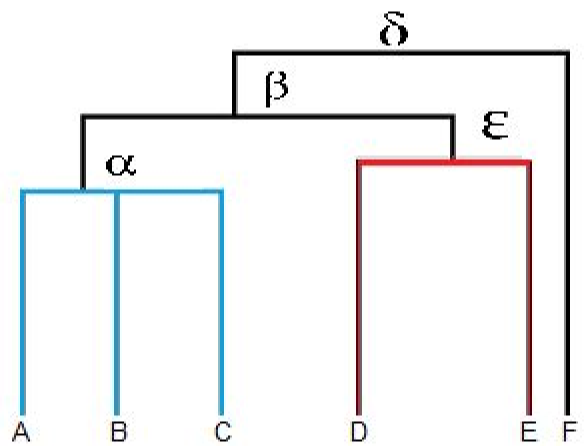
John M. Warlop Assignment #6 Due 11/20/2017

1 Describe how hierarchical clustering method works



In general, it works as follows:

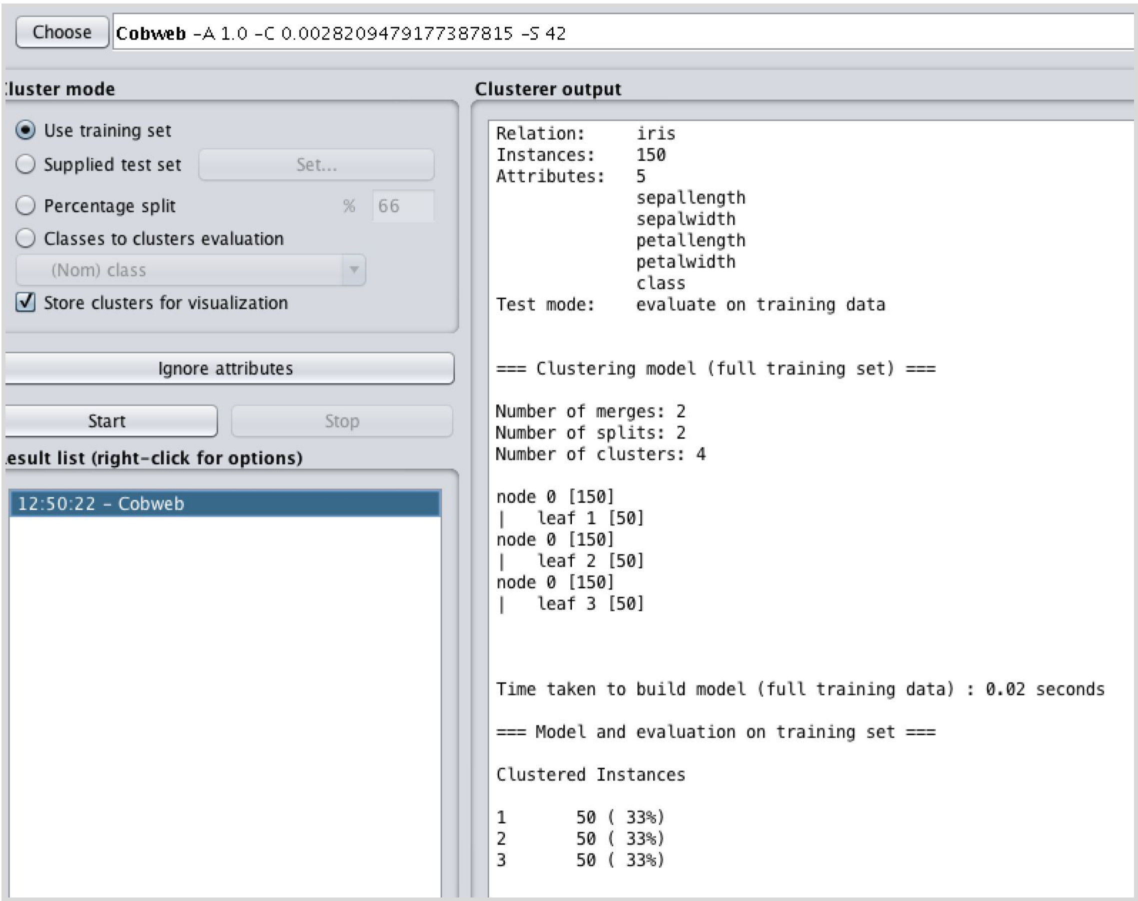
Given N data points(each data point is a cluster), find the two data points that are closest together. Now compute the centroid for these two data points, these two data points now represent a new cluster with data centroid. Now find next to nearest points, this could be two totally new points, or include the prior centroid.

You continue to merge clusters until you only have two clusters. You could stop at one huge cluster, but this is not very usefull.

To visualize this process, look at image to left. This is called a dendrogram. If this image represented the process of hierarchical clustering, the image would be interpreted as follows:
Datum A,B and C merged into one cluster alpha, Datums D and E merged into cluster epsilon and then alpha and epsilon clusters merged into beta -- leaving just clusters beta and F(F is its own cluster). If merged into one cluster, delta would be your final cluster.

