RadViz : The Visual Data Mining Tool

Abstract

RadViz, a radial visualization based on the spring paradigm, with appropriate interactive and layout controls, can effectively be used as a Classifier, a Clustering Tool, a Feature Reduction mechanism and an Association Rule generator. Thus, RadViz is a general machine learning tool that covers the main application areas in data mining. . In these applications, visual clustering is a predominant mechanism in achieving results. For example RadViz is a "gradual" classifier that can effectively classify multi-class problems by giving a "visual" indication of the probability of a record. This paper will discuss these applications of RadViz.

The RadViz Layout:

An example of the RadViz layout is illustrated in Figure 1. There are 16 variables or dimensions associated with the 1 point plotted (in red). Sixteen imaginary springs are anchored to the points on the circumference and attached to one data point. The data point is plotted where the sum of the forces are zero according to Hooke's law (F = Kx): where the force is proportional to the distance x to the anchor point. The value K for each spring is the value of the variable for the data point. In this example the spring constants (or dimensional values) are higher for the yellow springs and lower for the blue springs. Normally, many points are plotted without showing the spring lines. Generally, the dimensions (variables) are normalized to have values between 0 and 1 so that all dimensions have "equal" weights. This spring paradigm layout as some interesting features.

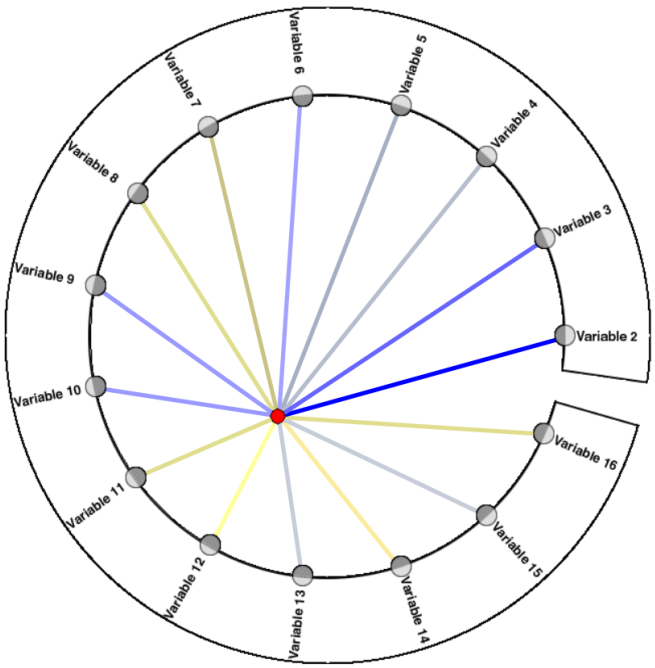


Figure 1 RadViz Basics

For example if all dimensions have the same normalized value the data point will lie exactly in the center of the circle. If the point is a unit vector then that point will lie exactly at the fixed point on the edge of the circle (where the spring for that dimension is

fixed). Many points can map to the same position. This represents a non-linear transformation of the data which preserves certain symmetries and which produces an intuitive display. Some features of this visualization include:

- it is intuitive, higher dimension values "pull" the data points closer to the dimension on the circumference
- points with approximately equal dimension values will lie close to the center
- points with similar values whose dimensions are opposite each other on the circle will lie near the center
- points which have one or two dimension values greater than the others lie closer to those dimensions
- the relative locations of the of the dimension anchor points can drastically affect the layout (the idea behind the "Class discrimination layout" algorithm)
- an n-dimensional line gets mapped to a line (or a single point) in RadViz
- Convex sets in *n*-space map into convex sets in RadViz
- Computation time is very fast
- 1000's of dimensions can be displayed in one visualization

N-Space vs RadViz space:

Phil this is yours!

The standard normalization used in RadViz is to scale and translate each dimension's maximum and minimum to the values 1 and 0. Thus we map an n-dimensional unit cube to a 2-dimensional unit circle.

The RadViz mapping function is:

$$R(\vec{v}) = A\vec{v}$$

$$\text{with } A = \begin{pmatrix} \cos\frac{2\pi}{n} & \cos\frac{4\pi}{n} & \ldots & \cos\frac{2(n-1)\pi}{n} \\ \sin\frac{2\pi}{n} & \sin\frac{4\pi}{n} & \ldots & \sin\frac{2(n-1)\pi}{n} \end{pmatrix},$$

Some basic properties of the mapping

- $R(\vec{0}) = \vec{0}$
- **A line in *n*-space maps into a line in the RadViz display (sometimes to a single point)**
- **Convex sets in *n*-space map into convex sets in RadViz**

Implications of these properties

- **All points on a line through the origin in *n*-space are mapped into a single point in RadViz**

