General Strategy and Clustering Process

The general strategy was to take a first clustering run on the complete flat file dataset and look at the results to decide how to refine the dataset or clustering algorithm. WEKA memory was increased to 4000m by starting WEKA using a command prompt to accommodate the full data set. To improve run times and general ease of use a smaller data set was mainly used for clustering, train-a. Numerical classes where changed to nominal where appropriate. Other clusters where generated by varying options within the clustering algorithm or by changing the clustering algorithm. Output from the clustering was examined for % of instances in each cluster and SSE values. Output was also visualized within WEKA.

Description of Clustering Algorithm

I used the simple K-means clustering algorithm the most. K-means clustering attempts to assign observations to a k number of specified clusters in which each observations belongs to a cluster with nearest mean. I kept the same similarity metric, Euclidean distance, for all clustering runs with a max iteration of 500. Having the max iteration set at 500 always was more than plenty with most clustering runs finishing under a 100 iterations. Out of the 79 classes only 15 were typically selected for clustering to help reduce noise in the data set as noted with the complete run on all 79 classes. Instances that were empty were typically left empty for clustering. K was varied, but in the end reduced to 5 clusters for better visualization. The first fun on the full dataset had 20 clusters which made for an extremely even distribution of instance distribution in the clusters. Another run was done with 16 clusters, number of unique disposition codes, but also was very noisy.

One clustering run was done using the DBscan method, which clusters based on differences in density of the data. Default values were used parameters epsilon and minPoints. A total of 295 clusters were generated using DBscan method which is less than optimal for visualization.

Weka Clustering Output

Scheme: weka.clusterers.SimpleKMeans -M -N 10 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Relation: train-a

Instances: 99999

Attributes: 79

Interval

Patient-Sex

Age

Patient-Race

Patient-Ethnicity

Patient-Disposition

Length of Stay

Admit-Type

Admit-Source

Hospital-ID

Region-ID

Principal-Dx-Code

Admit-Dx-Code

Principal-PR-Code

Severity

Cost Weight

Ignored:

RecordID

PatientID

Record Count

Other-Dx-Code-1

Other-Dx-Code-2

Other-Dx-Code-3

Other-Dx-Code-4

Other-Dx-Code-5

Other-Dx-Code-6

Other-Dx-Code-7

Other-Dx-Code-8

Other-Dx-Code-9

Other-Dx-Code-10

Other-Dx-Code-11

Other-Dx-Code-12

Other-Dx-Code-13

Other-Dx-Code-14

Other-PR-Code-1

Other-PR-Code-2

Other-PR-Code-3

Other-PR-Code-4

Other-PR-Code-5

Other-PR-Code-6

Other-PR-Code-7

Other-PR-Code-8

Other-PR-Code-9

Other-PR-Code-10

Other-PR-Code-11

Other-PR-Code-12

Other-PR-Code-13

Other-PR-Code-14

Cause-E-Code

Place-E-Code

Reimb DRG

Reimb MDC

AccomCharges

AncilCharges

TotalCharges

Serv-Class

Residence-Ind

Emergency-Dept-Ind

CMC1

CMC2

CMC3

CMC4

CMC5

CMC6

CMC7

CMC8

CMC9

CMC10

PL1

PL2

PL3

PL4

PL5

Pot Amb

Complication-Minor

Complication-Sever

Trauma-Minor

Trauma-Severe

Trauma-Severity

Nosocomial Inf

Test mode: evaluate on training

kMeans

======

Number of iterations: 61

Within cluster sum of squared errors: 129327.16416951646

Cluster centroids:

Cluster#

Attribute Full Data 0 1 2 3 4 5 6 7 8 9

(99999) (4796) (6947) (7917) (12056) (13288) (13965) (14856) (10339) (6945) (8890)

==============================================================================================================================================

Interval 135.3515 140.0847 114.2964 152.5772 142.5718 125.4759 135.4721 127.7603 146.7944 143.2531 134.3425

Patient-Sex F F M F F F M M M F F

Age 72.7984 77.0357 66.9703 75.0945 74.4055 72.8775 71.5985 70.9024 72.2965 75.0462 74.6048

Patient-Race 13.2684 4.5367 75.2595 2.4084 11.5896 12.7587 4.6396 5.634 1.6995 28.4903 10.1223

Patient-Ethnicity 2.5291 2.4802 3.8096 2.221 2.063 1.976 2.3567 1.9749 2.5581 5.5078 2.1238