Fundamentals of Data Mining

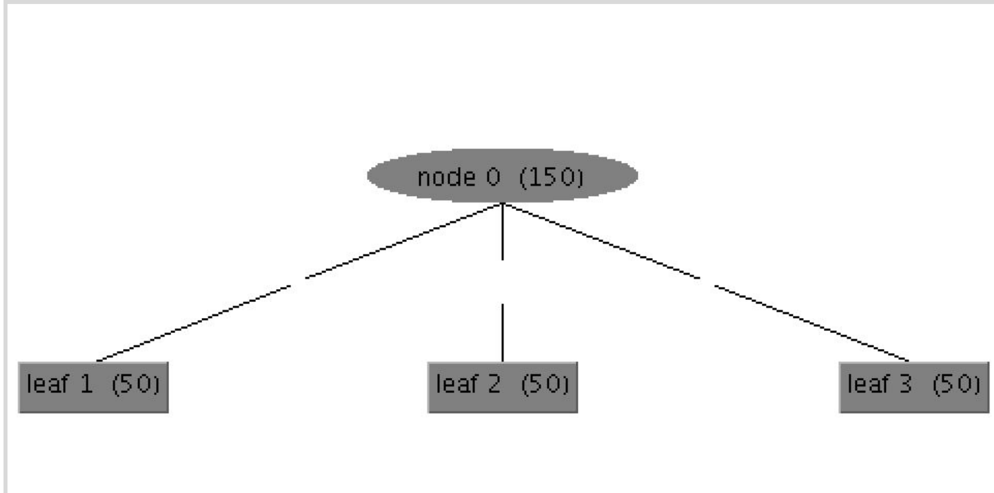
2 Produce a hierarchical clustering(COBWEB) model for iris data. How many clusters did it produce? Why? Did you expect that outcome?



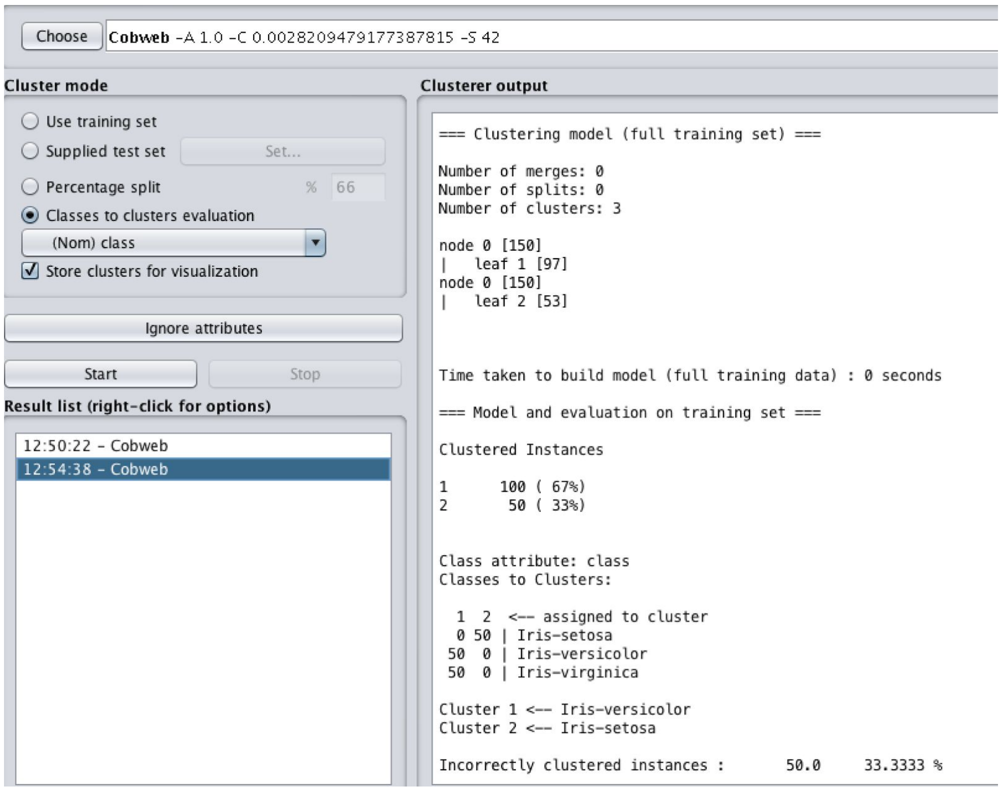
Input

Output

There are three classes, thus I'd expect 3 clusters. Running Cobweb with default settings, I get 4 clusters. Yet, I get 4 cluster instances. Don't know exactly the difference between a cluster and a cluster instance though.

