**Project 4 – Clustering**

**CS548 Knowledge Discovery and Data Mining - Fall 2016**

**Prof. Carolina Ruiz**

**Students:** Suchithra Balakrishnan, Mu Niu, Rohitpal Singh

|  |  |  |
| --- | --- | --- |
| **Dataset :**   * Dataset Description * Data Exploration * Initial Data Preprocessing (if any) | **Dataset** | |
| /05  /10  /05 | |
| **Code Description:** At leasttwo Clustering algorithms | **Weka**  /20 | **Python**  /10 |
| **Experiments:**   * Guiding Questions | /10 | |
| K-means - Sufficient & coherent set of experiments | /05 | /05 |
| * Objectives, Parameters, Additional Pre/Post-processing | /05 | /05 |
| * Presentation of results | /05 | /05 |
| * Analysis of individual experiments’ results | /05 | /05 |
| Hierarchical - Sufficient & coherent set of experiments | /05 | /05 |
| * Objectives, Parameters, Additional Pre/Post-processing | /05 | /05 |
| * Presentation of results | /05 | /05 |
| * Analysis of individual experiments’ results | /05 | /05 |
| DBSCAN - Sufficient & coherent set of experiments | N/A | /05 |
| * Objectives, Parameters, Additional Pre/Post-processing | N/A | /05 |
| * Presentation of results | N/A | /05 |
| * Analysis of individual experiments’ results | N/A | /05 |
| Quantitative Analysis of Results and Discussion | /10 | |
| Qualitative Analysis of Results, Discussion, and Visualizations | /20 | |
| Advanced Topic | /30 | |
| Total Written Report Project 4 | /220 = /100 | |

**Dataset Description, Exploration, and Initial Preprocessing: (at most 1 page)**

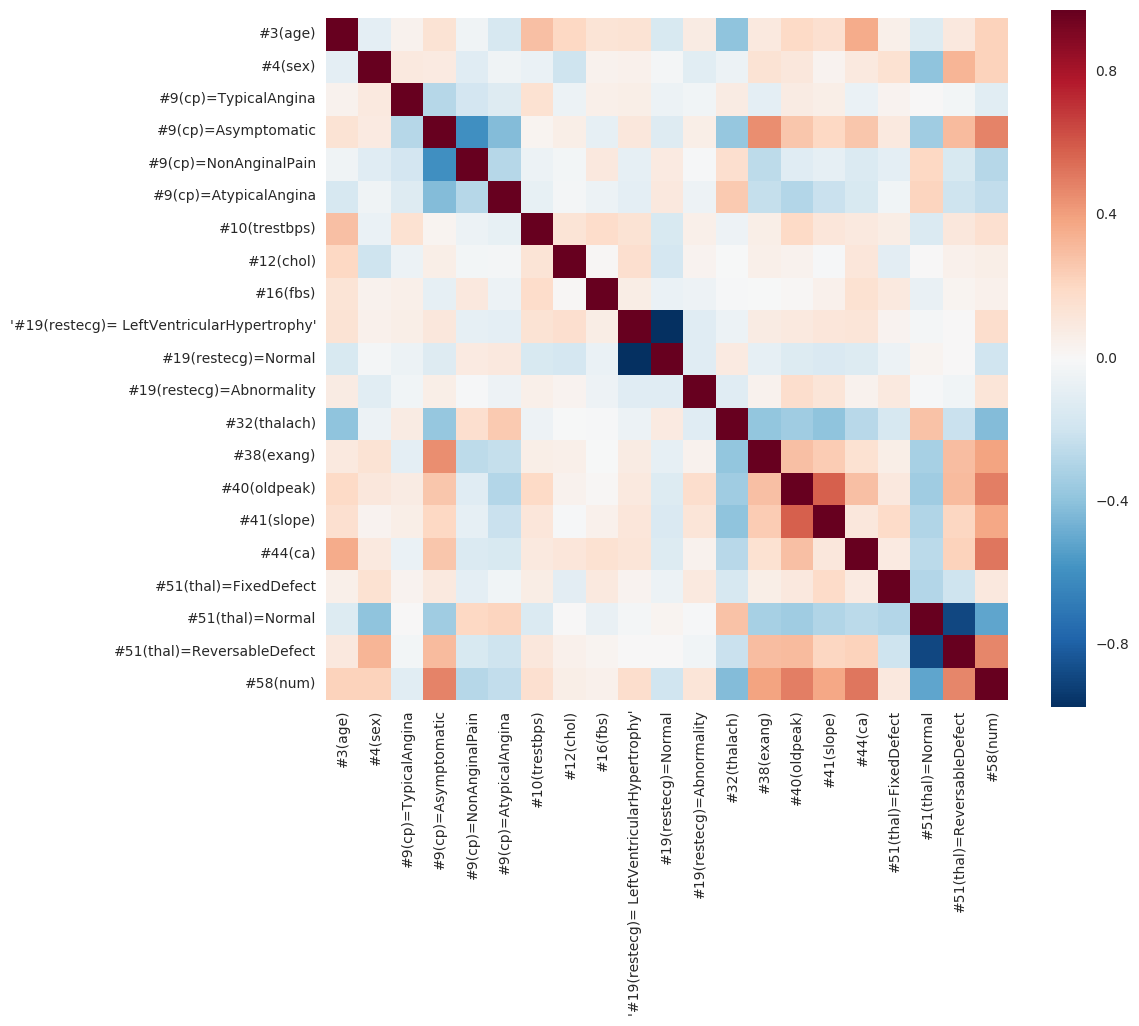
**[05 points] Dataset Description: (e.g., dataset domain, number of instances, number of attributes, distribution of target attribute, % missing values, …)**

