Data Mining **Assignment#4** **Manish Kumar**

(kumar20@iu.edu)

Problem:1 Given n data points with n_+ positive points and n_-negative points with individual posterior probability as p_i, we have to draw ROC curve and find the area under this curve.

Now, In the best case we can assume that all the posterior probability lie on a single line,i.e all points are collinear. In this case, if we get two points which are not same, we will be able to calculate the area under the ROC curve. **So, Minimum number of points to be able to calculate the area under the ROC curve should be 2.**

Now, In worst case, if no three points lie on a straight line then we need all the n points to calculate the area under the ROC curve. In general, the maximum number of points needed to calculate area under ROC curve should be total number of points subtracted by the number of extra points on each line which are not endpoints. That is we have to subtract all those points which are not endpoints and lie on straight line on ROC curve. That number should be the maximum number we need to calculate area under ROC curve.

Problem:2 We are given n data points which are clustered into m non overlapping groups where each cluster having centroid C_i.

By merging 2 clusters, our objective is to minimize the SSE. Whenever we group 2 clusters, we increase the SSE. In order to minimally increase it, we have to check which 2 clusters after joining gives the minimum increase. So, In order to check which 2 cluster will give the minimum increment, we have to see what is the resultant effects when we merge two clusters.

SSE of an individual cluster is given by:

$$SSE_j = \sum_{i=1}^{k}(C_i - x_i)^2$$

Where c_j is the centroid of cluster j and x are the points that belong to the cluster and k is the total number of points in that cluster.

Now, As given in the book, SSE can also be given as

$$SSE_j = \frac{1}{2p}\sum_{x\in Ci}\sum_{y\in C_i}(y_i - x_i)^2$$

Which is nothing but distances between each point in the cluster over which we do the summation and then divide it by total number of points in the cluster multiplied by 2.

Whenever we merge 2 clusters, we reduces the individual SSE of each cluster and adds SSE of the newly formed cluster.

So, We can write the total effect as:

$$SSE_{Effect} = SSE_{new} - \ (SSE_i + SSE_j)$$

Now, as per the question we are given the distance between each point in the n data points. So, Effectively we can calculate this term just by having the distance matrix between the points. To select 2 clusters which should be merged, we will see for which 2 clusters, this SSE effect will be minimu.

Algorithm:

```
Effect = infinity
points_save=()
For i in range(m):
    Calculate SSE(i)
    For j in range(i+1,m):
        Calculate SSE(j)
        Reduce_error = SSE(i) + SSE(j)
        Calculate SSE(new(i,j))
        Diff = SSE(i,j) - Reduce_error
        If Diff < Effect:
            Effect = Diff
            points_save(i,j)
```

In the end of program, points_save will give which 2 clusters should be merged.

Problem:3

A. In the program, I have generated the frequent itemset using $F_{k-1}*F_{k-1}$ as well as with $F_{k-1} * F_1$. To choose the respective method, enter the following option on the console screen. I am also generating the total number of candidate itemset and total number of frequent itemset generated by respective rule. To see all the generate frequent itemset, uncomment line No. 339 in the program.

NOTE: To set minimum support and confidence, set global variable Minsup and Min_confidence at the very beginning of the program.

B. Three data-sets that I have used to run this program on are:

- **Nursery Data Set**
- **Car Evaluation Data Set**

Data Mining **Assignment#4** **Manish Kumar**

(kumar20@iu.edu)

- **Tic-Tac-Toe Data Set**

For each dataset, I have includes a script (name of the script starts with “convert_dataset.py”) in python which converts the dataset into corresponding sparse matrix form on which program run. Nursery datasets have more than 12K rows as required by the question to have one data set more than 10k.