- **If two points in *n*-space are mapped into the same point in RadViz, all points on the line connecting them in *n*-space are mapped to that single point in RadViz**
- **The inverse image of a point on the RadViz display is a connected set in *n*-space.**
- **Considering points in *n*-space as vectors:**
- **If two linearly independent vectors in *n*-space are mapped into the same point in RadViz, every vector in the subspace spanned by them is mapped into that point.**
- **That is, any linear combination of the two vectors is mapped into the same point in RadViz.**
- **Considering points in *n*-space as vectors:**
- **If two linearly independent vectors in *n*-space are mapped into the same point in RadViz, every vector in the subspace spanned by them is mapped into that point. That is, any linear combination of the two vectors is mapped into the same point in RadViz.**

Inverse images

- **The inverse image of a point in the *n*-sided polygon in the RadViz display always contains (at least) a line through the origin in *n*-space.**

- **The inverse image of a point in the <u>interior</u> of the polygon is an $n - 2$ dimensional hyperplane in *n*-space.**

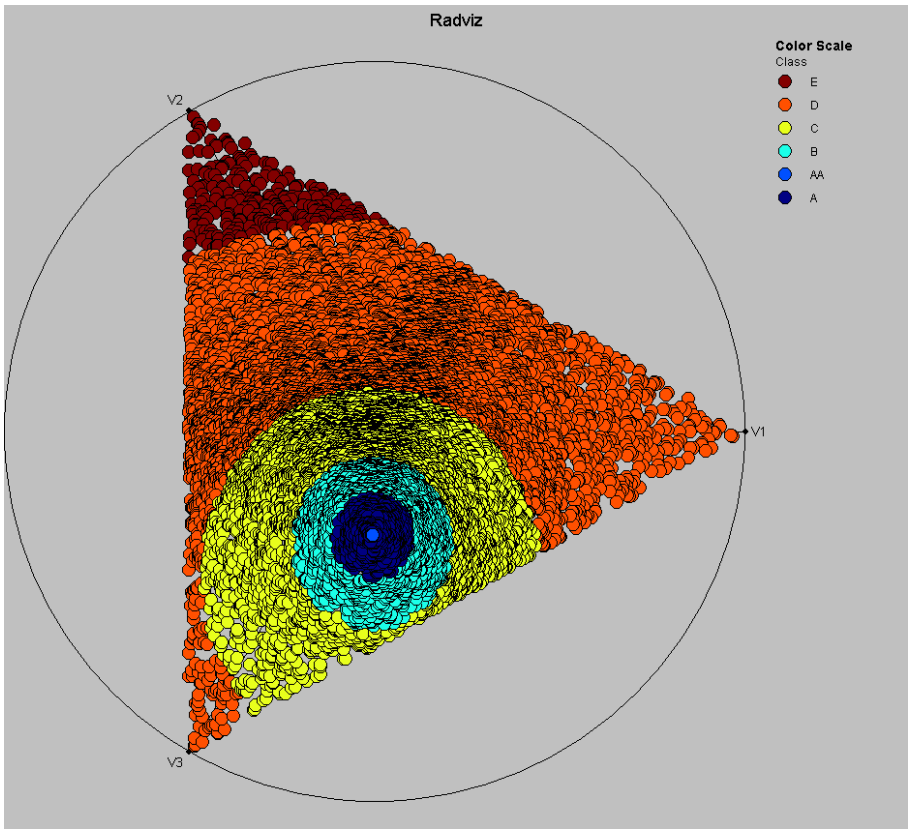


Figure 2 Cosine correlation values to the targeted point

Classification and Feature Reduction

Balanced Class Layout

A difficulty with RadViz is arranging the dimensions around the circumference to get meaningful results. As an example in **Figure 3**, over 7000 gene expression values are plotted for 38 patients. There are 38 dimensional values or variables for each of the 7000 genes. Eleven of the patients have one type of cancer (ALL) and 27 patients have another type (AML). The layout of the dimensions (patients) is random and the color of each gene is the value of the t-statistic comparing the expression values of the two types of cancer patients.