This dataset describes some of the mainly causes for heart disease, such as age, resting blood pressure, serum cholesterol and fasting blood sugar.

The target attribute is prediction of whether a particular person have potential risk to have heart disease or not (0 = false, 1,2,3,4 = true in different level). There are in total 303 instances, 14 attributes in the dataset. There are in total 164 instances are false in target attribute. The dataset has 6 missing values in total, 4 in attribute #44(ca) and 2 in attribute #51(thal).

There are binominal attributes like sex, continuous attributes like fbs and age, we all treat them as numerical attributes in our experiments. When we do classes to clusters evaluation, we change #58(num) to be nominal attributes with value 0, 1, 2, 3 and 4.

**[10 points] Data Exploration: (e.g., comments on interesting or salient aspects of the dataset, visualizations, correlation, issues with the data, …)**

Before we calculated the correlation matrix, we changed the three nominal attributes, #9(cp), #19(restecg) and #51(thal) into binominal attributes so that we can form correlation matrix treating all attributes as numerical ones. We found that the four attributes, #9(cp)=Asymptomatic, #40(oldpeak), #44(ca) and #51(thal)=ReversableDefect, are highly positively correlated with the target attribute, two attributes, \*32(thalach) and #51(thal)=Normal are highly negatively correlated with target attributes. Compared to last project’s result, we can easily see how it can be different it can be if we separated specific attributes to reflect heart disease condition.

A problem that is a little bit concern us is that we found some attributes may have significant impact on target attribute if treated as nominal attributes, but they don’t show high correlation with target attribute if considered numerical. This situation typically happened if these attributes are binominal.

**[05 points] Initial data preprocessing, if any, based on data exploration findings: (e.g., removing IDs, strings, necessary dimensionality reduction, …)**

We changed the three nominal attributes into binominal attributes and then normalize whole dataset in order to avoid errors caused by large values . For target attribute, we kept the original values 0, 1, 2, 3 and 4 to represent different level of heart disease. We removed the 6 instances with missing values in attribute #44(ca) and 2 in attribute #51(thal), so that there are in total 297 instances after preprocessing.

We didn’t do any dimensional reduction in initial data preprocessing since the 14 attributes had already been the result of their dimensional reduction from 76 attributes they got. And we didn’t want to lose much information on the initial preprocessing.

**Weka Code Description: Inputs, output, and process followed by Weka’s code for clustering (at most 2/3 page)**

**[10 points] Code Description of the K-means clustering algorithm implementation in Weka:**

**Inputs :** Preprocessed dataset , SimpleKMean as cluster algorithm with default parameters and Target attribute for cluster evaluation.

**Process :** SimpleKMean calls function buildClusters() which takes  instances as an input . Set iteration=0 and classindex= -1 (which means it assumes no class). Using default parameter this algorithm removes all missing values from nominal and numeric attributes by mode and mean from training dataset.

After this it calls function moveCentroid() which returns the mean for numeric attributes and first index of value of nominal attribute which appears the most as original centroids. Now from last instance in the dataset,  it starts adding instances in cluster\_centroids till (number of instances in cluster\_centroids is equal to number of clusters ( passed as a default parameter)) . So in first iteration each individual instance in cluster\_centroidsacts as separate cluster.