Data Set: Nursery	Minimum Support	Candidate Items	Frequent Items	Maximal Frequent	Closed Frequent
$F_{k-1} * F_{k-1}$	0.05	3785	903	600	792
$F_{k-1} * F_{k-1}$	0.10	652	131	104	121
$F_{k-1} * F_{k-1}$	0.20	351	27	27	27

Data Set: Nursery	Minimum Support	Candidate Items	Frequent Items	Maximal Frequent	Closed Frequent
$F_{k-1} * F_1$	0.05	3112	903	600	792
$F_{k-1} * F_1$	0.10	801	131	104	121
$F_{k-1} * F_1$	0.20	351	27	27	27

Data Set: Car	Minimum Support	Candidate Items	Frequent Items	Maximal Frequent	Closed Frequent
$F_{k-1} * F_{k-1}$	0.05	1264	204	183	190
$F_{k-1} * F_{k-1}$	0.10	264	48	39	43
$F_{k-1} * F_{k-1}$	0.20	210	21	21	21

Data Mining **Assignment#4** **Manish Kumar**

(kumar20@iu.edu)

Data Set: Car	Minimum Support	Candidate Items	Frequent Items	Maximal Frequent	Closed Frequent
$F_{k-1} * F_1$	0.05	1079	204	183	190
$F_{k-1} * F_1$	0.10	223	48	39	43
$F_{k-1} * F_1$	0.20	190	21	21	21

Data Set: Tic-tac-toe	Minimum Support	Candidate Items	Frequent Items	Maximal Frequent	Closed Frequent
$F_{k-1} * F_{k-1}$	0.05	6271	1545	905	1120
$F_{k-1} * F_{k-1}$	0.10	1565	325	187	243
$F_{k-1} * F_{k-1}$	0.20	391	42	26	35

Data Set: Tic-tac-toe	Minimum Support	Candidate Items	Frequent Items	Maximal Frequent	Closed Frequent
$F_{k-1} * F_1$	0.05	9275	1545	905	1120
$F_{k-1} * F_1$	0.10	2261	325	187	243
$F_{k-1} * F_1$	0.20	348	42	26	35

Observation: We can see from the data that both these methods generate same number of frequent Items as well as Maximal frequent and closed Frequent.

Data	Support	Confidence	Rule Generated	Savings
Nursery	0.10	0.03	208	400

Data Mining **Assignment#4** **Manish Kumar**

(kumar20@iu.edu)

Nursery	**0.05**	**0.03**	**1280**	**756**
Nursery	**0.03**	**0.03**	**3992**	**800**
Nursery	**0.10**	**0.02**	**310**	**475**
Nursery	**0.05**	**0.02**	**1464**	**883**
Nursery	**0.03**	**0.02**	**4342**	**971**
Nursery	**0.10**	**0.01**	**400**	**327**
Nursery	**0.05**	**0.01**	**1560**	**635**
Nursery	**0.03**	**0.01**	**4786**	**433**

Data	**Support**	**Confidence**	**Rule Generated**	**Savings**
Car	**0.10**	**0.03**	**42**	**194**
Car	**0.05**	**0.03**	**330**	**290**
Car	**0.03**	**0.03**	**438**	**338**
Car	**0.10**	**0.02**	**56**	**225**
Car	**0.05**	**0.02**	**360**	**341**
Car	**0.03**	**0.02**	**476**	**228**
Car	**0.10**	**0.01**	**74**	**427**
Car	**0.05**	**0.01**	**404**	**163**
Car	**0.03**	**0.01**	**524**	**241**

Data	**Support**	**Confidence**	**Rule Generated**	**Savings**
Tic-Tac	**0.10**	**0.03**	**944**	**74**

Tic-Tac	0.05	0.03	Rule Generated	20
Tic-Tac	0.03	0.03	649	88
Tic-Tac	0.10	0.02	943	210
Tic-Tac	0.05	0.02	771	183
Tic-Tac	0.03	0.02	330	253
Tic-Tac	0.10	0.01	407	110
Tic-Tac	0.05	0.01	205	237
Tic-Tac	0.03	0.01	358	194

E.)