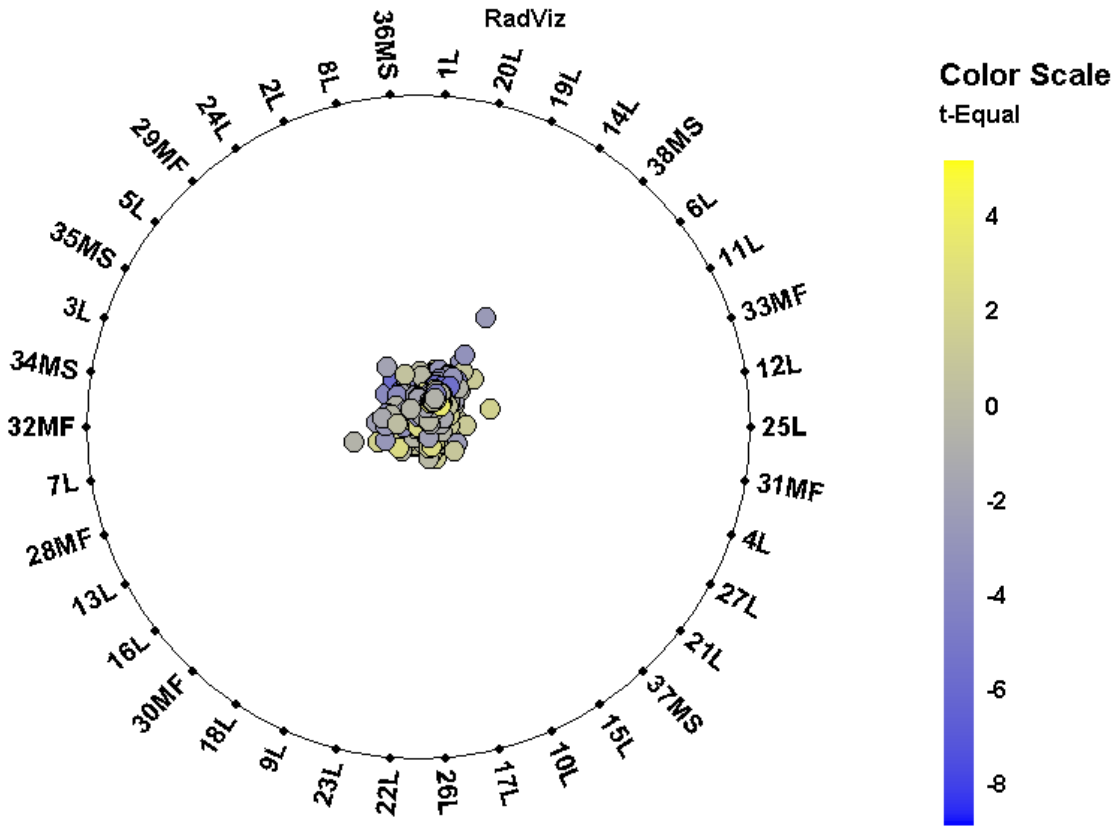


Figure 3 - Gene Expression two class problem

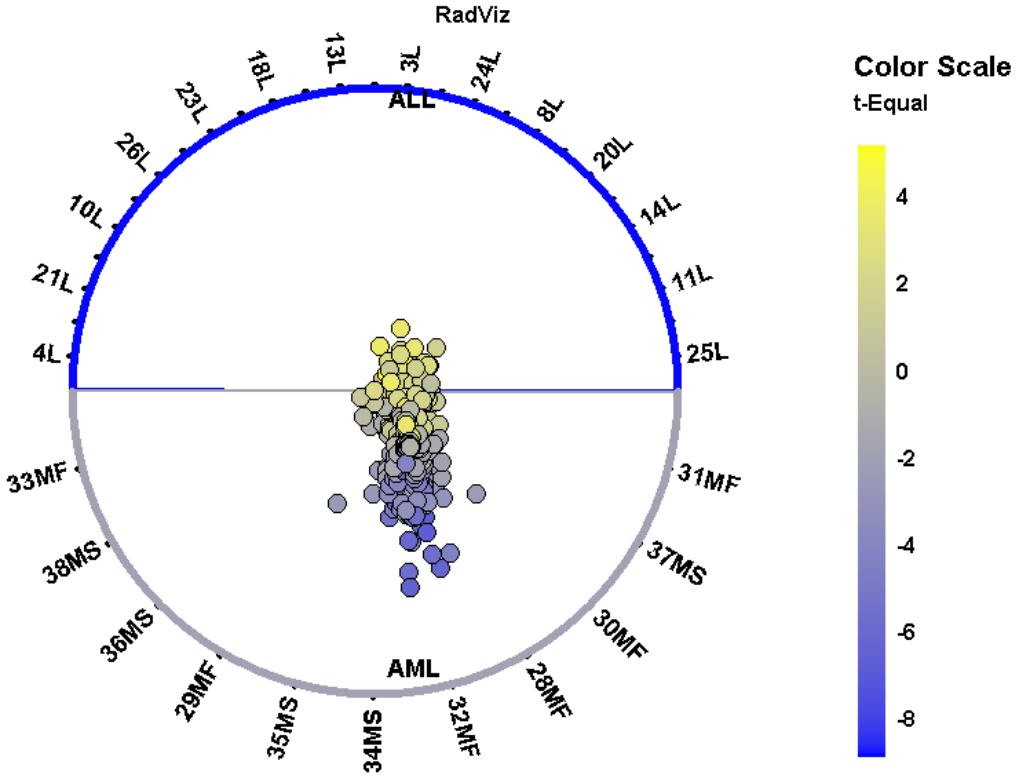


Figure 4 – Balancing the layout for two classes
In **Figure 4** all the ALL patients are arranged in the upper hemisphere while the AML patients are arranged in the lower hemisphere. A waiting factor for the spring constant is also used to balance the 27 and 11 dimensions. It can be seen by the different coloring of the t-statistic that genes with higher mean expression values for ALL or AML are "pulled" to that hemisphere. Simply selecting genes at the upper and lower extremes will give genes that that are differentially expressed for the two caner types.