Increases the iteration and for each instances in the dataset it calculate the distance using Euclidean distance method (passed as default approach)  from each instances in cluster\_centroids (as clusters). And assign all the instances in the dataset according to shortest Euclidean distance to respective clusters. Again it calculate the centroid for each clusters and repeat the process till (either max number of iteration or centroid of clusters changes). In the end it evaluate the performance of clustering against target attribute and displays the output.

**Output :** Number of iteration= 6, SSE=515.77 (within cluster), final cluster centroids , time taken to build  clusters and incorrectly clustered  instances .

**[10 points] Code Description of the hierarchical clustering algorithm in Weka:**

**Input :**Preprocessed dataset , hierarchicalClusterascluster algorithm with default parameters and Target attribute for cluster evaluation.

**Process :** Hierarchical Clustering calls function buildClusters() which takes  instances as an input. Initially it uses all instances as a separate cluster . Then It checks the value of the parameter ‘Linktype’ if it is not “neighbour joining” then Hierarchical Clustering uses variables nClusters= (#instances in the dataset) , ClusterNodes= (node type of size equal to #instances) and nClusterID (Vector of vectors with size equal to #instances) as an input to call the function doLinkClustering(). In doLinkClustering() , it creates a matrix having Euclidean distance ( as a default method) between each pair of instances and checks the condition if (nClusters < nNumClusters) where nNumClusters is by default 2 as dataset has two class values( 0 and 1). Till this condition is true it starts merging two clusters which are close to each other using Merge(). After performing one merge operation it again updates the distance matrix. This process continues till nClusters becomes equal to nNumClusters. At the end it evaluate the performance of clustering against target attribute and displays the output.

**Output :** List of Clusters , Time taken to build the model, Evaluation using Class values and Incorrectly Clustered instances

**[10 points] Python Packages and Functions used for Clustering. Describe inputs & outputs (at most 1/3 page)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Package** | **Function** | **Input** | **Output** |
| sklearn | cluster | Data, # of clusters, other parameters | cluster label, centroid |
| sklearn | metrics | Cluster label, predicted cluster label | score |
| scipy | Distance, hierarchy | Data, algorithm, eps, min sample | Dendrogram tree |
| seaborn | heatmap | Similarity matrix | Visualization |