There is no change in the top 10 rules with varying threshold of support and confidence. It will change only in the case when rule generated are less than 10 only. So, I am here just producing top 10 rules for all threshold of support and confidence which I have generated.

Top 10 Rules for Nursery:

health=not_recom	**-------->**	**rank=not_recom**
rank=not_recom	**-------->**	**health=not_recom**
children=1	**-------->**	**rank=not_recom**
children=1	**-------->**	**health=not_recom**
health=not_recom	**-------->**	**rank=not_recom**
children=2	**-------->**	**health=not_recom**
health=not_recom	**-------->**	**children=2**
children=3	**-------->**	**rank=not_recom**
children=more	**-------->**	**health=not_recom**
children=more	**-------->**	**rank=not_recom**

Data Mining Assignment#4 Manish Kumar
(kumar20@iu.edu)

Top 10 Rules for Car-Evaluation:

persons=2- --------> evaluation=unacc
safety=low --------> evaluation=unacc
maint=vhigh --------> evaluation=unacc
buying=high --------> evaluation=unacc
safety=low --------> evaluation=unacc
persons=2 --------> evaluation=unacc
buying=low,safety=low --------> evaluation=unacc
buying=vhigh,maint=high --------> evaluation=unacc
buying=vhigh,maint=vhigh --------> evaluation=unacc
buying=vhigh,persons=2 --------> evaluation=unacc

Top 10 Rules for Tic-Tac:

'X6' --------> 'positive'
'X0' --------> 'positive
‘B4' --------> ‘positive'
'O7' --------> 'positive
'O5' --------> 'positive'
'X4' --------> 'positive'
'X2' --------> 'positive
'X8' --------> 'positive'
'O3' --------> 'positive'
'O1' --------> 'positive'

F.) Using Lift:

Tic-Tac (support = 0.05) rules:

'X4' --------> 'o6', 'positive'
'X4' --------> 'positive', 'o4'
'X4' --------> 'positive', 'o3'
'X4' --------> 'positive', 'o2'
'X4' --------> positive', 'o6'
'Negative' --------> 'o4'
'O4' --------> 'negative'
'X4' --------> 'positive', 'o8'
'X4' --------> 'positive', 'o8'
'X4' --------> 'o6', 'positive'

Car-Evaluation(support=0.05) rules:

'Low1'	**-------->**	**'low5'**
'Low5’	**-------->**	**'low1'**
'4dash'	**-------->**	**'high'**
High'	**-------->**	**'4dash**
'4dash'	**-------->**	**'vhigh'**
'Vhigh'	**-------->**	**'4dash'**
'Low5'	**-------->**	**'small'**
'Small'	**-------->**	**'low5’**
'Small'	**-------->**	**'high'**
'High'	**-------->**	**'small'**

Nursery(support=0.05) rules:

'Slightly_prob'	**-------->**	**'recommended'**
'Recommended'	**-------->**	**'slightly_prob',’1’,**
'More'	**-------->**	**'convenient’**
'Nonprob'	**-------->**	**'inconv'**
'Inconv'	**-------->**	**'nonprob'**
'Incomplete'	**-------->**	**'recommended'**
'Recommended'	**-------->**	**'incomplete'**
'Convenient'	**-------->**	**'more'**
'Convenient'	**-------->**	**'1'**
'1'	**-------->**	**convenient',’more’**

Problem: 4

A. An Impossibility Theorem For Clustering

Abstract:

In this paper, Author Jon Kleinberg, who is a professor at Cornell University, pitches the idea that the notion of clustering naturally arises in a number of problem domains, but each domain has it’s own optimality function and implementation techniques. Because of these diverse principles, author argues that it is impossible to develop a unified framework for clustering. Author reasons that clustering function inherently involves trade off between three basic properties, namely scale invariance,richness and consistency. He also introduces a basis for implementation agnostic categorization of clustering function.