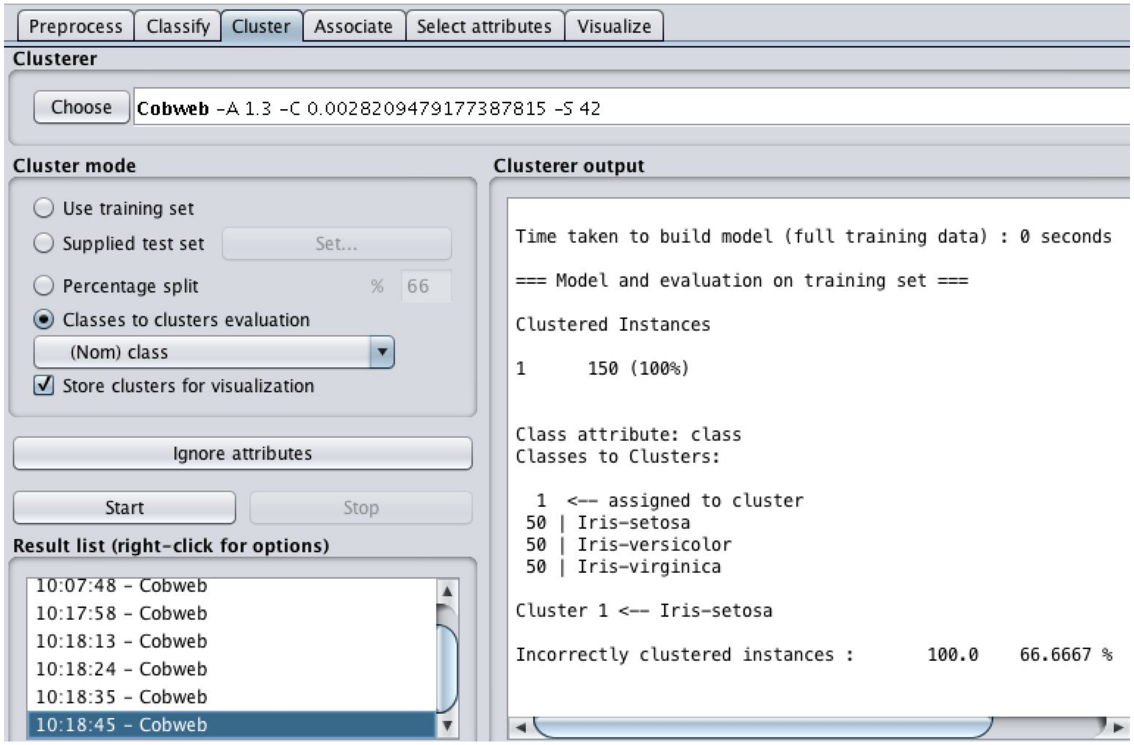

node 0 (150)
leaf 1 (50)
leaf 2 (50)
leaf 3 (50)

Change to use the classes to cluster evaluation. What can you conclude from it?



Output
Weka builds the clustering first and then during testing assigns classes to clusters. After the first step, Weka(using Cobweb with classes to clusters) created only two clusters. Yet, as we know the iris data has 3 classes, so the Iris-setosa class is considered an error.

Use the acuity and cutoff parameters in order to produce a model that clusters major Iris types together. What values of parameters worked the best? Examine your findings/understanding of the produced results.



I slowly increased acuity until I got each major Iris type into their own cluster. When acuity hit 1.3, I got the desired result. I did not change cutoff. The acuity is the minimum standard deviation between clusters. Thus, when standard deviation get up to 1.3, the classes are all in one cluster. Since there is only one cluster, 2/3(66.67%) are not in correct cluster.

3 Describe how EM clustering methods works.

The EM algorithm assigns each instance to a cluster, but unlike K-means, there is a probability distribution as to wheather or not a instance belongs in one cluster or the other. Thus, this means a data point could belong to more than one cluster.

The EM algorithm starts by w number of clusters and finds mean, and covariance and size of cluster and next goes to the E-step("Expectation"). In this step, for each data point, the probability that it belongs in a certain cluster is determined. Next with the M-step("Maximization") the parameters for each cluster(mean, covariance and size) are re-calculated using the new weighted parameters.

For each iteration of the EM algorithm, the log-likelihood get better(increases) and it can be proven that this algorithm does converge(log liklihood not changing), however it is not gauranteed that you will hit a global maximum, you may have hit a local optimum, thus usually it is best to try EM with various starting parameters.

4 Use the EM clustering method on either the basketball or the cloud data set. How many clusters did the algorithm decide to make? Describe the model produced. If you change from “Use Training set“ to “Percentage evaluation split – 66% train and 33% test” - how does the evaluation change? Discuss your findings/understanding of the produced results with respect to the specific dataset.