Pivoting the Data Table
In **Figure 5** the data has been pivoted so that the 7000 genes are dimensions or anchor points around the circumference of RadViz. Only some of the gene accession numbers are shown. The 38 sample patient points are in the center colored by the cancer type. The genes are positioned randomly and no grouping or clustering can be seen.

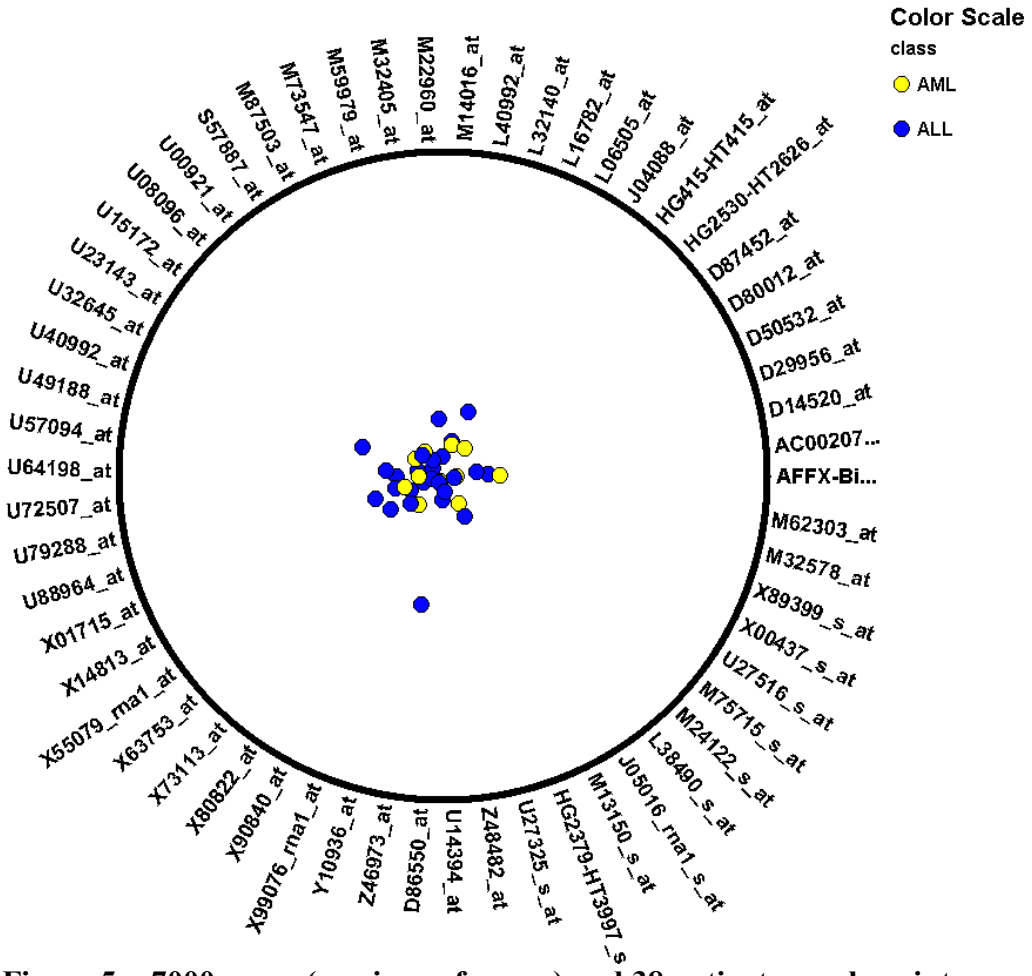


Figure 5 - 7000 genes (on circumference) and 38 patient sample points

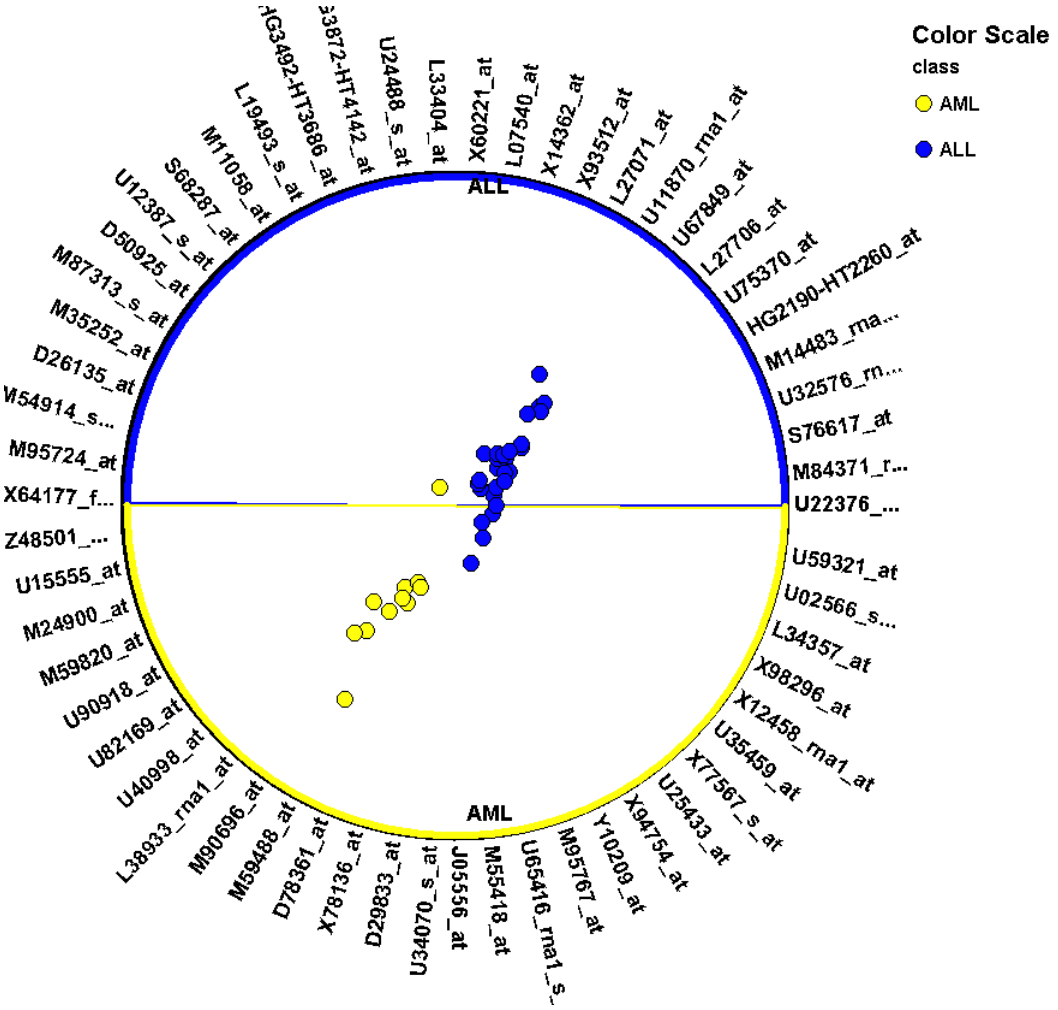


Figure 6 - 7000 genes (on circumference) and 38 patient sample points arranged by t-statistic