Patient-Disposition 5.0899 6.2965 4.1393 4.9924 5.4603 4.9692 5.1779 4.6392 4.8871 5.8472 5.2057

Length of Stay 7.8484 8.553 8.0557 7.1646 7.5804 8.6729 7.5702 8.0387 7.005 6.922 8.87

Admit-Type 1.296 1.2262 1.3417 1.3501 1.294 1.1455 1.3292 1.2241 1.4374 1.3754 1.3188

Admit-Source 5.7777 6.21 5.6801 5.9604 5.7789 6.064 5.6215 5.8927 5.4472 5.5696 5.6288

Hospital-ID 108.1512 154.9024 60.7406 188.3965 89.5349 49.0014 140.2579 34.289 174.7732 158.1126 118.6579

Region-ID C C F L H M C D K I C

Principal-Dx-Code 26999.5876 25277.9222 27969.0495 31609.6986 37367.8102 23745.0846 35784.9547 21371.0752 14820.0659 28583.6102 22410.7796

Admit-Dx-Code 28319.7758 25009.478 31179.1631 30721.4077 51454.1871 20321.5834 43256.7934 17059.6119 10556.4597 31226.4978 20073.5394

Principal-PR-Code 6688.8147 9190.8968 6732.693 6494.3904 6710.3049 8298.7397 6332.2855 5816.1887 6213.6322 6890.536 3239.6019

Severity 3.8595 4.1576 3.6472 4.0493 3.8438 3.9174 3.7436 3.7993 3.9229 3.8749 3.8278

Cost Weight 2.173 2.215 2.1338 2.2215 2.1214 2.0956 2.1914 2.1651 2.1924 2.1592 2.2963

Clustered Instances

0 4796 ( 5%)

1 6947 ( 7%)

2 7917 ( 8%)

3 12056 ( 12%)

4 13288 ( 13%)

5 13965 ( 14%)

6 14856 ( 15%)

7 10339 ( 10%)

8 6945 ( 7%)

9 8890 ( 9%)

