Stat645

Model visualisation

Hadley Wickham

1. Classify
2. Clusterfly

Classifly

Classification

In high-dimensional space, how can we find the boundary that does the best job of separating two (or more) groups.

Basic idea

How can we better understand how different classification algorithms work?

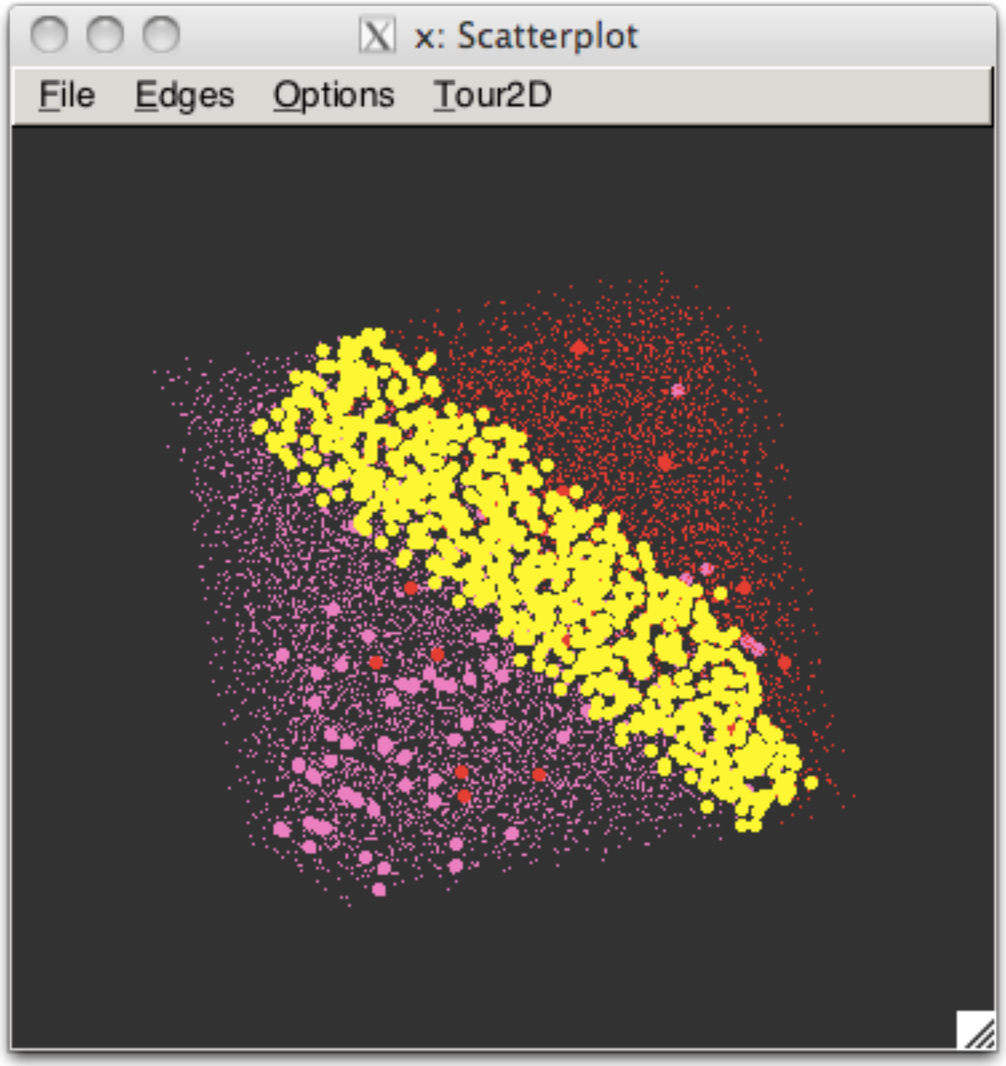
Advantage is the difference between the probability of the best and second best predictions - if small, must be close to boundary.