**[10 points] Three Guiding Questions about the dataset domain (at most 1/3 page):**

**1. What combination groups exist for people who have heart disease based on age and gender?**

**2. What groups occur with chest pain, fasting blood sugar level and maximum heart rate?**

**3. If a person suffers from severe heart problem, what other conditions exist in such clusters?**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[40 points] Summary of Experiments with Partitional Clustering (k-means).** *At most 1 page.* | | | | | | | | |
|  | **Tool** | **Pre-process** | **# clusters** | **Distance**  **function** | **#**  **ite** | **SSE** | **% of instances**  **per cluster** | **Observations about experiment**  **Observations about visualization**  **Interpretation of centroids**  **Classes to cluster evaluation?** |
| P1 | Weka | Subset selection  (sex, age num(0-4)) | 6 | Manhattan | 5 | 36.87 | 15%,16%,20%,13%,20%, 16% | The SSE decreases as the number of clusters increases and it is also observed that Euclidean distance function produces clusters with lesser SSE compared to Manhattan. In class to cluster evaluation, sex is used as nominal. When used higher number of clusters the classes to cluster evaluation gives higher percentage of incorrectly clustered instances. |
| 8 | Euclidean | 5 | 4.32 | 9%, 23%, 14%, 6%, 12%, 6%,12%, 20% |
| Python | 2 | Euclidean | 300 | 155.6 | 28%,72% |
| 4 | 60.7 | 16%,27%,30%,27% |
| 5 | 39.6 | 17%,24%,11%,22%,26% |
| 6 | 30 | 30%,12%,24%,12%,18%,4% |
| P2 | Weka | Subset selection  (cp(binominal),fbs,thalach) | 6 | Euclidean | 2 | 10 | 9%, 42%, 23%, 15%, 6%, 6% | The attribute fbs has maximum correlation with most attributes. So this experiment was to find the clusters that were formed with fbs values. thal is used as a nominal attributes for classes to cluster evaluation. The percentage of clusters in weka and python are almost equal. |
| Python | 2 | 300 | 135.9 | 48%,52% |
| 4 | 43.8 | 8%,48%,28%,16% |
| 5 | 29.6 | 42%,28%,16%,8%,5% |
| 6 | 16.07 | 42%,22%,16%,6%,5%,8% |
| P3 | Weka | NA | 5 | Manhattan | 7 | 653.8 | 16%,29%,12%,21%, 23% | This experiment was to find the other conditions that exist with several heart problems, that is people with eart diasease at 2, 3. When used 7 cluster, cluster 3, 5 and 6 had severe heart disease. And some of their conditions were cp=Asymptotic, thallch= (135, 142), chol= avg of 250, thal= reversible defect. |
| 7 | Euclidean | 7 | 337.8 | 13%,11%,12%,9%, 25%, 20%, 11% |
| Python | 2 | Euclidean | 300 | 775.9 | 35%,65% |
| 4 | 602 | 24%,20%,28%,28% |
| 5 | 567.1 | 18%,23%,11%,22%,26% |
| 6 | 523.9 | 28%,11%,25%,11%,13%,11% |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **[40 points] Summary of Experiments with Hierarchical Clustering (single link, complete link, average, centroid, Ward).** *At most 1 page.* | | | | | | | |
|  | **Tool** | **Pre-process** | **# clusters** | **Link**  **type** | **Time**  **taken** | **% of instances**  **per cluster** | **Observations about experiment**  **Observations about visualization**  **Classes to cluster evaluation?** |
| P1 | Weka | Subset selection  (sex, age num(0-4)) | 4 | Average | 0.13s | 55%,28%,14%,3% | From the experiment we can see that link type “Ward” gives us more separated clusters in this experiment, so that it can give us more specific information about each specific clusters. We have similar results from weka and python. Results show male in his 40s to 60s suffer the most from heart disease, 5 clusters would be appropriate. |
| 6 | Ward | 0.38s | 11%,10%,18%,19%,24%,17% |
| Python | 2 | Complete | 0.007 s | 74%,26% |
| Average | 0.007 s | 74%,26% |
| 5 | Complete | 0.007 s | 6%,45%,5%,19%,24% |
| Average | 0.008 s | 69%,19%,5%,3%,4% |
| P2 | Weka | Subset selection  (cp(binominal),fbs,thalach) | 5 | Single | 0.15s | 2%,42%,23%,28%,6% | In this experiment, we found complete link type is more suitable for forming clusters. We tried different cluster number and there’s always a huge cluster that contain instances with asymptomatic chest pain, low fbs and median maximum heart rate achieved. |
| 7 | Complete | 0.14s | 2%,42%,23%,21%,6%,6%,2% |
| Python | 5 | Single | 0.15s | 2%,42%,23%,28%,6% |
| 2 | Ward | 0.007 s | 52%,48% |
| Complete | 0.008 s | 52%,48% |
| 5 | Ward | 0.007 s | 28%,16%,42%,8%,5% |
| Complete | 0.008 s | 47%,8%,16%,6%,5%,22% |
| P3 | Weka | NA | 6 | Complete | 0.3s | 11%,34%,38%,5%,10%,3% | Link type ward is also better for this experiment, but due to multidimensional dataset, it gives us less than 0.5 NMI score no matter what. But the big clusters did show that higher heart disease problem are related with high level of ca and slope. |
| 9 | Ward | 0.76s | 7%,11%,17%,13%,15%,8%,8%,15%,7% |
| Python | 2 | Ward | 0.008 s | 50%,50% |
| Average | 0.007 s | 23%,77% |
| 5 | Ward | 0.008 s | 26%,22%,17%,24%,11% |
| Average | 0.008 s | 75.7%,22%,0.3%,1%,1% |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **[40 points] Summary of Experiments with DBSCAN in Python.** *At most 2/3 page.* | | | | | | | |
|  | **Pre-process** | **Epsilon** | **minPts** | **#**  **clusters** | **Time taken** | **% of instances**  **per cluster** | **Observations about experiment**  **Observations about visualization**  **Interpretation of means & std dev**  **Classes to cluster evaluation?** |
| D1 | Subset selection  (sex, age num(0-4)) | 0.05 | 5 | 6 | 0.005 s | 29%,8%,23%,13%,6%,3% | Distance should as small as possible to get separated, meaningful clusters. Big epsilon will make clusters mix up. In our experiment of epsilon=0.05 and min pts=5, only 18% of points are considered noise. We found that male in 50s have the highest heart problem. |
| 0.5 | 9 | 0.006 s | 30%,9%,15%,24%,2%,9%,4%,3%,2% |
| 1 | 1 | 0.005 s | 99% |
| 0.05 | 10 | 6 | 0.005 s | 25%,10%,7%,6%,5%,5% |
| 0.5 | 6 | 0.008 s | 30%,9%,15%,24%,9%,4% |
| 1 | 1 | 0.005 s | 99% |
| D2 | Subset selection  (cp(binominal),fbs,thalach) | 0.05 | 5 | 6 | 0.005 s | 42%,11%,2%,20%,1%,2% | These attributes are clearly separated more than attributes in question 1. There’s a major cluster contains instances have Asymptomatic, low fbs and maximum heart rate achieved to be 108. This cluster contained 42% instances, can be considered as a common situation. |
| 0.5 | 8 | 0.005 s | 2%,42%,22%,15%,5%,6%,6%,2% |
| 1 | 4 | 0.005 s | 8%,48%,28%,16% |
| 0.05 | 10 | 3 | 0.007 s | 39%,11%,19% |
| 0.5 | 6 | 0.006 s | 42%,22%,15%,5%,6%,6% |
| 1 | 4 | 0.005 s | 48%,28%,15%,7% |
| D3 | NA | 0.6 | 5 | 7 | 0.006 s | 3%,3%,5%,3%,3%,2%,2% |
| 0.8 | 14 | 0.006 s | 2%,3%,3%,2%,5%,2%,3%,2%,2%,3%,2%,2%,2%,2% | Results showed that the heart disease level is closely related to attributes like age, gender and cp level. But because of dimensions are too many, we can’t have very clear cluster groups with typical result. In one of our experiment, when radius is 1.25 and minPts is 8, we have 37% noise point and that’s our best estimation. |
| 1.25 | 10 | 0.007 s | 11%,11%,4%,7%,6%,11%,10%,8%,2%,2% |
| 0.6 | 10 | 2 | 0.006 s | 3%,5% |
| 0.8 | 2 | 0.006 s | 3%,5% |
| 1.25 | 8 | 0.007 s | 11%,4%,8%,5%,11%,8%,10%,6% |