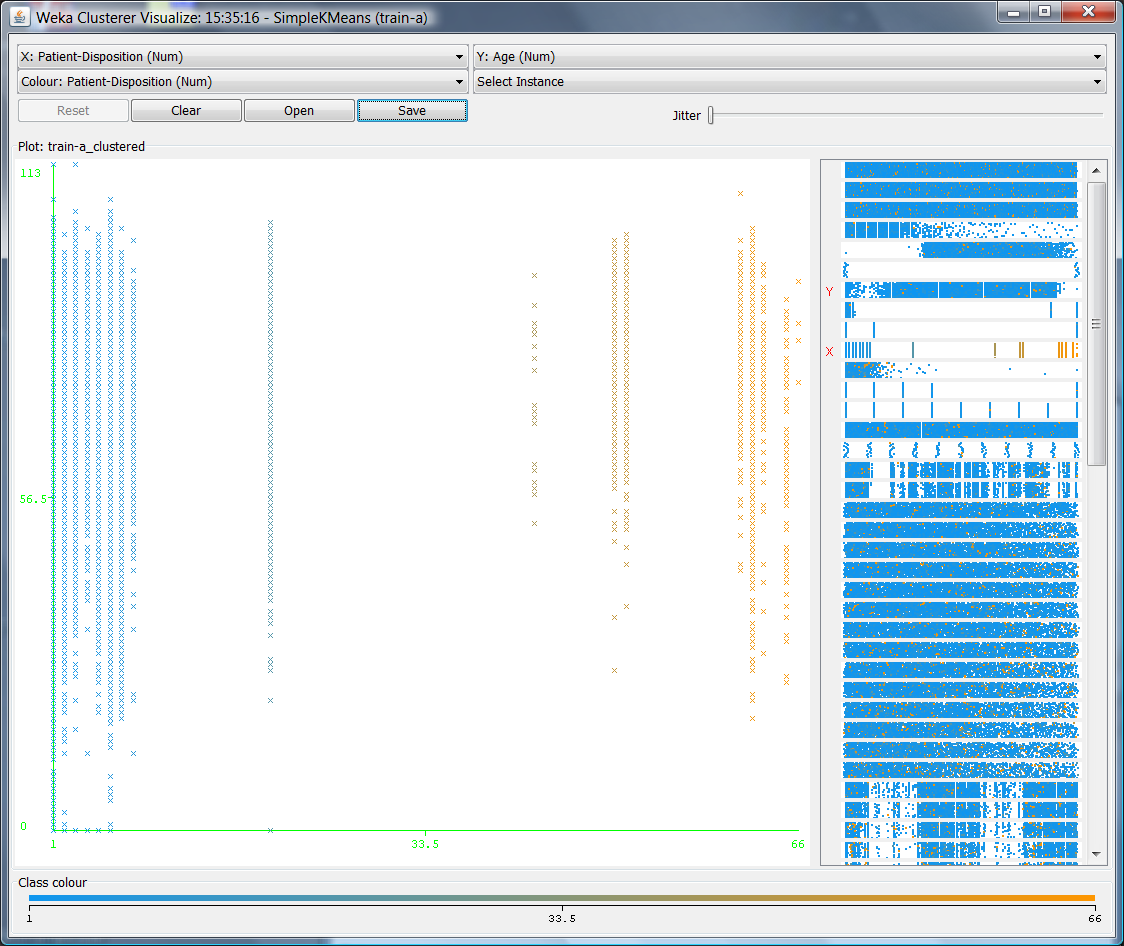


Figure 1 Visualization of clusters with patient disposition on the x-axis and age on y, clusters colored by patient disposition.

Clustered DataObjects: 99999

Number of attributes: 44

Epsilon: 0.9; minPoints: 6

Index: weka.clusterers.forOPTICSAndDBScan.Databases.SequentialDatabase

Distance-type: weka.clusterers.forOPTICSAndDBScan.DataObjects.EuclidianDataObject

Number of generated clusters: 295

Elapsed time: 26678.35



Figure 2 Visualization of the DBscan output with Patient Disposition on the X axis and Interval on the Y axis.

Scheme: weka.clusterers.SimpleKMeans -M -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Relation: train-a

Instances: 99999

Attributes: 79

Interval

Patient-Sex

Age

Patient-Race

Patient-Ethnicity

Patient-Disposition

Length of Stay

Principal-Dx-Code

Admit-Dx-Code

Principal-PR-Code

Severity

Cost Weight

Ignored:

RecordID

PatientID

Record Count

Admit-Type

Admit-Source

Hospital-ID

Region-ID

Other-Dx-Code-1

Other-Dx-Code-2

Other-Dx-Code-3

Other-Dx-Code-4

Other-Dx-Code-5

Other-Dx-Code-6

Other-Dx-Code-7

Other-Dx-Code-8

Other-Dx-Code-9

Other-Dx-Code-10

Other-Dx-Code-11

Other-Dx-Code-12

Other-Dx-Code-13

Other-Dx-Code-14

Other-PR-Code-1

Other-PR-Code-2

Other-PR-Code-3

Other-PR-Code-4

Other-PR-Code-5

Other-PR-Code-6

Other-PR-Code-7

Other-PR-Code-8

Other-PR-Code-9

Other-PR-Code-10

Other-PR-Code-11

Other-PR-Code-12

Other-PR-Code-13

Other-PR-Code-14

Cause-E-Code

Place-E-Code

Reimb DRG

Reimb MDC

AccomCharges

AncilCharges

TotalCharges

Serv-Class

Residence-Ind

Emergency-Dept-Ind

CMC1

CMC2

CMC3

CMC4

CMC5

CMC6

CMC7

CMC8

CMC9

CMC10

PL1

PL2

PL3

PL4

PL5

Pot Amb

Complication-Minor

Complication-Sever

Trauma-Minor

Trauma-Severe

Trauma-Severity

Nosocomial Inf

Test mode: evaluate on training data

=== Model and evaluation on training set ===

kMeans

======

Number of iterations: 16

Within cluster sum of squared errors: 62946.410136152714

Cluster centroids:

Cluster#

Attribute Full Data 0 1 2 3 4

(99999) (22342) (43713) (16625) (9156) (8163)

=======================================================================================

Interval 135.3515 136.1489 132.8607 143.9729 121.1814 145.1991

Patient-Sex F F M F F F

Age 72.7984 76.06 70.5736 74.5248 70.693 74.6302

Patient-Race 13.2684 1.2658 9.5916 1.2992 92.4091 1.4174

Patient-Ethnicity 2.5291 2.331 2.2124 2.303 5.0236 2.4293

Patient-Disposition 5.0899 5.121 4.9236 4.9867 4.5959 6.6596

Length of Stay 7.8484 8.1268 7.6887 7.9166 7.5331 8.1564

Principal-Dx-Code 26999.5876 8637.2772 26434.047 39359.0078 28015.4926 53994.3844