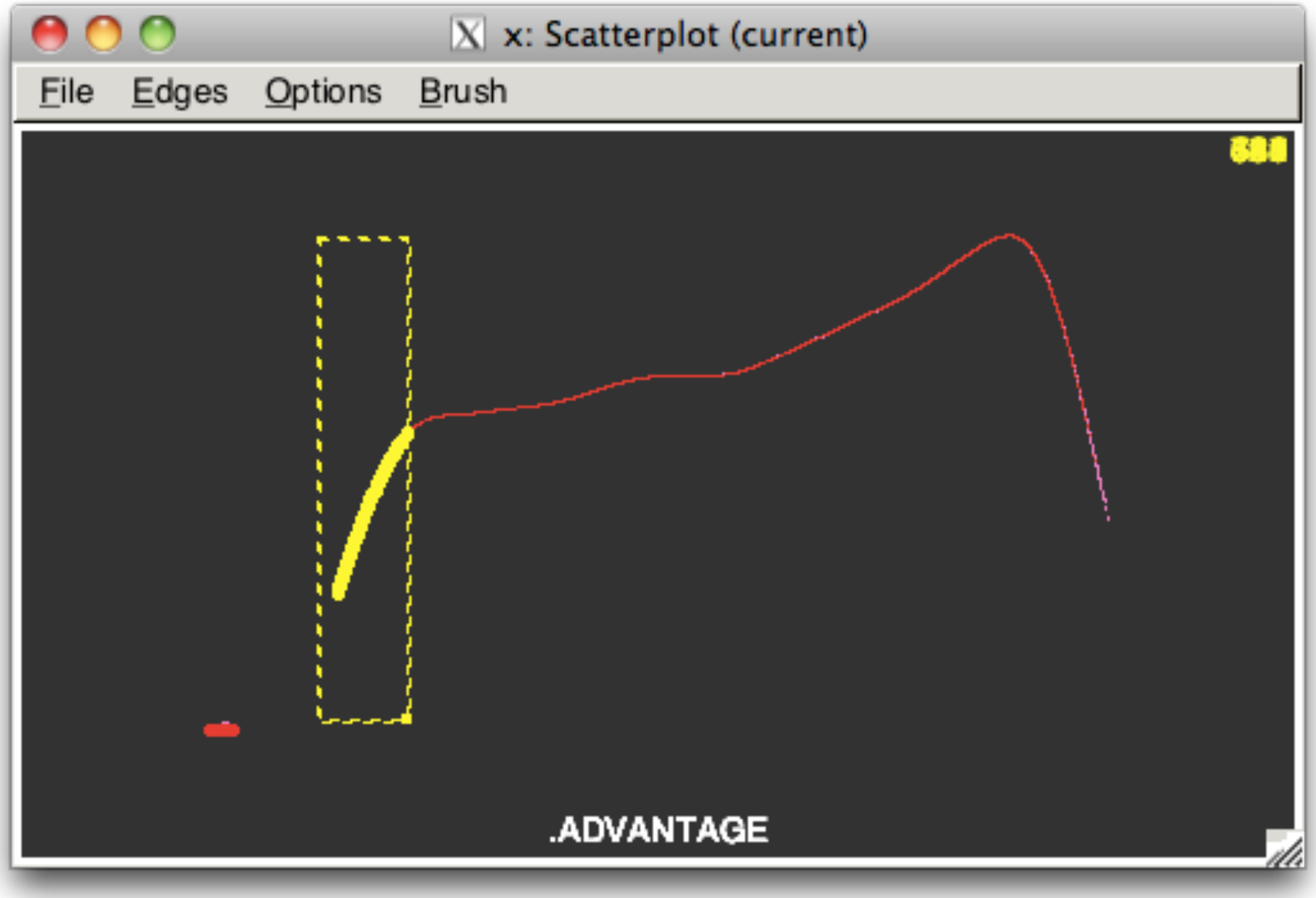
Big points are data, little points are predictions.

