# Targeted Projection Pursuit

## Quick Start

Targeted Projection Pursuit (TPP) is a technique for visualising and exploring large complex data sets. Its a version of [Projection Pursuit](http://en.wikipedia.org/wiki/Projection_pursuit) that allows the user to find interesting projections of their data using a simple and intuitive interface.

What is TPP?

Targeted Projection Pursuit is a way of interactively exploring complex data. To understand what it does, first suppose we have a simple two-dimensional set of data. That is, we have two measurements for each sample (imagine an Excel spreadsheet with 2 columns and lots of rows). We could picture this data using a conventional scatter plot like this:

Here we can see that the points are divided into three groups. And we can color the points to see the groups more clearly:

But now suppose we had three measurements for each sample. We can’t put three dimensions on a flat page or screen, but can see a [linear projection](http://en.wikipedia.org/wiki/Projection_(linear_algebra)) of it – a view of the data from one particular angle

One standard linear projection is [Principal Components](http://en.wikipedia.org/wiki/Principal_component_analysis), which can be thought of as that view of a data set in which the points are most spread out. But no single linear projection will show all the detail in the data. By changing the angle of the view we can see different aspects.

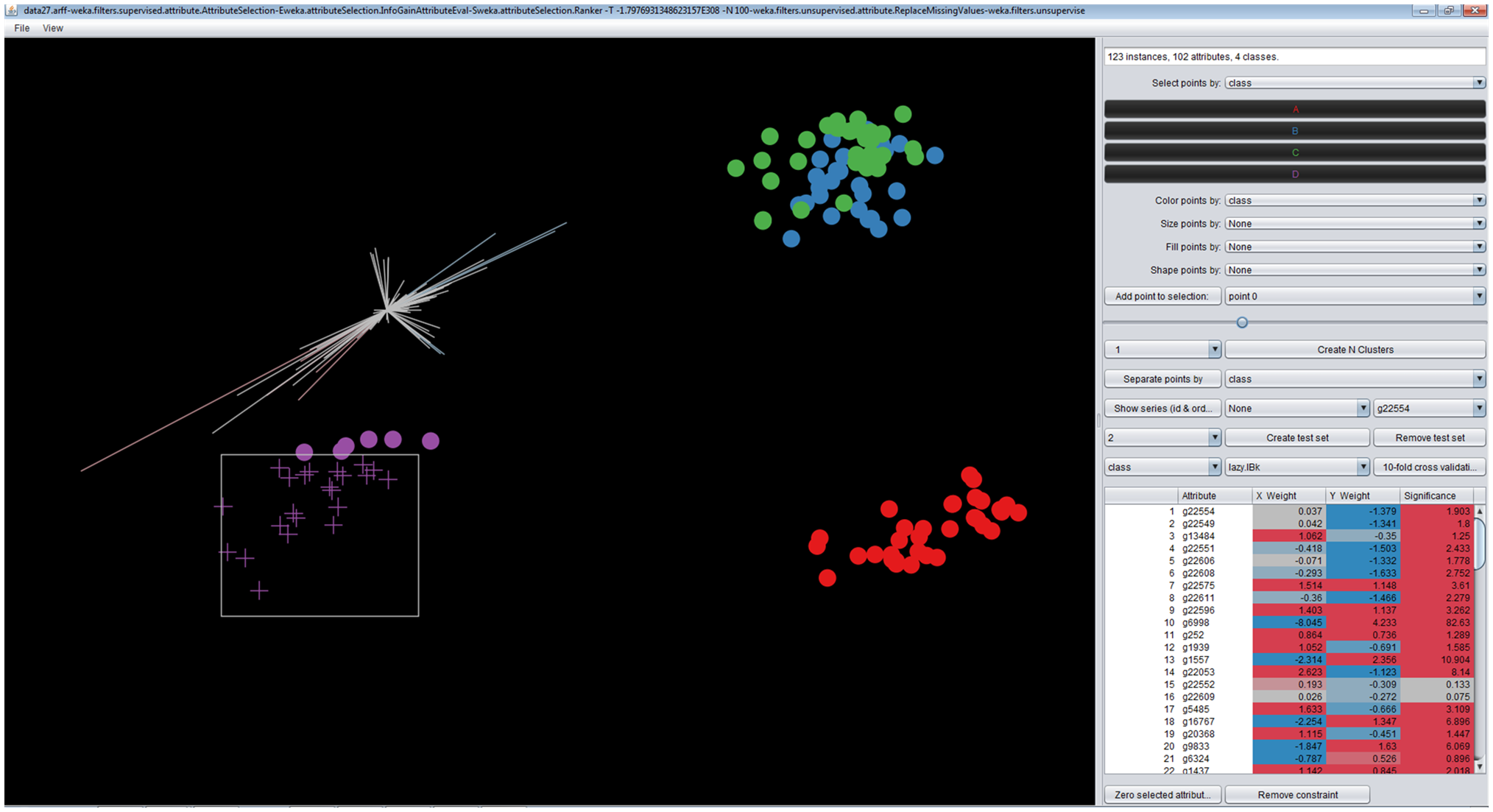
For example, from this angle we can see how the different classes in the data can be separated – and the length of the axes tells us which variables play the greatest role in separating those classes.

Now suppose that we had not two or three, but hundreds of dimensions in our data. We can still use a linear projection to view that high dimensional space on our screen.

But the problem now is how to explore the data to see the many different views of the data. Targeted Projection Pursuit solves this problem by allowing the user to drag points to find new views. These views can tell us:

* Which classes can be separated and which can not
* Which attributes are most associated with each class and in what combinations
* Which points are outliers, and to what extent
* The sources of error in classification algorithms
* The existence of clusters in the data: the number, extent, and separability

See the video for a demonstration.



**The axes – one for each attribute. The length indicates the importance of the axis in separating the classes. The colour indicates the relative value of that attribute for the selected points. (Red indicates higher than mean; blue indicates lower than mean).**

**Points can be selected by area, then dragged to find a new view of the data**

**These classes of points could not be separated, indicating that there was little difference between them**

**Points can also be selected by class.**

**These controls allow you to change the attribute used to determine the colour, size, fill, and shape of each point. The slider changes the mean size.**

**The projection table shows the position and overall length of each attribute. The table can be ordered by any column (click on the heading)**

**These buttons show the effect of removing attributes by setting selected axes/attributes to zero**

## Loading Data

TPP prefers its data in [ARFF format](http://weka.wikispaces.com/ARFF+%28book+version%29), but can also accept data in tab-separated values format. To load a data file, click on the File menu item and choose what format data file you want to load from.

## Initial View

The data is initially shown projected onto the first two principal components (X=PC1, Y=PC2). The axes show the positions of the attributes and the projection table shows their numerical values. The ‘Significance’ column shows the overall length of each axis. The points are coloured using the first nominal attribute (classification) found in the data.

## Exploring the Data

TPP allows you to see different views of the data by grabbing and dragging points. If a linear projection can be found that matches the new position of the points then it will be shown

The

exploring data

selecting points

by button

selecting by different attributes

by rectangle

separating points

moving axes

zeroing axes

fitting points to window

the projection table

changing the view

color by

size by

slider

fill by

shape by

other actions

unsupervised clustering

hierarchical clustering

show series

create test sets

supervised classification

exporting data

images

data

proteins