Admit-Dx-Code 28319.7758 5694.3418 27684 42327.0424 29818.8059 63464.7093

Principal-PR-Code 6688.8147 7764.2755 6492.3584 3255.7033 6886.7672 8699.8531

Severity 3.8595 3.9693 3.8172 3.7506 3.76 4.1193

Cost Weight 2.173 2.1137 2.193 2.1887 2.1431 2.2305

Clustered Instances

0 22342 ( 22%)

1 43713 ( 44%)

2 16625 ( 17%)

3 9156 ( 9%)

4 8163 ( 8%)

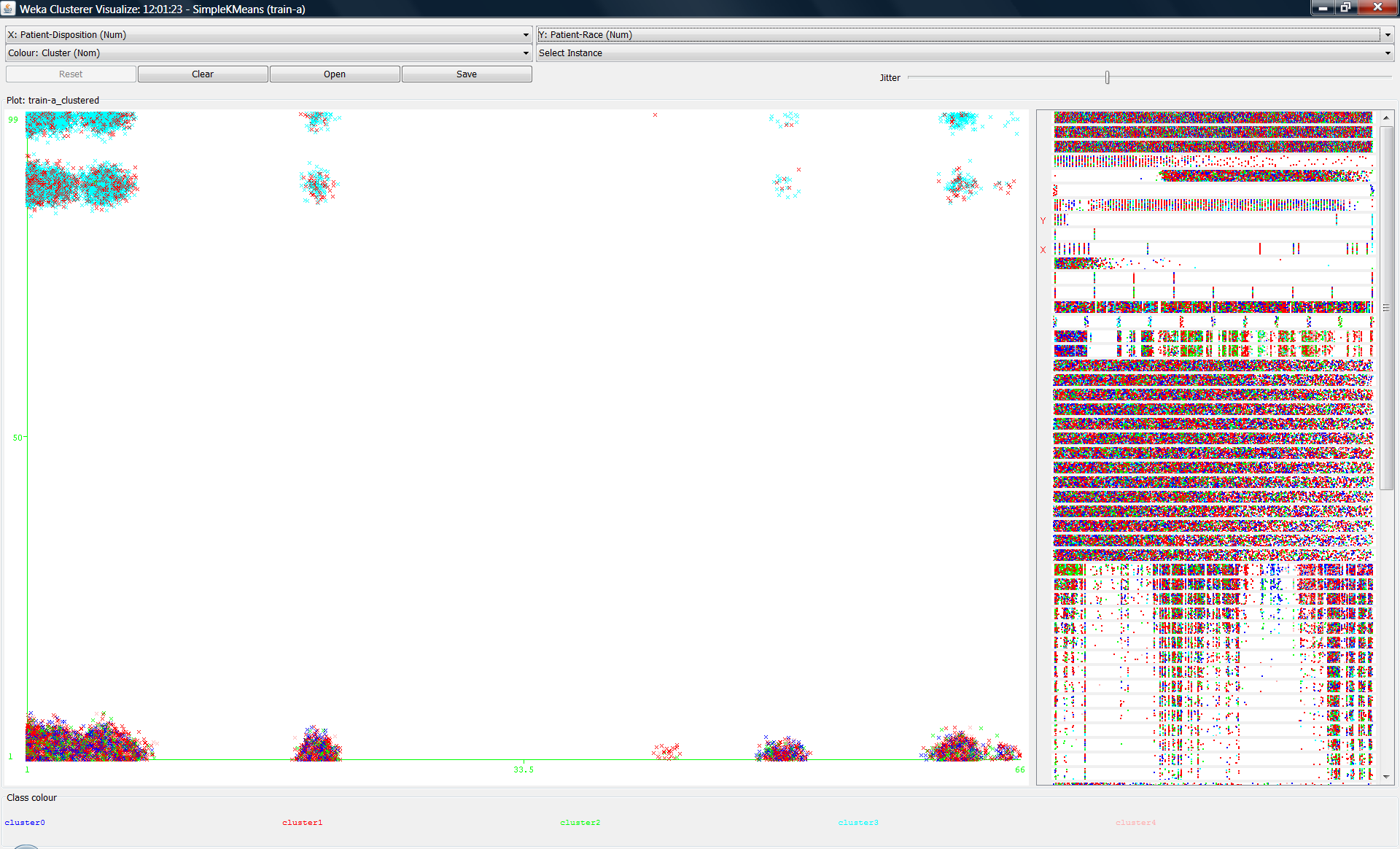


Figure 3 Visualization of the above K-means 5 cluster run with patient disposition on the X axis and Race on the Y axis, colored by cluster.