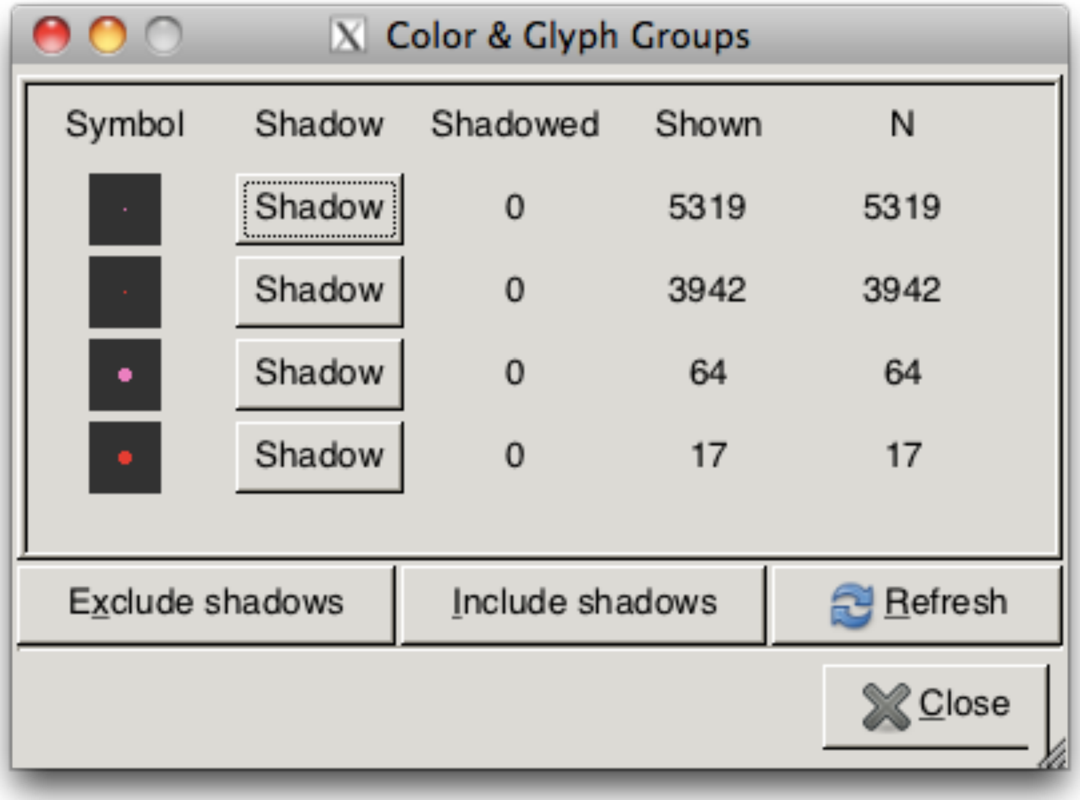
Techniques

Look at the boundaries: kyphosis-lda.xml. Open 1d view of .ADVANTAGE and brush.