In **Figure 6** the genes are arranged in order of the t-statistic value when comparing the gene expression value between the two classes. The genes with higher mean expression values for ALL are in the upper hemisphere arranged from right to left and the genes with higher mean expression values for AML are positioned in the lower hemisphere, arranged from right to left. This arrangement gives the tilted class separation shown in **Figure 6**. From the layout it is seen that the gene arrangement clearly separates the classes except for the one yellow (light) sample in the center.

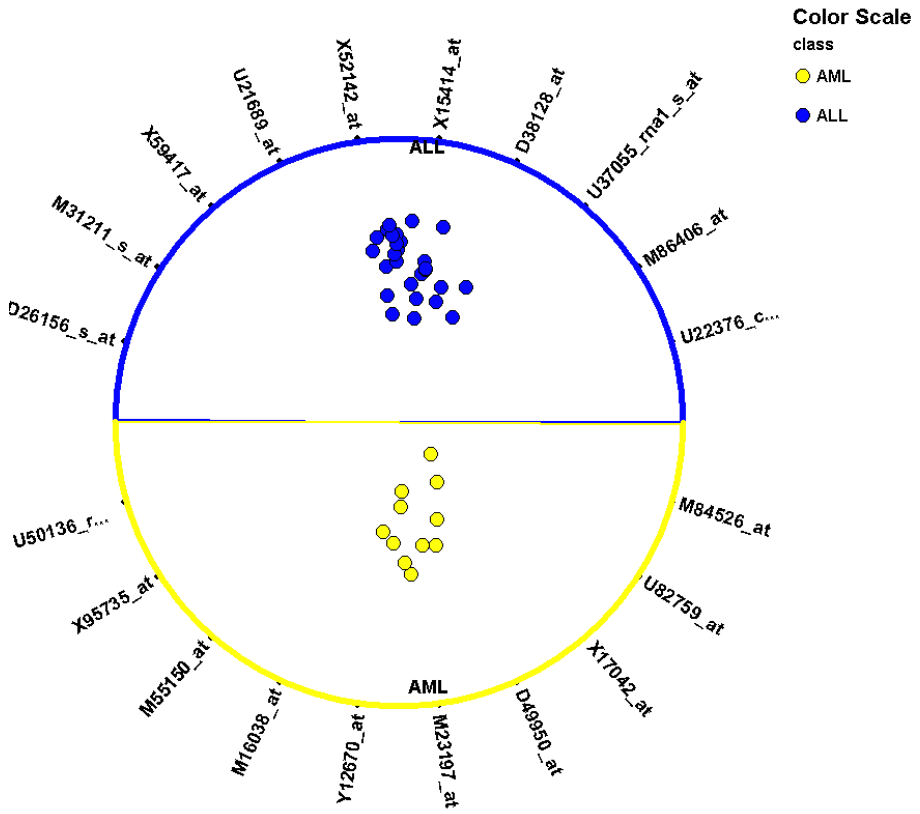


Figure 7 - Top 10 genes for AML and ALL based on the t-statistic between the two classes

