ['4', '1'] --> ['1', '4', '0'] ['4', '1', '1'] --> ['4', '0'] ['4', '1', '1', '4'] --> ['0'] ['4'] --> ['1', '1', '4', '0'] ['4', '1'] --> ['1', '4', '0'] ['4', '1', '1'] --> ['4', '0']
	0.5	
	0.7	
10%	0.3	['1', '1'] --> ['0'] ['1'] --> ['1', '1'] ['1', '1'] --> ['1'] ['1'] --> ['2', '0'] ['1', '2'] --> ['0'] ['1'] --> ['3', '0'] ['1', '3'] --> ['0'] ['1'] --> ['1', '0'] ['1', '1'] --> ['0'] ['1'] --> ['3', '0'] ['1'] --> ['0', '1']
	0.5	
	0.7	

Comment: [wife education -1, husband education – 1 → contraception – low]
Here it means if both husband and wife are uneducated, the n contraceptives are of no use. There is a high chance that this could be the case for valid rule.

Nursery

Minimum Support Threshold	Minimum Confidence Threshold	Top 10 Association Rules generated	Comments
5%	0.7	i. ['housing-less_conv', 'finance-convenient', 'health-not_recom'] --> ['classAttr-not_recom'] ii. ['housing-less_conv', 'finance-inconv', 'health-not_recom'] --> ['classAttr-not_recom'] iii. ['housing-critical', 'finance-convenient', 'health-not_recom'] --> ['classAttr-not_recom'] iv. ['housing-critical', 'finance-inconv', 'health-not_recom'] --> ['classAttr-not_recom'] v. ['parents-usual', 'health-not_recom'] --> ['classAttr-not_recom'] vi. ['parents- pretentious', 'health-not_recom'] --> ['classAttr-not_recom'] vii. ['parents-great_pret', 'health-not_recom'] --> ['classAttr-not_recom'] viii. ['has_nurs-proper', 'health-not_recom'] --> ['classAttr-not_recom'] ix. ['has_nurs-less_proper', 'health-not_recom'] --> ['classAttr-not_recom'] x. ['has_nurs-improper', 'health-not_recom'] --> ['classAttr-not_recom']	The top 10 rules generated for threshold confidence of 0.7, 0.5 and 0.3 remain same as there are more than 10 rules of confidence 1.0. All these rules will be part of the top rules generated for 0.3,0.5 and 0.7 confidence.
	0.5		
	0.3		
7%	0.7	i. ['housing-less_conv', 'finance-convenient', 'health-not_recom'] --> ['classAttr-not_recom'] ii. ['housing-less_conv', 'finance-inconv', 'health-not_recom'] --> ['classAttr-not_recom'] iii. ['housing-critical', 'finance-convenient', 'health-not_recom'] --> ['classAttr-not_recom'] iv. ['housing-critical', 'finance-inconv', 'health-not_recom'] --> ['classAttr-not_recom'] v. ['parents-usual', 'health-not_recom'] --> ['classAttr-not_recom'] vi. ['parents- pretentious', 'health-not_recom'] --> ['classAttr-not_recom'] vii. ['parents-great_pret', 'health-not_recom'] --> ['classAttr-not_recom'] viii. ['has_nurs-proper', 'health-not_recom'] --> ['classAttr-not_recom'] ix. ['has_nurs-less_proper', 'health-not_recom'] --> ['classAttr-not_recom']	The top 10 rules generated for threshold confidence of 0.7, 0.5 and 0.3 remain same as there are more than 10 rules of confidence 1.0. All these rules will be part of the top rules generated for 0.3,0.5 and 0.7 confidence.
	0.5		
	0.3		

		x. ['has_nurs-improper', 'health-not_recom'] --> ['classAttr-not_recom']	
10%	0.7	i. ['housing-less_conv', 'finance-convenient', 'health-not_recom'] --> ['classAttr-not_recom'] ii. ['housing-less_conv', 'finance-inconv', 'health-not_recom'] --> ['classAttr-not_recom'] iii. ['housing-critical', 'finance-convenient', 'health-not_recom'] --> ['classAttr-not_recom'] iv. ['housing-critical', 'finance-inconv', 'health-not_recom'] --> ['classAttr-not_recom'] v. ['parents-usual', 'health-not_recom'] --> ['classAttr-not_recom'] vi. ['parents- pretentious', 'health-not_recom'] --> ['classAttr-not_recom'] vii. ['parents-great_pret', 'health-not_recom'] --> ['classAttr-not_recom'] viii. ['has_nurs-proper', 'health-not_recom'] --> ['classAttr-not_recom'] ix. ['has_nurs-less_proper', 'health-not_recom'] --> ['classAttr-not_recom'] x. ['has_nurs-improper', 'health-not_recom'] --> ['classAttr-not_recom']	The top 10 rules generated for threshold confidence of 0.7, 0.5 and 0.3 remain same as there are more than 10 rules of confidence 1.0. All these rules will be part of the top rules generated for 0.3,0.5 and 0.7 confidence.
	0.5		
	0.3		