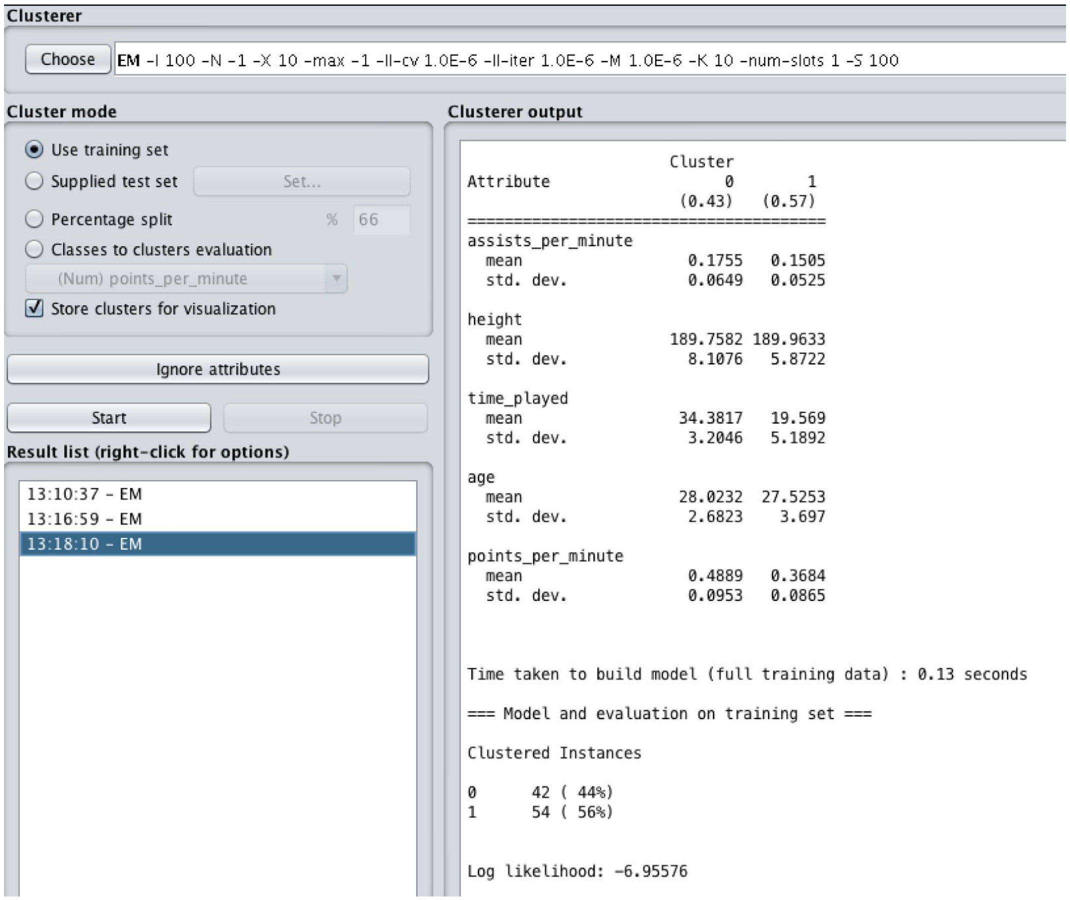


Image 1

Input
Output

Weka/EM produced 2 clusters. By default the numberOfClusters parameter is set to -1, which means EM tries to determine the best number of clusters.

It should be noted that using training set data for training and testing leads to overfitting or too good of an estimate. So even though the log liklihood is better with using training set, the split is more realistic.

I thought the most interesting look at the data was when I looked at # of points per minute and # of assist per minute.

When I clicked on many red and blue data points. The blue points are better because at this represents players that score more or assist in scoring.

I looked at the other

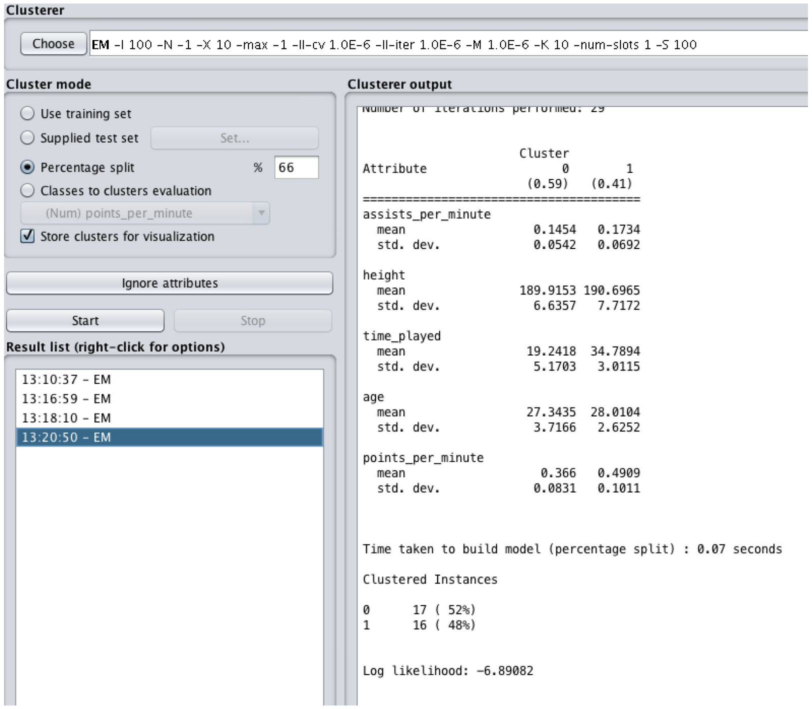
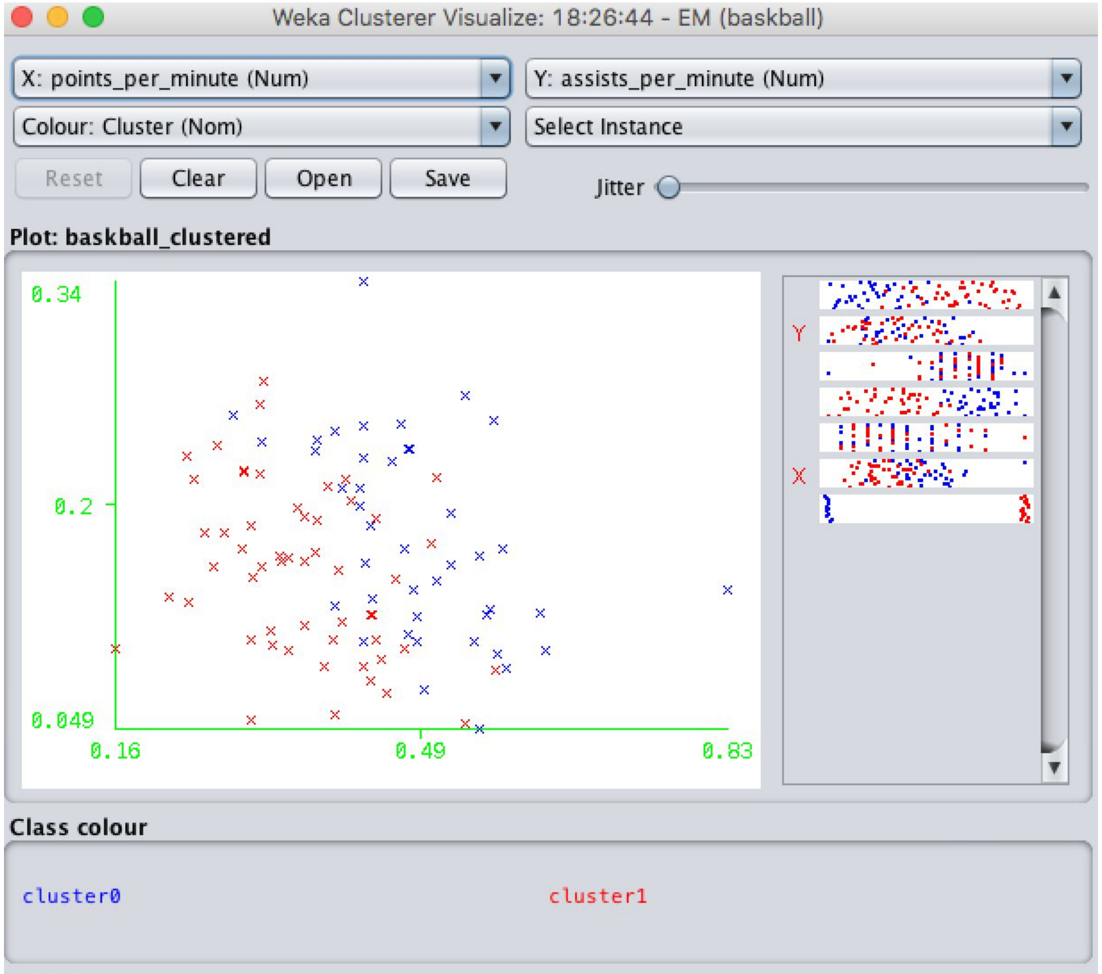