In **Figure 7** only the genes with the highest t-statistic for each class(10) are used in the RadViz arrangement. This layout suggests that the 20 genes could produce a good classification for the two different cancers and using 14 classifiers from the Weka package [xxx] all but one correctly classified the 38 patients using 10 fold validation. We have used several statistics for feature reduction and RadViz layout arrangements and found that the layouts can quickly show whether a good classifier can be found and whether sub classes and/or outliers exist.

Row Name	Assigned Class	ALL	AML
Column Count Per Class		10	10
U22376_cds2_s_at (t-stat. value)	ALL	5.16759	-5.16759
M86406_at (t-stat. value)	ALL	4.96878	-4.96878
U37055_rna1_s_at (t-stat. value)	ALL	4.76987	-4.76987
D38128_at (t-stat. value)	ALL	4.72355	-4.72355
X15414_at (t-stat. value)	ALL	4.64553	-4.64553
X52142_at (t-stat. value)	ALL	4.63855	-4.63855
U21689_at (t-stat. value)	ALL	4.58085	-4.58085
X59417_at (t-stat. value)	ALL	4.57263	-4.57263
M31211_s_at (t-stat. value)	ALL	4.45531	-4.45531
D26156_s_at (t-stat. value)	ALL	4.45253	-4.45253
U50136_rna1_at (t-stat. value)	AML	-8.86979	8.86979
X95735_at (t-stat. value)	AML	-8.66966	8.66966
M55150_at (t-stat. value)	AML	-8.32267	8.32267
M16038_at (t-stat. value)	AML	-7.40101	7.40101
Y12670_at (t-stat. value)	AML	-7.39384	7.39384
M23197_at (t-stat. value)	AML	-7.25707	7.25707
D49950_at (t-stat. value)	AML	-6.96854	6.96854
X17042_at (t-stat. value)	AML	-6.79233	6.79233
U82759_at (t-stat. value)	AML	-6.74278	6.74278
M84526_at (t-stat. value)	AML	-6.6705	6.6705

t-statistic for top 10 genes distinguishing ALL from AML cancer tissue

Multi-Class Problems

The RadViz layout can also handle multi-class problems by using a pair wise t-statistic for arranging the dimensions. That is the dimensions for each class are found by a t-statistic comparing each class to all other classes combined. Both positive and negative t-statistics can be used in the layout arrangement. In **Figures 8 and 9** a three class and 6 class layout is shown.

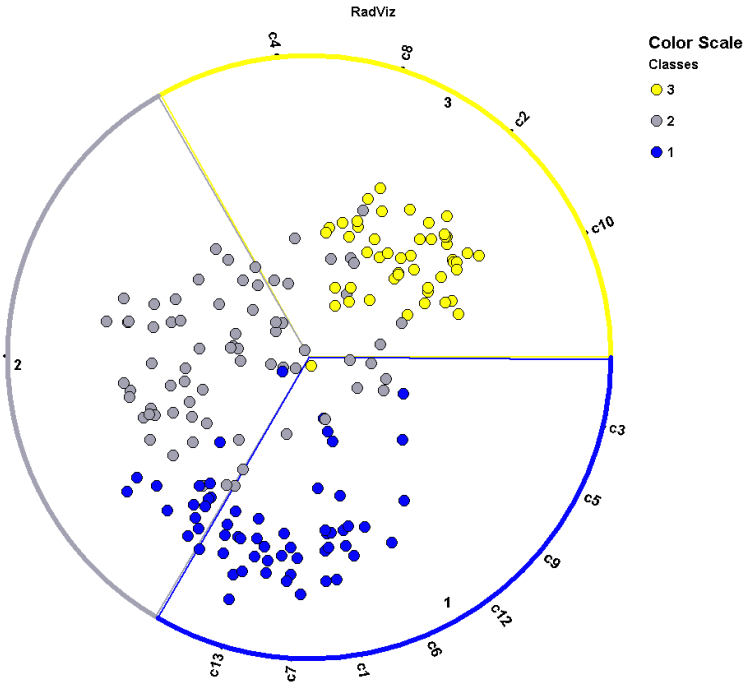


Figure 8 – a 3class problem (Wine from UCI Dataset)