Summary:

As per author, Need of clustering arises in different problem domains as a method for grouping objects from heterogeneous groups on the basis of some similarity measure. Clustering function f takes a set s of n objects and reduces into a set of m clusters with the objective of minimizing intra-cluster distance and maximizing inter-cluster distance. Author goes on to make the claim that apart from this objective at the abstract level, there is no unification in the clustering framework when applied on different domains at the level of implementation and concrete methods. Each of the implementation in these domains can come up with distinct results, because of optimality and correctness definition of each domain. Author claims that there has been less work done in the past that reasons about implementation and algorithm agnostic clustering. To support his point, author gives the analogy of research work in mathematical economics domain where in spite of different technical setting, researchers were able to formalize broad framework. Taking the cue from this work, author proposes an axiomatic framework for formalizing clustering. He came up with 3 simple properties for this framework on the basis of which clustering function can be measured. These properties are as follows:

- **Scale-invariance:** As per this property, change in measurement unit should not affect clustering function.
- **Richness:** This property states that all combination of partitions should be achievable.
- **Consistency:** According to this property, if intra-cluster distances are decreased or inter-cluster distances increases, clustering function should not change.

With all the three properties defined, Author defines Impossibility theorem. As per this theorem, there cannot be any clustering function f that satisfies all the three properties. Author extends his support for the impossibility theorem by taking a single linkage cluster and defining the various stops conditions and explains the points where the stopping condition fails to satisfy all the three conditions together. The stopping conditions considered are: (1). K-cluster stopping condition, (2). Distance-r stopping condition and (3) Scale-alpha stopping condition.

The paper strengthen the above drawn conclusion by introducing the concept of antichain. Antichain property states that if a clustering function f satisfies Scale-invariance and consistency then Range(f) is an antichain. To further simplify the concept of antichain, paper presents another term called refinement. A partition Γ’ is called a *refinement* of a partition Γ, if all the set of partition of Γ’ are subset of any of the partition of Γ. Now with refinement term defined, Antichain can also be viewed as a collection of partitions such that there cannot exist two distinct member partition which are refinement of other partition in the collection. In a nutshell, above mentioned antichain theorem implies that richness property cannot be satisfied if the other two properties are satisfied. Author presents an argument that the satisfaction of other two properties leads to formation of set of partition which are in fact antichain and since anti-chain restricts the no. of partitions, it results into conflict with richness property.

Having made the argument about impossibility theorem of clustering theorem, author uses two classes of clustering functions to show that there exist multitude of clustering functions which partially satisfies the impossibility theorem by satisfying two out of three properties. Those two classes are: (1) Single linkage procedure and (2). Centroid based clustering.

In the concluding remarks, the author sheds light on importance of the effect of relaxing the impossibility theorem properties. The relaxation of these properties lead to nearly satisfying the impossibility properties. Author writes about relaxation on consistency property at two levels. In first level, the requirement can be relaxed by allowing formation of substructure, which can be the subset of one of the original cluster. In the next level, author stresses on further relaxation of the requirement to permit having transformation of one partition to be the refinement of other, which leads to the construction of clustering function which not only satisfy scale-invariance and richness, but also meet the partial requirement of consistency.

Critique:

At that it was well thought after paper, as no one had previously tried to develop unified framework for Clustering function. He draws the inspiration from different field and tries to develop broad notion for unification of clustering. He proposes the new framework. He also defines well defined property for this framework. He also gave the mathematical proofs for the axioms developed for the framework. I liked this paper as it was substantially ahead of it's time and was kind of first in it's field. This paper was also first of it's kind to formalize the clustering function.

Author was also naive in this paper in certain sense. He goes on to proof his impossibility theorem by concluding that it was error from the part of clustering function. He failed to assume that it could be his framework drawbacks as well.

But, Overall, it was good paper and I personally liked it for it's unique thought process.