Image 2

attributes(age, height, etc) and I found the biggest association to be players around 30 years old are the most productive.

If players are to old or too young, they are not as productive.

This goes along with what I’ve always heard about athletes in that they are most productive around 30 years old.



Weka Clusterer Visualize: 18:26:44 - EM (baskball)
X: points_per_minute (Num)
Y: assists_per_minute (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: baskball_clustered
0.34
0.2
0.049
0.16
0.49
0.83
Y
X
Class colour
cluster0
cluster1

Image 3

Appendix

Home

Iris.ARFF

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
% 3. Past Usage:
%    - Publications: too many to mention!!!  Here are a few.
%    1. Fisher,R.A. "The use of multiple measurements in taxonomic problems"
%       Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions
%       to Mathematical Statistics" (John Wiley, NY, 1950).
%    2. Duda,R.O., & Hart,P.E. (1973) Pattern Classification and Scene Analysis.
%       (Q327.D83) John Wiley & Sons.  ISBN 0-471-22361-1.  See page 218.
%    3. Dasarathy, B.V. (1980) "Nosing Around the Neighborhood: A New System
%       Structure and Classification Rule for Recognition in Partially Exposed
%       Environments".  IEEE Transactions on Pattern Analysis and Machine
%       Intelligence, Vol. PAMI-2, No. 1, 67-71.
%       -- Results:
%          -- very low misclassification rates (0% for the setosa class)
%    4. Gates, G.W. (1972) "The Reduced Nearest Neighbor Rule".  IEEE
%       Transactions on Information Theory, May 1972, 431-433.
%       -- Results:
%          -- very low misclassification rates again
%    5. See also: 1988 MLC Proceedings, 54-64.  Cheeseman et al's AUTOCLASS II
%       conceptual clustering system finds 3 classes in the data.
```

```
%
% 4. Relevant Information:
%    --- This is perhaps the best known database to be found in the pattern
%        recognition literature.  Fisher's paper is a classic in the field
%        and is referenced frequently to this day.  (See Duda & Hart, for
%        example.)  The data set contains 3 classes of 50 instances each,
%        where each class refers to a type of iris plant.  One class is
%        linearly separable from the other 2; the latter are NOT linearly
%        separable from each other.
%    --- Predicted attribute: class of iris plant.
%    --- This is an exceedingly simple domain.
%
% 5. Number of Instances: 150 (50 in each of three classes)
%
% 6. Number of Attributes: 4 numeric, predictive attributes and the class
%
% 7. Attribute Information:
%    1. sepal length in cm
%    2. sepal width in cm
%    3. petal length in cm
%    4. petal width in cm
%    5. class:
%       -- Iris Setosa
%       -- Iris Versicolour
%       -- Iris Virginica
%
% 8. Missing Attribute Values: None
%
% Summary Statistics:
%                 Min  Max   Mean    SD   Class Correlation
%    sepal length: 4.3  7.9   5.84  0.83    0.7826
%     sepal width: 2.0  4.4   3.05  0.43   -0.4194
%    petal length: 1.0  6.9   3.76  1.76    0.9490  (high!)
%     petal width: 0.1  2.5   1.20  0.76    0.9565  (high!)
%
% 9. Class Distribution: 33.3% for each of 3 classes.