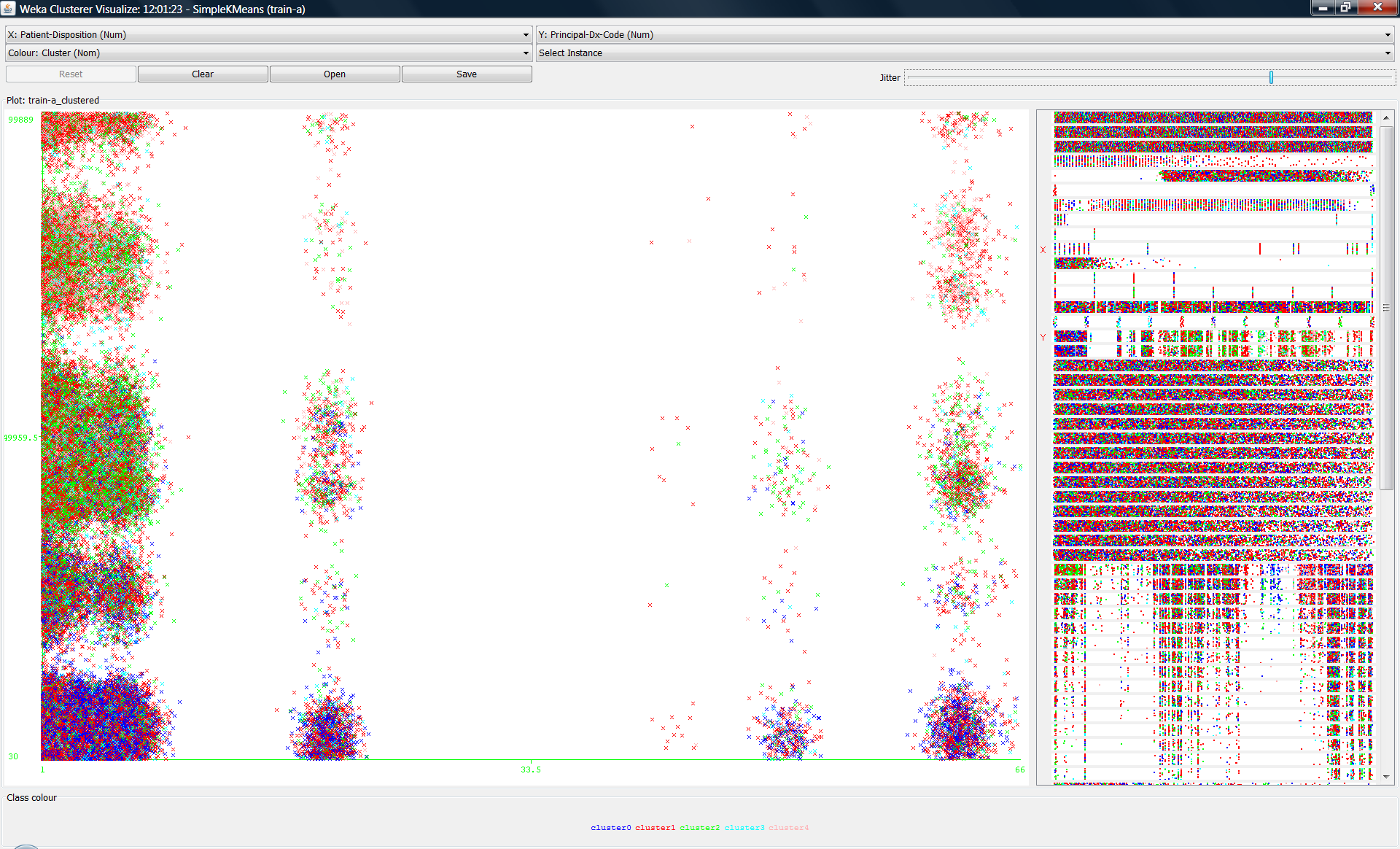


Figure 4. Visualization of the above K-means 5 cluster run with patient disposition on the X axis and Principle DX code on the Y axis, colored by cluster.