Comment:

['parents- pretentious', 'health-not_recom'] --> ['classAttr-not_recom']

In this rule, if parents are prententious and health is not recommendable, then the child is not recommended for nursery education.

3f)

Lift vs Confidence-

Lift generates comparatively better rules than confidence. It takes into account both left and right side probabilities. These rules generated are more reliable. We do not use pruning in Lift based rule generation, and generate all possible rules from which we can select the best rule.

- Car Evaluation:

Minimum Support Threshold	Top 10 Association Rules
1%	i. ['more', 'big', 'high'] --> ['vgood'] ii. ['more', 'big'] --> ['high', 'vgood'] iii. ['more', 'big', 'high'] --> ['acc'] iv. ['more', 'big', 'med'] --> ['acc'] v. ['more', 'med', 'high'] --> ['acc'] vi. ['more', 'med', 'med'] --> ['acc'] vii. ['more', 'small', 'high'] --> ['acc'] viii. ['4', 'big'] --> ['high', 'vgood'] ix. ['4', 'med', 'high'] --> ['acc'] x. ['4', 'med', 'med'] --> ['acc']
10%	i. ['2'] --> ['small', 'unacc'] ii. ['2', 'small'] --> ['unacc'] iii. ['2'] --> ['med', 'unacc'] iv. ['2'] --> ['big', 'unacc'] v. ['2'] --> ['low', 'unacc'] vi. ['2'] --> ['med', 'unacc'] vii. ['2'] --> ['high', 'unacc'] viii. ['4'] --> ['low', 'unacc'] ix. ['big'] --> ['low', 'unacc'] x. ['big', 'low'] --> ['unacc']
5%	i. ['4', 'med'] --> ['acc'] ii. ['4', 'high'] --> ['acc'] iii. ['more', 'med'] --> ['acc'] iv. ['more', 'high'] --> ['acc'] v. ['2'] --> ['high', 'unacc'] vi. ['2', '2'] --> ['unacc'] vii. ['vhigh', 'low'] --> ['unacc'] viii. ['high', '2'] --> ['unacc'] ix. ['high', 'low'] --> ['unacc'] x. ['med', '2'] --> ['unacc']

Contraceptive Method Choice:

Minimum Support Threshold	Top 10 Association Rules generated
5%	['4', '4', '1'] --> ['1', '1', '4', '0'] ['4', '4', '1', '1'] --> ['1', '4', '0'] ['4', '4', '1', '1', '1'] --> ['4', '0'] ['4', '4', '1', '1', '1', '4'] --> ['0'] ['4'] --> ['4', '1', '1', '4', '0', '2'] ['4', '4'] --> ['1', '1', '4', '0', '2'] ['4', '4', '1'] --> ['1', '4', '0', '2'] ['4', '4', '1', '1'] --> ['4', '0', '2'] ['4', '4', '1', '1', '4'] --> ['0', '2']
10%	['4'] --> ['4', '1', '1', '4', '0'] ['4'] --> ['4', '1', '1', '4', '0'] ['4', '4'] --> ['1', '1', '4', '0'] ['4', '4', '1'] --> ['1', '4', '0'] ['4', '4', '1', '1'] --> ['4', '0'] ['4', '4', '1', '1', '4'] --> ['0'] ['4'] --> ['1', '1', '1', '4', '0'] ['4', '1'] --> ['1', '1', '4', '0'] ['4', '1', '1'] --> ['1', '4', '0']
12%	['4'] --> ['4', '1', '1', '4', '0'] ['4', '4'] --> ['1', '1', '4', '0'] ['4', '4', '1'] --> ['1', '4', '0'] ['4', '4', '1', '1'] --> ['4', '0'] ['4', '4', '1', '1', '4'] --> ['0'] ['4'] --> ['4', '1', '1', '4', '0'] ['4', '4'] --> ['1', '1', '4', '0'] ['4', '4', '1'] --> ['1', '4', '0'] ['4', '4', '1', '1'] --> ['4', '0'] ['4', '4', '1', '1', '4'] --> ['0']