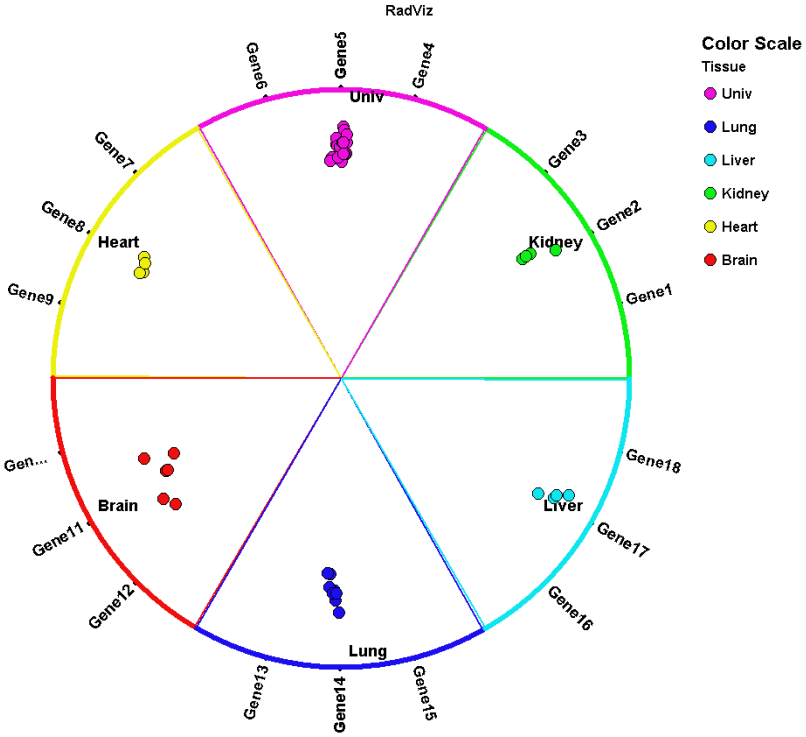
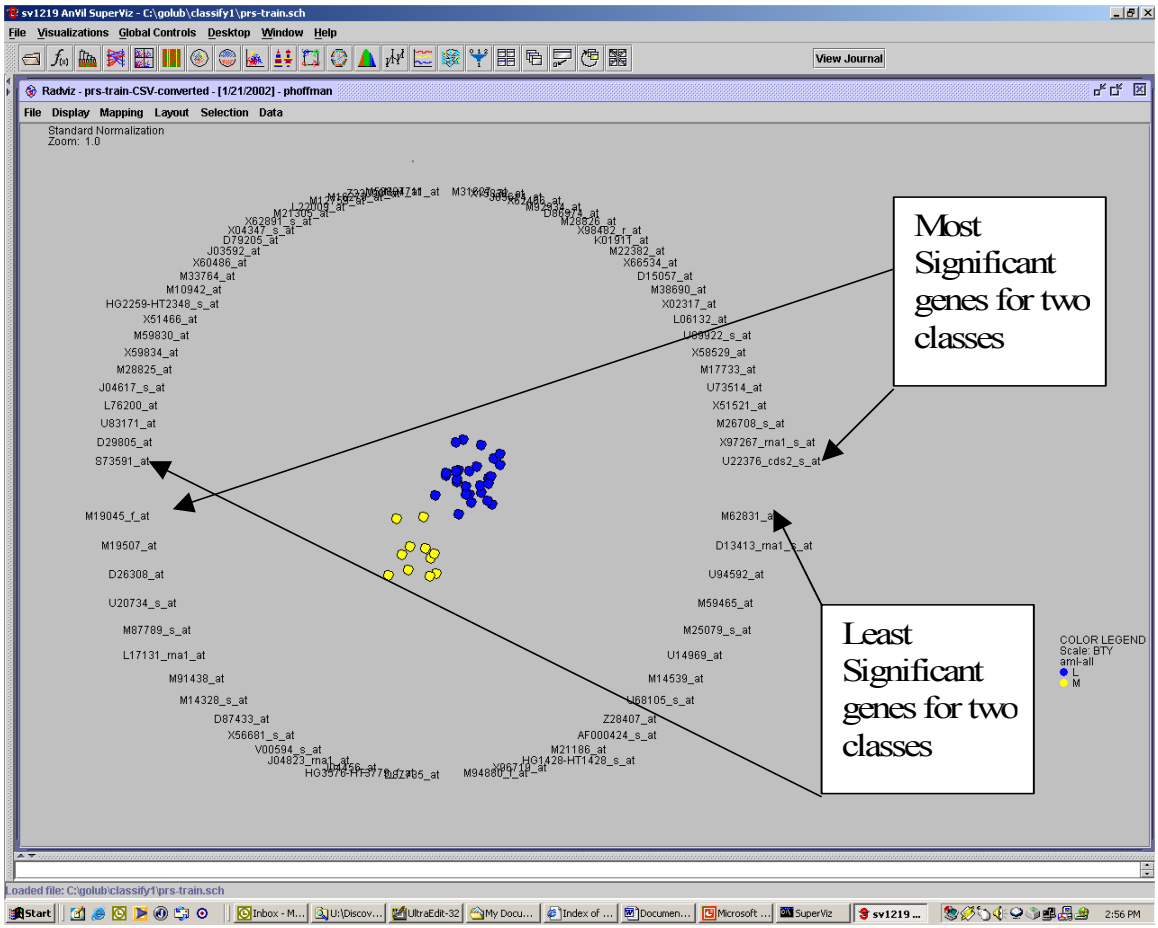


Figure 9 – genes distinguishing 6 tissues



A layout using t-statistic with equal variance

The t-statistic is calculated for each column (gene) for all the ALL (L) values comparing with all the AML (M) values in that column. The t-statistic is a standard statistical test comparing two groups using the means and standard deviations. The t-statistic for each column determines the order of the columns around the RadViz perimeter.