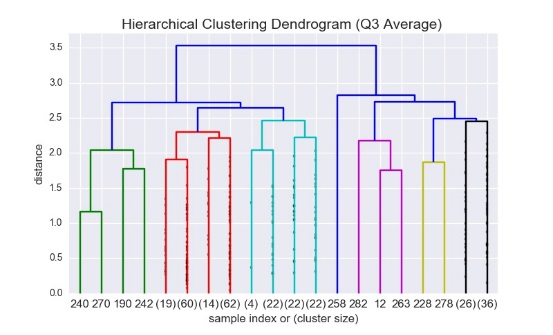
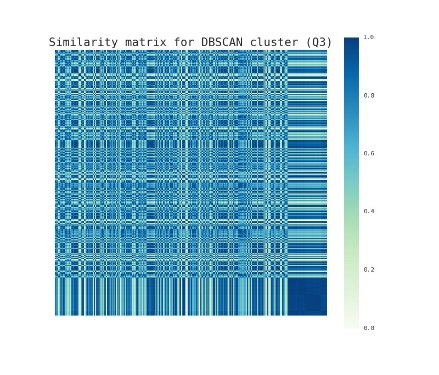
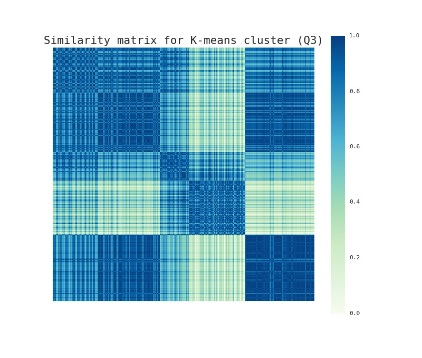
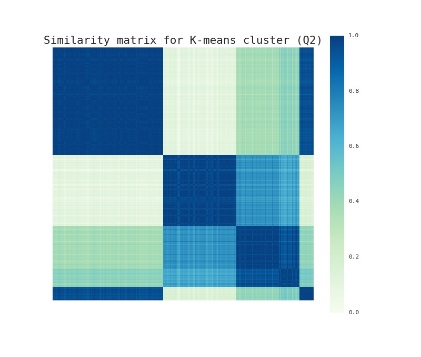
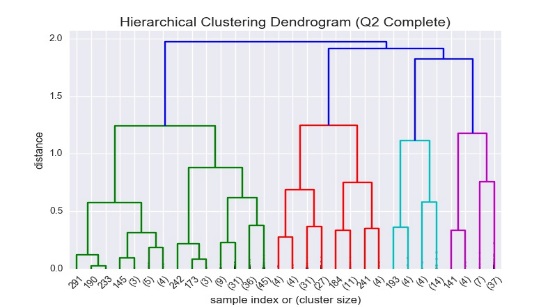
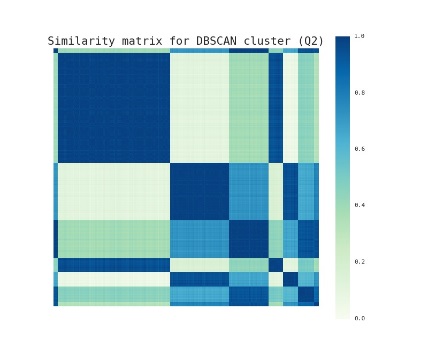
**[10 points] Quantitative Analysis of Weka and Python Results and Discussion (at most 1/3 page)**