@RELATION iris

@ATTRIBUTE sepallength    REAL
@ATTRIBUTE sepalwidth     REAL
```

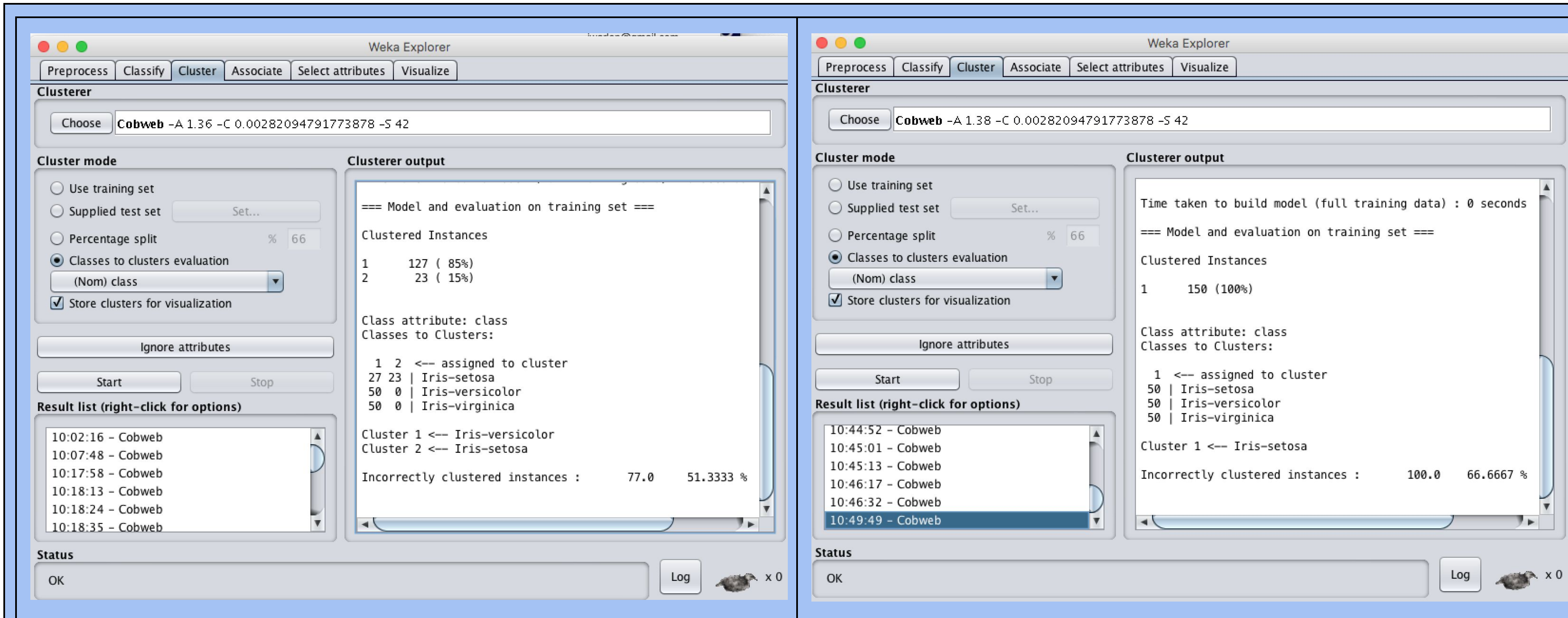
```
@ATTRIBUTE petallength    REAL
@ATTRIBUTE petalwidth     REAL
@ATTRIBUTE class    {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
4.6,3.6,1.0,0.2,Iris-setosa
5.1,3.3,1.7,0.5,Iris-setosa
4.8,3.4,1.9,0.2,Iris-setosa
5.0,3.0,1.6,0.2,Iris-setosa
5.0,3.4,1.6,0.4,Iris-setosa
5.2,3.5,1.5,0.2,Iris-setosa
5.2,3.4,1.4,0.2,Iris-setosa
4.7,3.2,1.6,0.2,Iris-setosa
4.8,3.1,1.6,0.2,Iris-setosa
5.4,3.4,1.5,0.4,Iris-setosa
5.2,4.1,1.5,0.1,Iris-setosa
5.5,4.2,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.0,3.2,1.2,0.2,Iris-setosa
```

```
5.5,3.5,1.3,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
4.4,3.0,1.3,0.2,Iris-setosa
5.1,3.4,1.5,0.2,Iris-setosa
5.0,3.5,1.3,0.3,Iris-setosa
4.5,2.3,1.3,0.3,Iris-setosa
4.4,3.2,1.3,0.2,Iris-setosa
5.0,3.5,1.6,0.6,Iris-setosa
5.1,3.8,1.9,0.4,Iris-setosa
4.8,3.0,1.4,0.3,Iris-setosa
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.5,2.8,4.6,1.5,Iris-versicolor
5.7,2.8,4.5,1.3,Iris-versicolor
6.3,3.3,4.7,1.6,Iris-versicolor
4.9,2.4,3.3,1.0,Iris-versicolor
6.6,2.9,4.6,1.3,Iris-versicolor
5.2,2.7,3.9,1.4,Iris-versicolor
5.0,2.0,3.5,1.0,Iris-versicolor
5.9,3.0,4.2,1.5,Iris-versicolor
6.0,2.2,4.0,1.0,Iris-versicolor
6.1,2.9,4.7,1.4,Iris-versicolor
5.6,2.9,3.6,1.3,Iris-versicolor
6.7,3.1,4.4,1.4,Iris-versicolor
5.6,3.0,4.5,1.5,Iris-versicolor
5.8,2.7,4.1,1.0,Iris-versicolor
6.2,2.2,4.5,1.5,Iris-versicolor
5.6,2.5,3.9,1.1,Iris-versicolor
5.9,3.2,4.8,1.8,Iris-versicolor
6.1,2.8,4.0,1.3,Iris-versicolor
6.3,2.5,4.9,1.5,Iris-versicolor
6.1,2.8,4.7,1.2,Iris-versicolor
6.4,2.9,4.3,1.3,Iris-versicolor
6.6,3.0,4.4,1.4,Iris-versicolor
6.8,2.8,4.8,1.4,Iris-versicolor
```