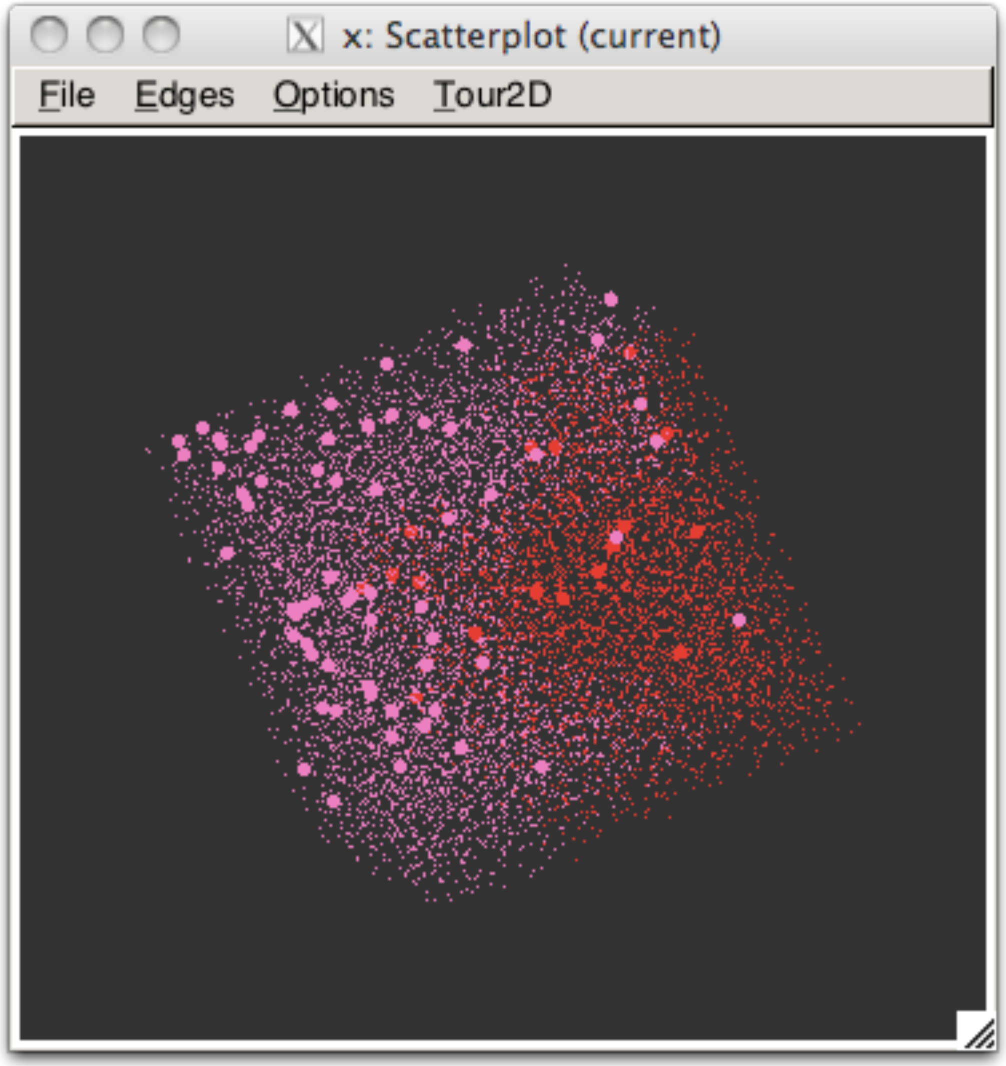
Look at the groups: use colour and glyph groups to show only one region.