In k means if size of clusters is 5 then SSE is quite low in both python and weka.

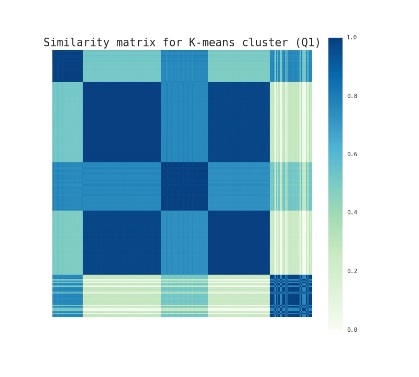
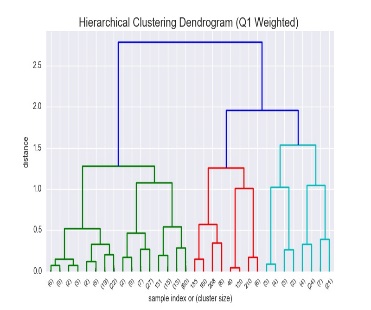
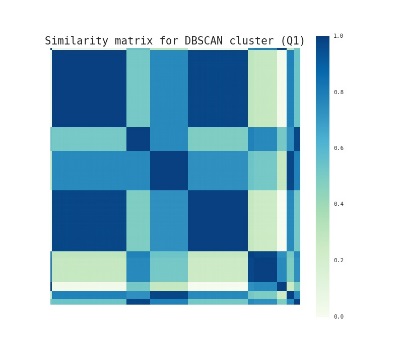
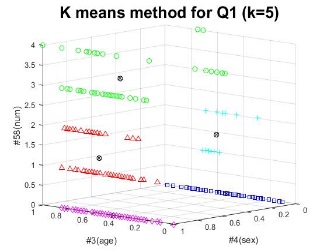
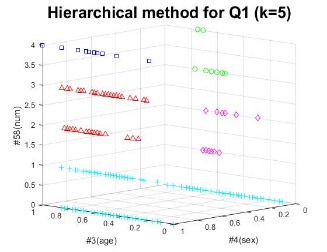
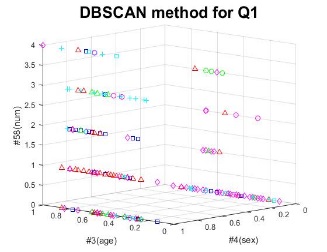
|  |  |  |  |
| --- | --- | --- | --- |
| **K means(5 cluster)** | **NMI** | **V-measure** | **Silhouette Coefficient** |
| **Q1** | 0.84 | 0.837 | 0.62,0.8,0.59,0.76,0.47 |
| **Q2** | 0.13 | 0.13 | 0.8,0.69,0.78,0.66,0.82 |
| **Q3** | 0.5 | 0.494 | 0.06,0.19,0.18,0.18, |

In Hierarchy cluster, % of instances per cluster is also 2% indicates that few clusters are not having all representative data points and unnecessary lead to creation of a particular cluster**.** If Epsilon radius decreases and minpts is large then all data points that must fall in this radius are excluded so lead to many clusters and all are so less separated from other clusters based on silhouette coefficient. **[20 points] Qualitative Analysis of Weka and Python Results on and Visualizations (at most 1 page)**

(Remember also to analyze the results from the point of view of the dataset domain, and discuss the answers that the experiments provided to your guiding questions.)



Q3 Q2



**Q1**

From all these visualization results we can see that when there are fewer attributes to do clusters, we got more clear clusters to view.

For guiding question 1, the biggest group comes from instances that are males at age of fifty, which can be considered as a very dangerous group in having heart disease. Also, females tend to have higher chance to have heart disease as age growth, but it’s slower than male and not as serious.