```
6.7,3.0,5.0,1.7,Iris-versicolor
6.0,2.9,4.5,1.5,Iris-versicolor
5.7,2.6,3.5,1.0,Iris-versicolor
5.5,2.4,3.8,1.1,Iris-versicolor
5.5,2.4,3.7,1.0,Iris-versicolor
5.8,2.7,3.9,1.2,Iris-versicolor
6.0,2.7,5.1,1.6,Iris-versicolor
5.4,3.0,4.5,1.5,Iris-versicolor
6.0,3.4,4.5,1.6,Iris-versicolor
6.7,3.1,4.7,1.5,Iris-versicolor
6.3,2.3,4.4,1.3,Iris-versicolor
5.6,3.0,4.1,1.3,Iris-versicolor
5.5,2.5,4.0,1.3,Iris-versicolor
5.5,2.6,4.4,1.2,Iris-versicolor
6.1,3.0,4.6,1.4,Iris-versicolor
5.8,2.6,4.0,1.2,Iris-versicolor
5.0,2.3,3.3,1.0,Iris-versicolor
5.6,2.7,4.2,1.3,Iris-versicolor
5.7,3.0,4.2,1.2,Iris-versicolor
5.7,2.9,4.2,1.3,Iris-versicolor
6.2,2.9,4.3,1.3,Iris-versicolor
5.1,2.5,3.0,1.1,Iris-versicolor
5.7,2.8,4.1,1.3,Iris-versicolor
6.3,3.3,6.0,2.5,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
7.1,3.0,5.9,2.1,Iris-virginica
6.3,2.9,5.6,1.8,Iris-virginica
6.5,3.0,5.8,2.2,Iris-virginica
7.6,3.0,6.6,2.1,Iris-virginica
4.9,2.5,4.5,1.7,Iris-virginica
7.3,2.9,6.3,1.8,Iris-virginica
6.7,2.5,5.8,1.8,Iris-virginica
7.2,3.6,6.1,2.5,Iris-virginica
6.5,3.2,5.1,2.0,Iris-virginica
6.4,2.7,5.3,1.9,Iris-virginica
6.8,3.0,5.5,2.1,Iris-virginica
5.7,2.5,5.0,2.0,Iris-virginica
5.8,2.8,5.1,2.4,Iris-virginica
6.4,3.2,5.3,2.3,Iris-virginica
6.5,3.0,5.5,1.8,Iris-virginica
7.7,3.8,6.7,2.2,Iris-virginica
```

```
7.7,2.6,6.9,2.3,Iris-virginica
6.0,2.2,5.0,1.5,Iris-virginica
6.9,3.2,5.7,2.3,Iris-virginica
5.6,2.8,4.9,2.0,Iris-virginica
7.7,2.8,6.7,2.0,Iris-virginica
6.3,2.7,4.9,1.8,Iris-virginica
6.7,3.3,5.7,2.1,Iris-virginica
7.2,3.2,6.0,1.8,Iris-virginica
6.2,2.8,4.8,1.8,Iris-virginica
6.1,3.0,4.9,1.8,Iris-virginica
6.4,2.8,5.6,2.1,Iris-virginica
7.2,3.0,5.8,1.6,Iris-virginica
7.4,2.8,6.1,1.9,Iris-virginica
7.9,3.8,6.4,2.0,Iris-virginica
6.4,2.8,5.6,2.2,Iris-virginica
6.3,2.8,5.1,1.5,Iris-virginica
6.1,2.6,5.6,1.4,Iris-virginica
7.7,3.0,6.1,2.3,Iris-virginica
6.3,3.4,5.6,2.4,Iris-virginica
6.4,3.1,5.5,1.8,Iris-virginica
6.0,3.0,4.8,1.8,Iris-virginica
6.9,3.1,5.4,2.1,Iris-virginica
6.7,3.1,5.6,2.4,Iris-virginica
6.9,3.1,5.1,2.3,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
6.8,3.2,5.9,2.3,Iris-virginica
6.7,3.3,5.7,2.5,Iris-virginica
6.7,3.0,5.2,2.3,Iris-virginica
6.3,2.5,5.0,1.9,Iris-virginica
6.5,3.0,5.2,2.0,Iris-virginica
6.2,3.4,5.4,2.3,Iris-virginica
5.9,3.0,5.1,1.8,Iris-virginica
%
%
%
```



Acuity 1.36 & 1.38 Notice how number of clusters is @ 2 w/1.36 and then back to one @ 1 w/1.38

[Home]

baskball.arff

```
% Dataset from Smoothing Methods in Statistics
% (ftp stat.cmu.edu/datasets)
%
% Simonoff, J.S. (1996). Smoothing Methods in Statistics. New York: Springer-Verlag.
%
%
% Points scored per minute is being treated as
% the class attribute.

@relation baskball
```

```
@attribute assists_per_minute real
@attribute height integer
@attribute time_played real
@attribute age integer
@attribute points_per_minute real

@data
0.0888,201,36.02,28,0.5885
0.1399,198,39.32,30,0.8291
0.0747,198,38.8,26,0.4974
0.0983,191,40.71,30,0.5772
0.1276,196,38.4,28,0.5703
0.1671,201,34.1,31,0.5835
0.1906,193,36.2,30,0.5276
0.1061,191,36.75,27,0.5523
0.2446,185,38.43,29,0.4007
0.167,203,33.54,24,0.477
0.2485,188,35.01,27,0.4313
0.1227,198,36.67,29,0.4909
0.124,185,33.88,24,0.5668
0.1461,191,35.59,30,0.5113
0.2315,191,38.01,28,0.3788
0.0494,193,32.38,32,0.559
0.1107,196,35.22,25,0.4799
0.2521,183,31.73,29,0.5735
0.1007,193,28.81,34,0.6318
0.1067,196,35.6,23,0.4326
0.1956,188,35.28,32,0.428
0.1828,191,29.54,28,0.4401
0.1627,196,31.35,28,0.5581
0.1403,198,33.5,23,0.4866
0.1563,193,34.56,32,0.5267
0.2681,183,39.53,27,0.5439
0.1236,196,26.7,34,0.4419
0.13,188,30.77,26,0.3998
0.0896,198,25.67,30,0.4325
0.2071,178,36.22,30,0.4086
0.2244,185,36.55,23,0.4624
0.3437,185,34.91,31,0.4325
0.1058,191,28.35,28,0.4903
0.2326,185,33.53,27,0.4802
```