Nursery:

Minimum Support Threshold	Top 10 Association Rules generated
10%	['usual', 'not_recom'] --> ['not_recom'] ['great_pret', 'not_recom'] --> ['not_recom'] ['convenient', 'not_recom'] --> ['not_recom'] ['less_conv', 'not_recom'] --> ['not_recom'] ['critical', 'not_recom'] --> ['not_recom'] ['convenient', 'not_recom'] --> ['not_recom'] ['inconv', 'not_recom'] --> ['not_recom'] ['nonprob', 'not_recom'] --> ['not_recom'] ['slightly_prob', 'not_recom'] --> ['not_recom'] ['problematic', 'not_recom'] --> ['not_recom']
8%	['usual', 'not_recom'] --> ['not_recom'] ['pretentious', 'not_recom'] --> ['not_recom'] ['great_pret'] --> ['not_recom', 'not_recom'] ['great_pret', 'not_recom'] --> ['not_recom'] ['complete', 'not_recom'] --> ['not_recom'] ['completed'] --> ['not_recom', 'not_recom'] ['completed', 'not_recom'] --> ['not_recom'] ['incomplete'] --> ['not_recom', 'not_recom'] ['incomplete', 'not_recom'] --> ['not_recom'] ['foster', 'not_recom'] --> ['not_recom']
15%	['convenient'] --> ['not_recom', 'not_recom'] ['convenient', 'not_recom'] --> ['not_recom'] ['inconv'] --> ['not_recom', 'not_recom'] ['inconv', 'not_recom'] --> ['not_recom'] ['convenient'] --> ['not_recom', 'not_recom'] ['convenient', 'not_recom'] --> ['not_recom'] ['inconv'] --> ['not_recom', 'not_recom'] ['inconv', 'not_recom'] --> ['not_recom'] ['convenient'] --> ['not_recom', 'not_recom'] ['convenient', 'not_recom'] --> ['not_recom'] ['inconv'] --> ['not_recom', 'not_recom'] ['inconv', 'not_recom'] --> ['not_recom']

Question 4

4a) (15 points) Kleinberg J. The impossibility theorem for clustering. Advances in Neural Information Processing Systems, NIPS 2002.

Solution:

Impossibility Theorem Summary

Clustering means grouping similar objects from a diverse set of objects. The similarity between different objects is usually measured in the form of distance between two objects. Smaller the distance, greater the similarity between objects. Though similarity is one of the measures for forming clusters, it is vague. Every dataset has its own definition of similarity. Clustering is more of an intuitive approach when we look at it from the algorithmic point of view. There is a diverse range of clustering algorithms, which lead into different type of result qualities.

Earlier work in clustering analysis involves axiomatic approaches by Jardine and Sibson, who proposed that in hierarchical clustering, single linkage is a consistent unique function that measures clustering. Puzicha on the other hand, focuses more on the cost efficiency of the partition function for clustering. Whereas few researchers have recently proposed properties that are sufficient to uniquely specify any clustering function.

In this paper, Kleinberg has given an axiomatic framework for clustering, which will analyze the clustering function irrespective of the algorithm used. In order to measure clustering, Kleinberg has proposed that clustering obeys *the impossibility theorem* which states that- There is no clustering function which satisfies all the three properties of –scale invariance, richness, consistency. A clustering function may easily obey two of the earlier stated properties, but not all three.

The Impossibility Theorem:

Consider a clustering function *f* that operates on Set S having n points, where n >= 2. This function *f* takes *d* on a Set *S* to return partition *T* in *S*.

- **Scale-Invariance:** This property states that any clustering function should be independent of the scaling in distance. Thus for given function *d* and *α* >0, *f(d) = f(α.d)*
- **Richness:** If Range (f) denotes all partitions of set S which is T, then richness means to be able to construct distance function *d* even when we are not aware of the distances between points.
- **Consistency:** Clustering function is said to be consistent if we get the same result even after shrinking or expanding the distance between points in same or different clusters respectively.