For guiding question 2, the largest group has the highest thalach value. Cluster results also showed that there is very high possibility that the person could also have non angina pain in his or her record. This could because that if a person’s maximum heart rate is very high, he or she can be a very healthy person so that he or she will not suffer from any kind of angina.

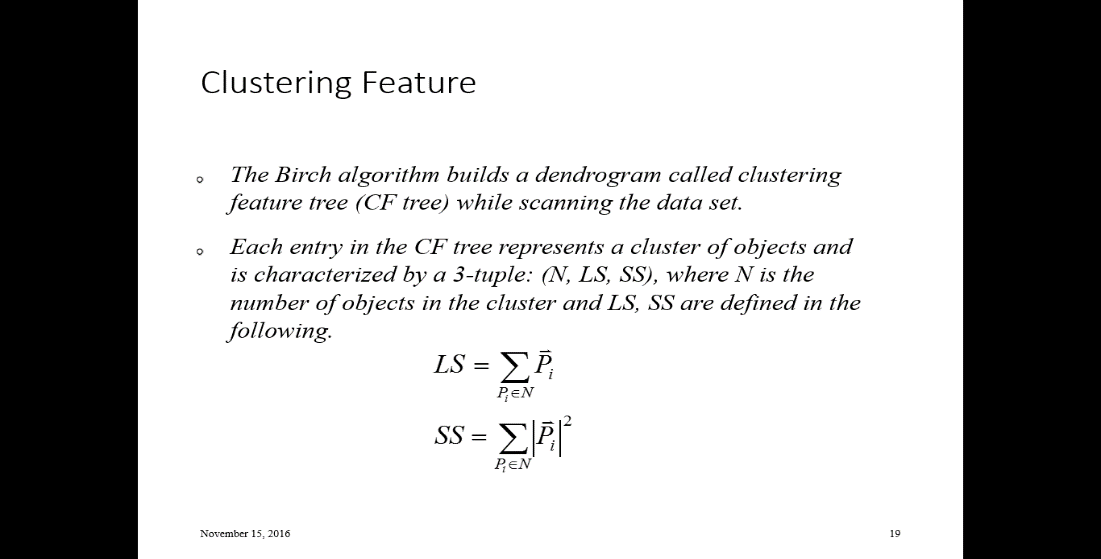
For guiding question 3, although the similarity matrix shows very confusing blocks, we can still get the result that when a person has serious heart problem, he or she will tend to have conditions such as high level of thalch, oldpeak and number of vessel colored, and there is also more chance that this person to be male.

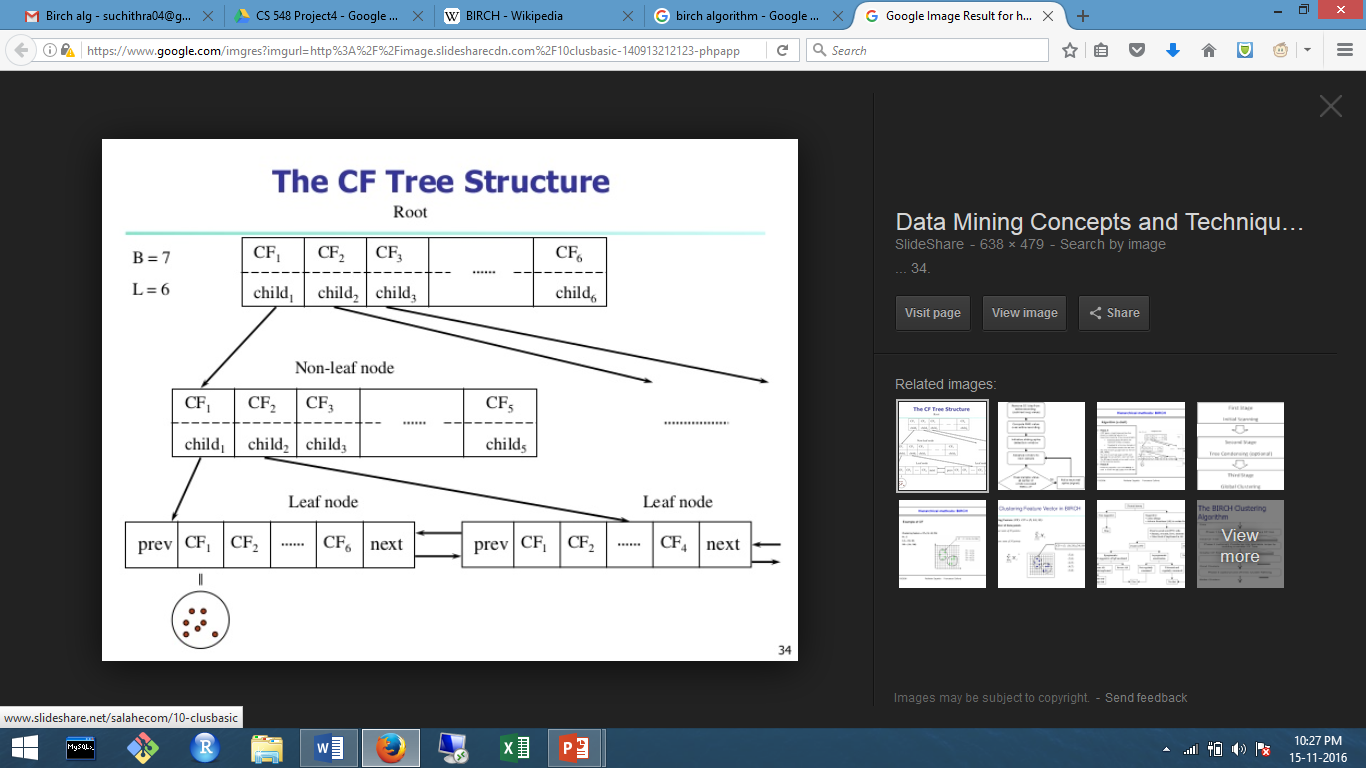
**Advanced Topic: BIRCH – Balanced Iterative Reducing and Clustering using Hierarchies**

