EM Cluster run on Breast-Cancer Data:

=== Run information ===

Scheme:weka.clusterers.EM -I 100 -N -1 -M 1.0E-6 -S 100

Relation: breast-cancer

Instances: 286

Attributes: 10

age

menopause

tumor-size

inv-nodes

node-caps

deg-malig

breast

breast-quad

irradiat

Class

Test mode:evaluate on training data

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

EM

==

Number of clusters selected by cross validation: 3

Cluster

Attribute 0 1 2

(0.36) (0.4) (0.24)

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

age

10-19 1 1 1

20-29 1.0045 1.991 1.0045

30-39 1.5147 25.3553 12.1299

40-49 4.2193 66.4737 22.307

50-59 48.3473 25.4929 25.1598

60-69 46.933 1.0403 12.0267

70-79 6.9597 1.0219 1.0184

80-89 1 1 1

90-99 1 1 1

[total] 111.9785 124.3751 76.6463

menopause

lt40 7.1309 1.8604 1.0087

ge40 96.909 3.3298 31.7612

premeno 1.9386 113.1849 37.8765

[total] 105.9785 118.3751 70.6463

tumor-size

0-4 4.96 5.0364 1.0036

5-9 2.9757 3.0213 1.003

10-14 14.137 14.852 2.011

15-19 17.2442 11.4875 4.2683

20-24 20.0289 21.3245 11.6465

25-29 14.7357 29.3171 12.9473

30-34 18.5737 20.0607 24.3656

35-39 5.6307 8.0983 8.271

40-44 8.6745 8.1057 8.2198

45-49 2.0343 1.9869 1.9788

50-54 4.9838 3.0847 2.9315

55-59 1 1 1

[total] 114.9785 127.3751 79.6463

inv-nodes

0-2 97.3074 108.7517 9.9409

3-5 5.1672 6.2416 27.5912

6-8 1.0946 2.5147 16.3907

9-11 2.6243 1.0988 9.2769

12-14 1.0005 1.7551 3.2444

15-17 1.7803 1.0133 6.2063

18-20 1 1 1

21-23 1 1 1

24-26 1.0041 1 1.9959

27-29 1 1 1

30-32 1 1 1

33-35 1 1 1

36-39 1 1 1

[total] 115.9785 128.3751 80.6463

node-caps

yes 3.0105 2.566 53.4235

no 101.968 114.8091 16.2228

[total] 104.9785 117.3751 69.6463

deg-malig

1 38.9172 33.9963 1.0865

2 39.5245 64.6194 28.8562

3 27.5369 19.7595 40.7037

[total] 105.9785 118.3751 70.6463

breast

left 56.7788 62.1931 36.0281

right 48.1997 55.182 33.6182

[total] 104.9785 117.3751 69.6463

breast-quad

left\_up 42.386 34.7548 22.8593

left\_low 39.6158 45.9394 28.4448

right\_up 10.4952 13.9286 11.5761

right\_low 4.7102 15.7882 6.5016

central 10.7714 9.9641 3.2646

[total] 107.9785 120.3751 72.6463

irradiat

yes 14.4847 18.5795 37.9357

no 90.4938 98.7956 31.7106

[total] 104.9785 117.3751 69.6463

Class

no-recurrence-events 86.87 89.3144 27.8156

recurrence-events 18.1085 28.0608 41.8307

[total] 104.9785 117.3751 69.6463

Time taken to build model (full training data) : 4.76 seconds

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

Clustered Instances

0 104 ( 36%)