Kleinberg went ahead and showed that any two of the three properties can be satisfied by a given clustering function using single-linkage procedure. Initially singe-linkage assigns each point to be a cluster in its own. It then combines points with minimum distances till a stopping condition is reached. Stopping conditions can be any of the suitable stopping condition among *k-cluster*, *distance-r* or *scale-α.*

The author has also given a beautiful analysis of how centroid based clustering, which is one of the most common clustering methods, is contrastingly related to the property of Consistency. It is very difficult to generate a consistent k-means or k-median clustering approach as choosing the centroids initially in fairly based on intuition. There is no specific method for generating initial set of centroids.As the impossibility theorem suggests, it is impossible for all the three properties namely, scale variance, richness and consistency, to be satisfied for any given clustering function. Hence every time two properties are satisfied we can say that the remaining third property is relaxed. For example, Single linkage approach satisfies both richness and consistency but, relaxes the scale-invariance property.

Strengths:

- The paper has very well explained the importance of scale-invariance, richness and consistency on the generalized form of clustering. The impossibility theorem, explains that two of the above properties could be easily satisfied for any given clustering algorithm.
- The author has also proved the impossibility theorem by giving simple and easily understandable proof. He has also gone ahead to give examples of relaxation of every property and how other two properties are satisfied for particular clustering function.

Weakness:

- Kleinberg has not mentioned why he chose only three properties – scale invariance, richness and consistency and what is the importance of other properties such as accuracy, efficiency, etc.
- Kleinberg's explanation to the consistency model does not seem to be plausible. If the intra-cluster distances increase, there could be a possibility of new cluster being formed but Kleinberg does not consider this case.

4b) Ackerman M, Ben-David S. Measures of clustering quality: a working set of axioms for clustering. Advances in Neural Information Processing Systems, NIPS 2008.

Measures of Clustering Quality

The main idea of this paper is to change Kleinberg's idea of formulation based on clustering functions to clustering quality measures. Ackerman and David take a step ahead of Kleinberg's thought of generalizing the clustering theory. Kleinberg formalizes the clustering function by measuring it based on three principles whereas in this paper, the authors propose formalizing clustering quality function and not clustering function. Their focus is more on analyzing how strong and correct the clustering is. Clustering Quality Measure (CQM) helps in quantifying how good a cluster is formed or how correct the partition has been made. The author further explain how obvious the impossibility theorem is by giving an example of 3- clustering and 6-clustering. Whenever 3 clusters are scaled to 6, the consistency is affected. On the contrary, the authors explain how scale invariance, consistency and richness form a plausible set of axioms for CQM. They have also explained how the how they easily adhere to functions like Relative Point Margin, Representative Set, Relative Margin. In order to measure quality of Clustering, we have to make sure the axioms satisfy properties of both soundness and completeness. As we do not have a crisp definition of clustering, we cannot really say that only the axioms – scale invariance, richness and consistency suffice to measure quality of clustering. Contrastingly, there are also few non clustering functions which satisfy these three axioms. This contradicts soundness of axioms. Soundness fails with only earlier mentioned three axioms when formalizing the clustering quality measure. To overcome this, David and Ackerman introduced a new axiom – Isomorphism Invariance. This property is similar to permutation invariance, where clustering is not affected by individual clustered elements. The authors also mention that quality of clustering can be computed in polynomial time which again favors the efficiency of the approach. The authors have not only mentioned the correctness of the axioms they have specified for measuring cluster quality, but they have also provided alternative approach where the axioms will satisfy or not solely depends on the number of the clusters. They have thus added one more perspective of clustering quality measure which is number of clusters. In case of loss function, we know that both richness and scale invariance are hampered. By fixing the number of clusters, the authors have stated that scaling can be handled by normalization when the number of clusters are fixed whereas richness can be handled using refinement and coarse preference.

In conclusion, Ackerman and David have beautifully translated Kleinberg's axioms of measuring clustering function into a revised set of axioms analyzing Clustering Quality Measure. They have shown the consistency and accuracy of their proposal by providing different examples in support of their theory. They have not only proved consistency, but also showed that clustering quality can be computed in polynomial time.

References:

1. Contraceptive Method Choice
   https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice
2. Nurseryy
   https://archive.ics.uci.edu/ml/datasets/Nursery
3. Car Evaluation
   https://archive.ics.uci.edu/ml/datasets/Car+Evaluation