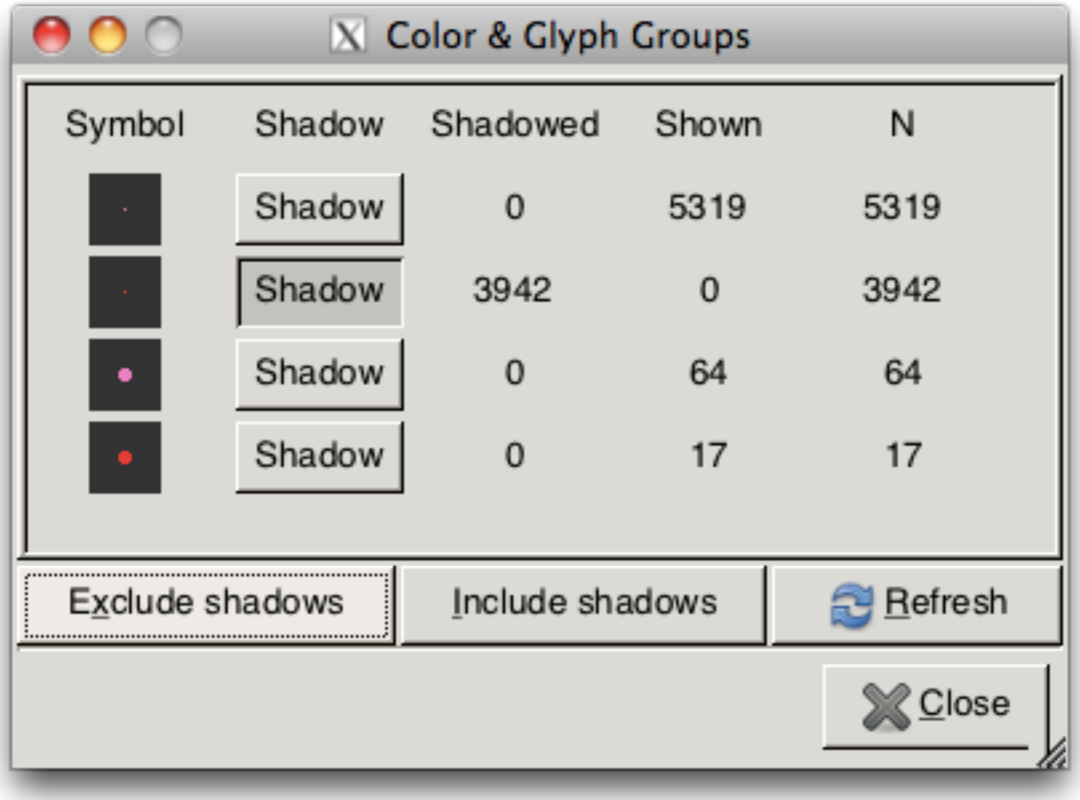

x: Scatterplot
File
Edges
Options
Tour2D


x: Scatterplot (current)
File
Edges
Options
Brush
.ADVANTAGE

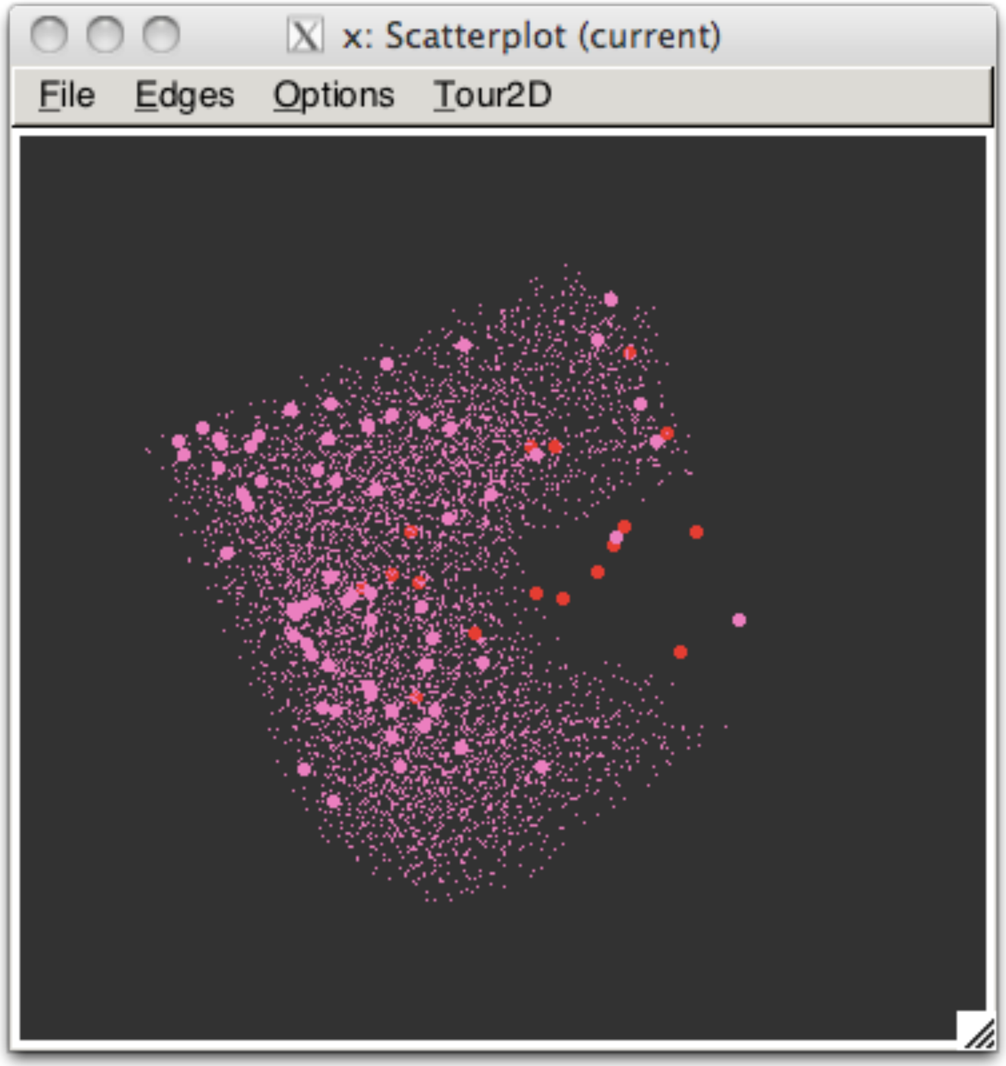


Symbol	Shadow	Shadowed	Shown	N
·	Shadow	0	5319	5319
·	Shadow	0	3942	3942
●	Shadow	0	64	64
●	Shadow	0	17	17





Symbol	Shadow	Shadowed	Shown	N
·	Shadow	0	5319	5319
·	Shadow	3942	0	3942
●	Shadow	0	64	64
●	Shadow	0	17	17



Your turn

Use these techniques to explore the wine classification data. What shapes are the boundaries?