**[7 points] List of sources/books/papers used for this topic (include URLs if available):**

* https://en.wikipedia.org/wiki/BIRCH
* *Tian Zhang, Raghu Ramakrishnan, Miron Livny, Data Mining and Knowledge Discovery, Volume 1, Issue 2, 1997, pp. 141-182*
* Davitkov Mireslav, University of Belgrade
* https://www.youtube.com/watch?v=LK\_XKbCOVh0

**[20 points] In your own words, provide an in-depth, yet concise, description of your chosen topic. Make sure to cover all relevant data mining aspects of your topic.**

BIRCH is unsupervised data mining algorithm used to perform hierarchical clustering in particularly large datasets. It is one of the most efficient algorithm to handle noise and to fit large dataset patterns in small memory, while minimizing I/O costs. BIRCH algorithm works on two attributes, one being the clustering feature (CF) and a CF Tree built with several of these CFs. The CF has three parameters, namely, N- number of data points, LS – Linear sum and SS- square sum of data points [**CF= (N, LS, SS)**]. The height of the CF Tree is balanced with two parameters, branching factor B and threshold T. The BIRCH algorithm takes as input n data points and desired number of clusters K.

The BIRCH algorithm consists of four phases. In phase 1, the construction of the CF Tree begins by inserting points based on the initial threshold. If it runs out of memory before it finishes scanning the data, then it increases the threshold value and rebuilds a new smaller tree, by reinserting the leaf entries from the older tree and then other values. So in this phase a good threshold values is necessary.

In phase 2, the leaf entries of the CF Tree are scanned to build smaller CF Tree. Moreover, outliers are removed in this phase and larger clusters are created from several crowded subclusters. This phase is optional.

In phase 3, any existing clustering algorithm is used to cluster the leaf nodes. An agglomerative hierarchical clustering method is applied to the subclusters represented by the CF vectors. In this step the user can provide the desired number of clusters.

Phase 4 is also an optional step. Additional scans are performed in this step and centroids are calculated. Based on the centroids, the points are redistributed. Outliers can also be eliminated in this step, as the points too far from the centroid can be treated as an outlier.

**[3 points] How does this topic relate to clustering?**

BIRCH algorithm performs faster than any existing algorithm on large datasets. It uses agglomerative hierarchical clustering. It is superior to existing clustering algorithms in scalability and stability.

**Authorship:** Although each student on the team is expected to be involved in every aspect of the project, describe in detail here the main contributions that each of the team members made to this project. This authorship description must accurately reflect the work done by each team member, and must be approved by all of the members of the team (at most 1/3 page)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Initial preprocessing** | **K mean code** | **Hierarchical code** | **Python package** | **Clustering** | | | **advanced topic** |
| **K means** | **hierarchical** | **DBSCAN** |
| **Rohitpal Singh** | **√** | **√** | **√** | **√** | **√** |  |  | **√** |
| **Suchithra Balakrishnan** | **√** | **√** | **√** | **√** |  | **√** | **√** | **√** |
| **Mu Niu** | **√** | **√** | **√** | **√** | **√** |  | **√** | **√** |