B. Measures of Clustering Quality: A Working Set of Axioms for Clustering

Abstraction:

In this paper, Authors refutes the argument presented by Kleinberg in his paper: "An Impossibility theorem for clustering". In that paper, Kleinberg made claims that given the framework, clustering function cannot satisfy all the properties. But Authors of this paper make an argument that impossibility theorem result were characteristics of kleinberg's framework and not that of clustering function. On advancing upon the work done by Kleinberg's impossibility theorem, authors develops a new consistent unification technique for clustering functions, and in this process, they change the fundamental focus from clustering function to clustering quality measure (CQM). To drive the point home, authors give various examples to demonstrate that unification over clustering function is attainable.

Summary:

Data Mining Assignment#4 Manish Kumar

(kumar20@iu.edu)

Authors start the paper with the discussion of the Kleinberg's work, Impossibility theorem. They disagree with the conclusion drawn by Kleinberg and claim that consistency property defined in his work make the clustering function unsatisfiable. So, to make an improvement, they propose new axioms which serve as axioms for measuring clustering quality measures and cluster overall. They also go on to define the clustering measure quality(CQM). The CQM is defined as a measure that yields the quality(goodness) of the clustering, given the dataset and the clustering methods implemented. Now, they claim under the proposed changes, unification is attainable if we change the focus from clustering function to CQM and by drawing a parallel set of axioms for it. The new revised axioms can be defined as follows:

1. **Scale Invariance:** Change in units of distance measurement should not impact the CQM.
2. **Consistency:** Consistency measure should not be impacted if the distance between intra-cluster points is decreased or the distance between inter-cluster points are increased.
3. **Richness:** As per this property, there always exist a distance function in the universe of set such that CQM is maximized for it.

Authors then go on to develop and define the formalization of CQM. By presenting an example, they convey that for CQM we find the ratio between a point and the centroid of it's own cluster and the centroid of it's nearest cluster. Smaller of the ratio values, better the clustering. Average of these ratio is considered as clustering quality measure for the clustering method.

Author extends their writing by stating different lemmas and defining soundness and completeness for writing. Soundness and completeness is required for the axioms to fit for clustering. Soundness implies that all data set satisfies all the properties defined by axiom and completeness states all properties are implied on each data set which a rules generates. The author discusses these cases because when we talk about clustering, it is an ill defined problem where we do not have any concrete definition of the same. So, they imply that we need to relax on soundness constraint. On Jon's axiom, all three property never satisfy each of them together thus fail to satisfy soundness. Hence we include a property of Isomorphism Invariance that is cluster should be neutral to each data set of three cluster and also deal with the issue of soundness constraint.

At the end of paper, they also moot a new evaluation technique which rely on the number of clusters in order to be computed.These new quality measures have it's origin from common loss function. But it fails on two axioms count. Further, Normalization of common loss function is done in order to correct where it fails. But even normalized common loss function suffers from biases toward purer/rough clustering. Jon's axiom all three property never satisfy each of them together thus fail to satisfy soundness. Hence we include a property of Isomorphism Invariance that is cluster should be neutral to each data set of three cluster. In the final section, author proposes the new definition of refinement preference but it also fail to meet the objective as it could not satisfy the Richness Axiom.

Data Mining **Assignment#4** **Manish Kumar**

(kumar20@iu.edu)

Critique:

This paper came after Jon’s paper and it was already 6 years past. So, it needed really attention to find the drawback’s ih jon’s paper. They conclusively proved that error was not on the part of clustering function but instead it was in his framework. They proposed the new framework in which clustering function was unificable across different domain. They also proposed new axioms and gave the mathematical proof for it. They formalized the parameter well and showed that clustering function can be unified at abstract level. The way they have defined CQM and it’s axiom, I really liked that approach. It was innovative and well thought after.
Unlike Jon’s paper, they focus on broad term and not concentrated on minute level,
I liked this paper in it’s layout style.

h

T