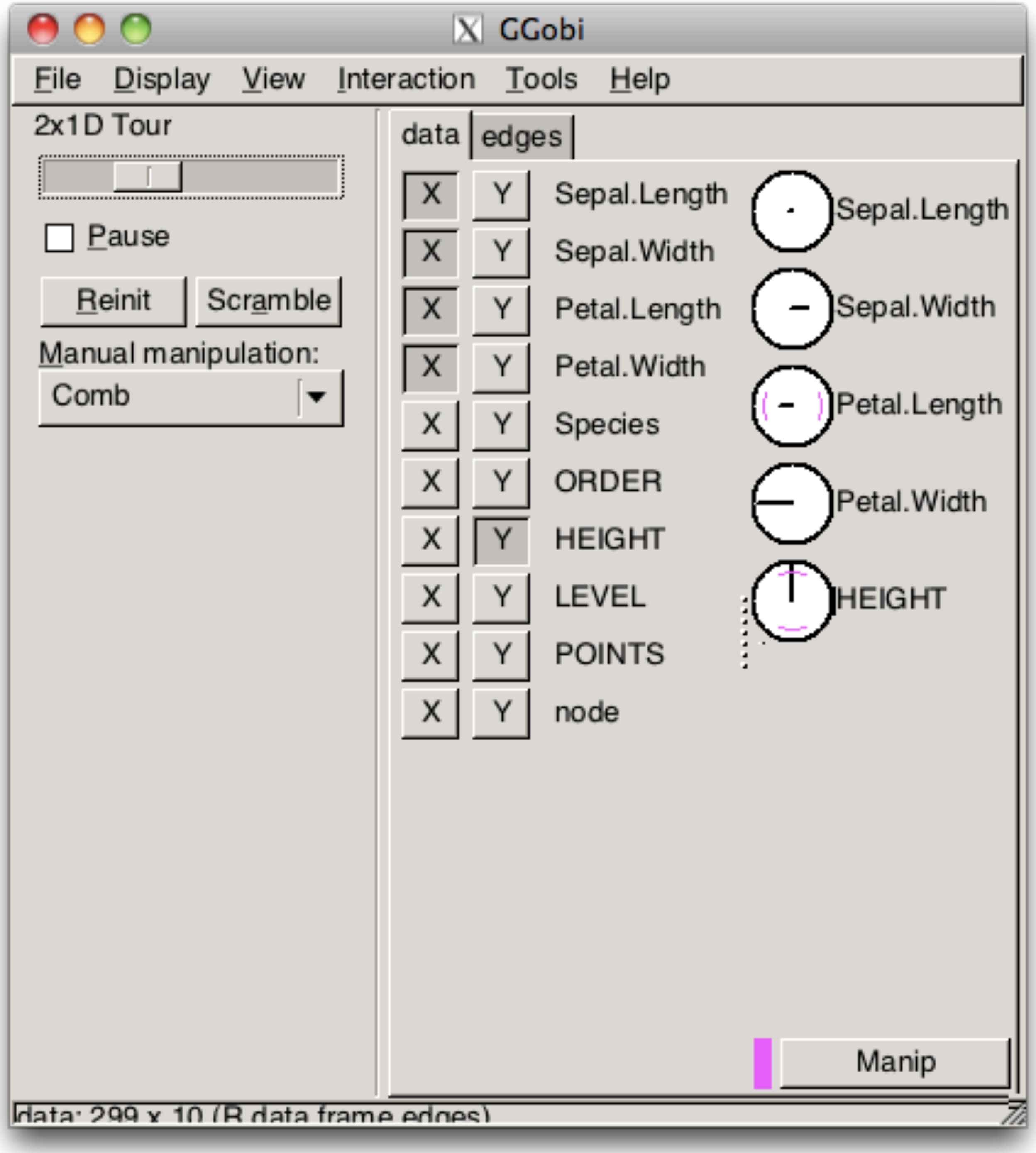
Clusterfly

Hierarchical clustering

Iteratively join closest points/clusters

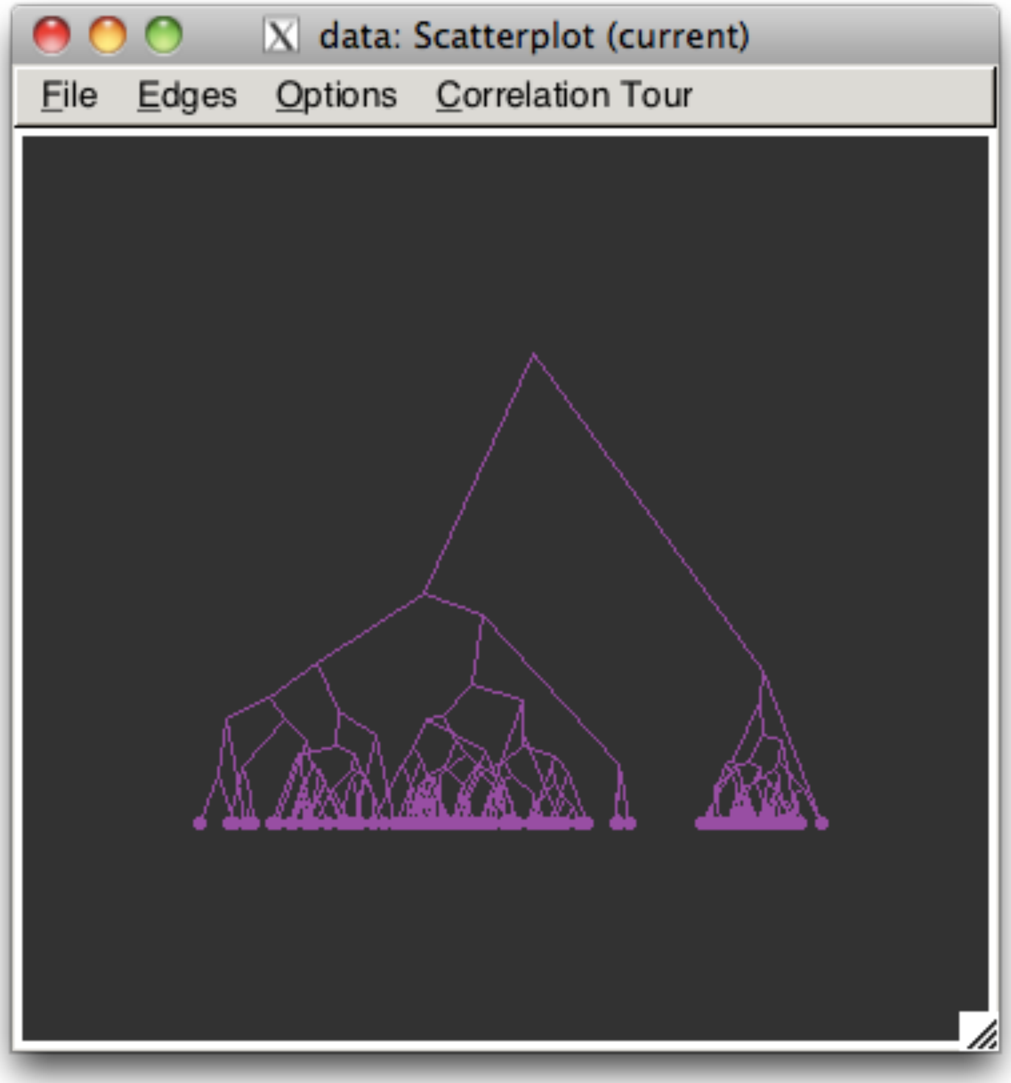
Two main parameters: what distance metric to use, definition of distance between two clusters

iris.xml



Manual manipulation: Comb

data: 299 x 10 (R data frame edges)


data: Scatterplot (current)
File
Edges
Options
Correlation Tour