```
0.1577,193,31.07,25,0.4345
0.2327,185,36.52,32,0.4819
0.1256,196,27.87,29,0.6244
0.107,198,24.31,34,0.3991
0.1343,193,31.26,28,0.4414
0.0586,196,22.18,23,0.4013
0.2383,185,35.25,26,0.3801
0.1006,198,22.87,30,0.3498
0.2164,193,24.49,32,0.3185
0.1485,198,23.57,27,0.3097
0.227,191,31.72,27,0.4319
0.1649,188,27.9,25,0.3799
0.1188,191,22.74,24,0.4091
0.194,193,20.62,27,0.3588
0.2495,185,30.46,25,0.4727
0.2378,185,32.38,27,0.3212
0.1592,191,25.75,31,0.3418
0.2069,170,33.84,30,0.4285
0.2084,185,27.83,25,0.3917
0.0877,193,21.67,26,0.5769
0.101,193,21.79,24,0.4773
0.0942,201,20.17,26,0.4512
0.055,193,29.07,31,0.3096
0.1071,196,24.28,24,0.3089
0.0728,193,19.24,27,0.4573
0.2771,180,27.07,28,0.3214
0.0528,196,18.95,22,0.5437
0.213,188,21.59,30,0.4121
0.1356,193,13.27,31,0.2185
0.1043,196,16.3,23,0.3313
0.113,191,23.01,25,0.3302
0.1477,196,20.31,31,0.4677
0.1317,188,17.46,33,0.2406
0.2187,191,21.95,28,0.3007
0.2127,188,14.57,37,0.2471
0.2547,160,34.55,28,0.2894
0.1591,191,22,24,0.3682
0.0898,196,13.37,34,0.389
0.2146,188,20.51,24,0.512
0.1871,183,19.78,28,0.4449
0.1528,191,16.36,33,0.4035
```

```
0.156,191,16.03,23,0.2683
0.2348,188,24.27,26,0.2719
0.1623,180,18.49,28,0.3408
0.1239,180,17.76,26,0.4393
0.2178,185,13.31,25,0.3004
0.1608,185,17.41,26,0.3503
0.0805,193,13.67,25,0.4388
0.1776,193,17.46,27,0.2578
0.1668,185,14.38,35,0.2989
0.1072,188,12.12,31,0.4455
0.1821,185,12.63,25,0.3087
0.188,180,12.24,30,0.3678
0.1167,196,12,24,0.3667
0.2617,185,24.46,27,0.3189
0.1994,188,20.06,27,0.4187
0.1706,170,17,25,0.5059
0.1554,183,11.58,24,0.3195
0.2282,185,10.08,24,0.2381
0.1778,185,18.56,23,0.2802
0.1863,185,11.81,23,0.381
0.1014,193,13.81,32,0.1593
```

Home

Cluster.EM Eval Training Data

```
=== Run information ===

Scheme:       weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Relation:     baskball
Instances:    96
Attributes:   5
              assists_per_minute
              height
              time_played
              age
              points_per_minute
Test mode:    evaluate on training data
```

```
=== Clustering model (full training set) ===


EM
==

Number of clusters selected by cross validation: 2
Number of iterations performed: 13


                      Cluster
Attribute                   0        1
                       (0.43)   (0.57)
=======================================
assists_per_minute
  mean                 0.1755   0.1505
  std. dev.            0.0649   0.0525

height
  mean               189.7582 189.9633
  std. dev.            8.1076   5.8722

time_played
  mean                34.3817   19.569
  std. dev.            3.2046   5.1892

age
  mean                28.0232  27.5253
  std. dev.            2.6823    3.697

points_per_minute
  mean                 0.4889   0.3684
  std. dev.            0.0953   0.0865



Time taken to build model (full training data) : 0.13 seconds

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

```
Clustered Instances

0      42 ( 44%)
1      54 ( 56%)


Log likelihood: -6.95576
```

```
=== Run information ===

Scheme:       weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Relation:     baskball
Instances:    96
Attributes:   5
              assists_per_minute
              height
              time_played
              age
              points_per_minute
Test mode:    split 66% train, remainder test


=== Clustering model (full training set) ===


EM
==

Number of clusters selected by cross validation: 2
Number of iterations performed: 13


                      Cluster
Attribute                   0        1
                       (0.43)   (0.57)
=======================================
assists_per_minute
```

```
  mean                   0.1755   0.1505
  std. dev.              0.0649   0.0525

height
  mean                 189.7582 189.9633
  std. dev.              8.1076   5.8722

time_played
  mean                  34.3817   19.569
  std. dev.              3.2046   5.1892

age
  mean                  28.0232  27.5253
  std. dev.              2.6823    3.697

points_per_minute
  mean                   0.4889   0.3684
  std. dev.              0.0953   0.0865



Time taken to build model (full training data) : 0.15 seconds

=== Model and evaluation on test split ===

EM
==

Number of clusters selected by cross validation: 2
Number of iterations performed: 29


                      Cluster
Attribute                   0        1
                       (0.59)   (0.41)
=======================================
assists_per_minute
  mean                   0.1454   0.1734
  std. dev.              0.0542   0.0692

height
```

```
  mean                189.9153 190.6965
  std. dev.             6.6357   7.7172

time_played
  mean                 19.2418  34.7894
  std. dev.             5.1703   3.0115

age
  mean                 27.3435  28.0104
  std. dev.             3.7166   2.6252

points_per_minute
  mean                   0.366   0.4909
  std. dev.             0.0831   0.1011



Time taken to build model (percentage split) : 0.08 seconds

Clustered Instances

0      17 ( 52%)
1      16 ( 48%)


Log likelihood: -6.89082
```