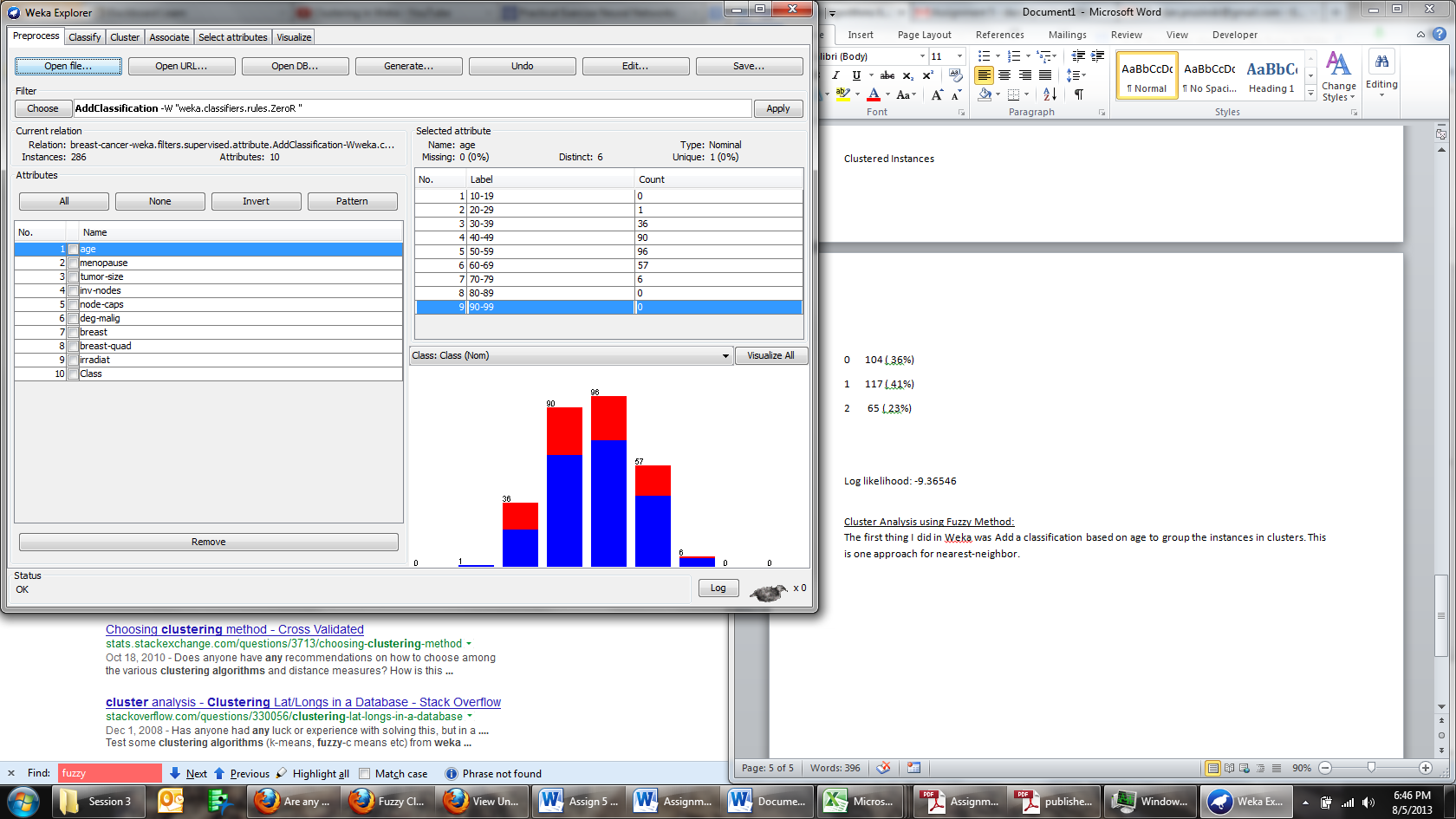
1 117 ( 41%)

2 65 ( 23%)

Log likelihood: -9.36546

Cluster Analysis using Fuzzy Method:

The first thing I did in Weka was Add a classification based on age to group the instances in clusters. This is one approach for nearest-neighbor. The output is below:



I then ran the Select Attributes CfsSubsetEval, which generated the result below:

=== Run information ===

Evaluator: weka.attributeSelection.CfsSubsetEval

Search:weka.attributeSelection.BestFirst -D 1 -N 5

Relation: breast-cancer-weka.filters.supervised.attribute.AddClassification-Wweka.classifiers.rules.ZeroR

Instances: 286

Attributes: 10

age

menopause

tumor-size

inv-nodes

node-caps

deg-malig

breast

breast-quad

irradiat

Class

Evaluation mode:evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:

Best first.

Start set: no attributes

Search direction: forward

Stale search after 5 node expansions

Total number of subsets evaluated: 47

Merit of best subset found: 0.097

Attribute Subset Evaluator (supervised, Class (nominal): 10 Class):

CFS Subset Evaluator

Including locally predictive attributes

Selected attributes: 3,4,5,6,9 : 5

tumor-size

inv-nodes

node-caps

deg-malig

irradiat

The merit of best subset found: 0.097, which I am assuming is the density.

EM utilized 3 clusters based on age, which is similar to the initial analysis of the grouping based on age.