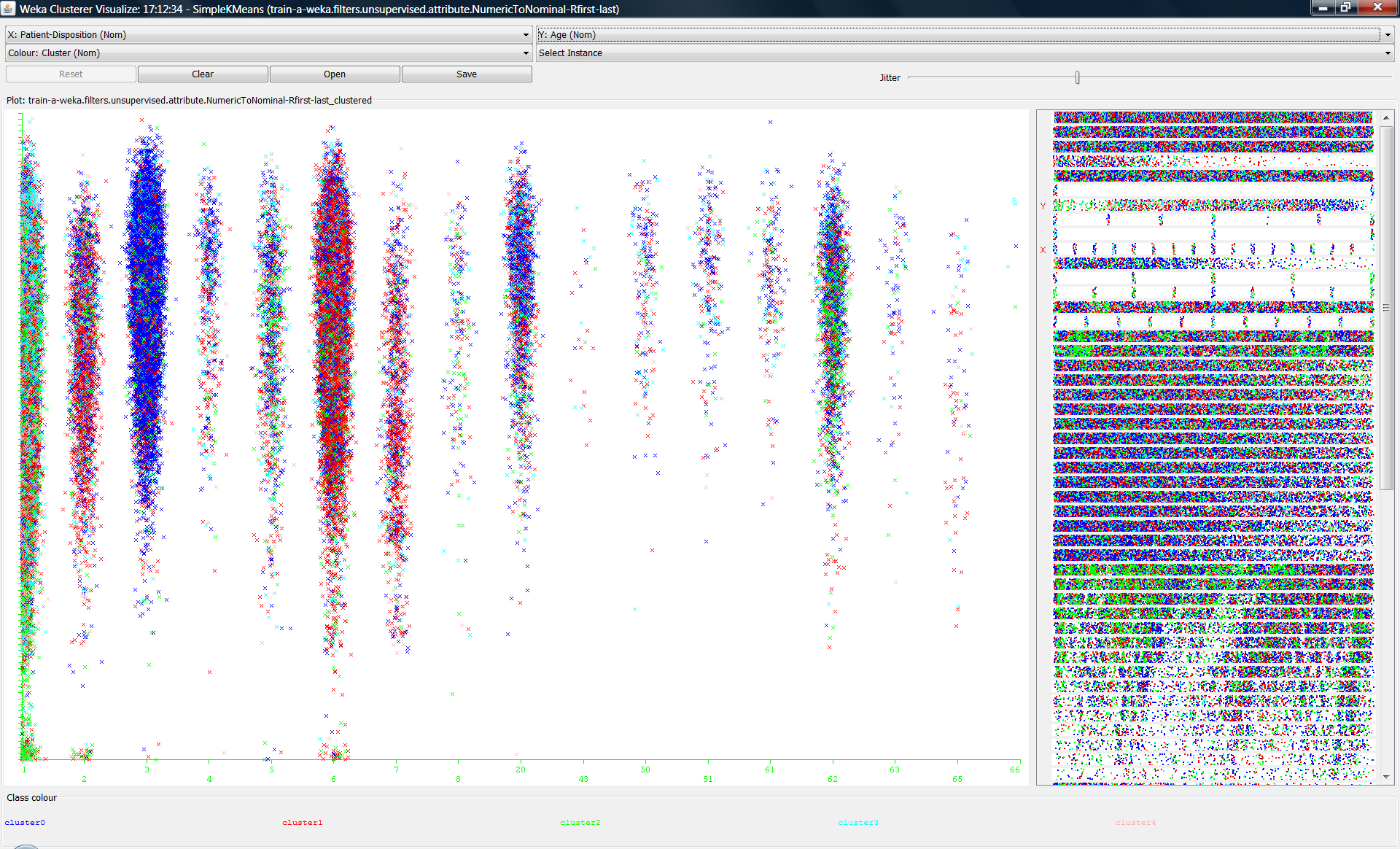


Figure 5. Visualization of the above K-means 5 cluster run with patient disposition on the X axis and Age on the Y axis, colored by cluster.

Discovered Clusters

The data set was very noisy resulting in undefined clusters, which made for rather unimpressive visualizations. There are some perceptual clusters generated with K set at 5 comparing patient disposition to Race, Principle Dx code, and Age as displayed in figures 3, 4 and 5 respectively. In figure 3, cluster 3 stand out as having a higher set of race codes. Cluster 0 in the figure 4 stands out has having a general lower Principle Dx code. In figure 4, patient disposition code 3 is predominate in cluster 0 with and age comparison.