The genes or columns that have higher values for ALL (L) are layed out in the top half of the RadViz circle, the genes or columns that have a higher values for AML (M) are layed out in the bottom half of the RadViz circle. The order of the genes are by t-statistic values and in the top half, the genes are ordered right to left. In the bottom half the genes are ordered with significance going from left to right.

The columns (genes) are layed out around the RadViz perimeter with the column that has the highest t-statistic (negative) value at about 3 o'clock in the diagram. U22376_cds2_s_at is the gene that is most significant for having a higher mean for ALL (L) than AML (M). Gene or column M19045_f_at is the gene that is most significant for having higher values for AML (M) than ALL(L) and it is placed about 9 O'clock in the bottom semi-circle.

Association Rules:

In addition to a visual classifier RadViz can also be used to visually find/and or verify association rules. In **Figure 10** a RadViz layout is used find which calculated chemical descriptors can help predict active or inactive chemicals. 960 chemicals are displayed with 11 calculated chemical properties. R1 through R4 are different chemical subgroups attached to a fixed ring structure. The layout uses the t-statistic but it can be seen that the lift (increased percentage) for the active chemicals in the lower hemisphere is only marginal. In the layout in **Figure 10** the R1-R4 subgroups are simply assigned consecutive numbers for the RadViz spring layout.

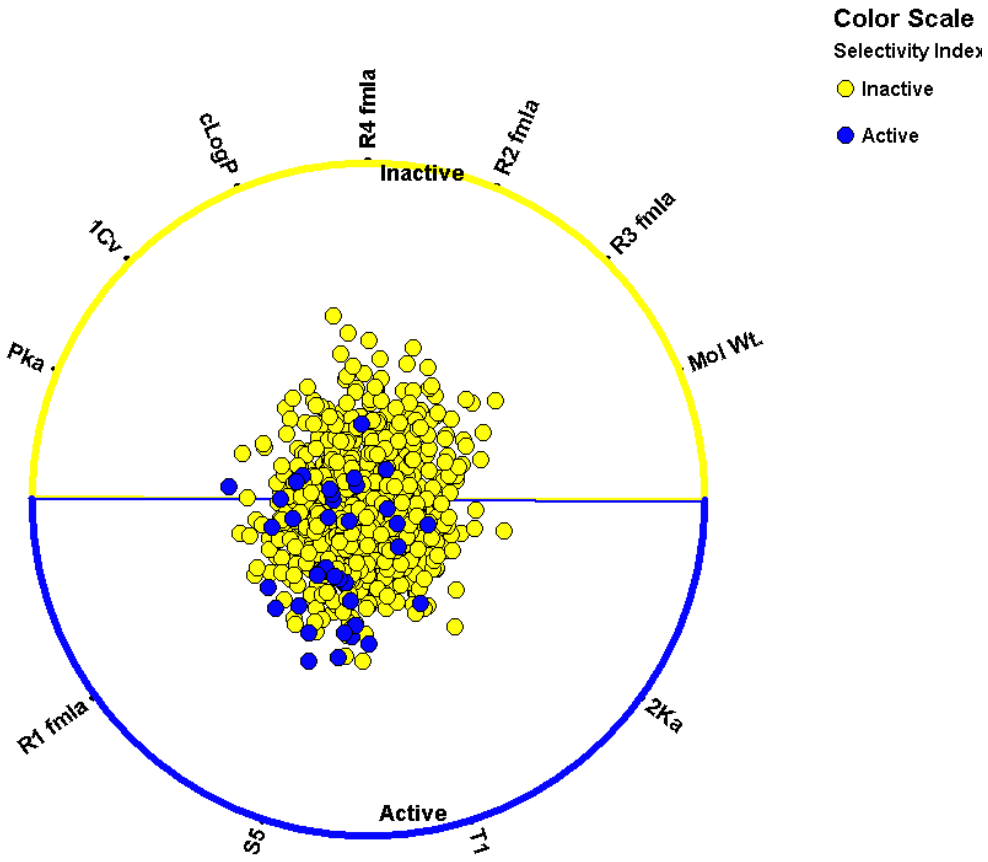


Figure 10 - Chemical descriptors predicting Active or Inactive compounds

A better numerical assignment of the R-groups for machine learning is to expand the 4 R-group columns to one column for each possible value making the value in the column a 1 for the particular R-group and a 0 for all other values. This is normally called

"flattening". The number of R-group descriptors is now increased to 25 instead of the original 4 and the class separation is better (See **Figure 11)**

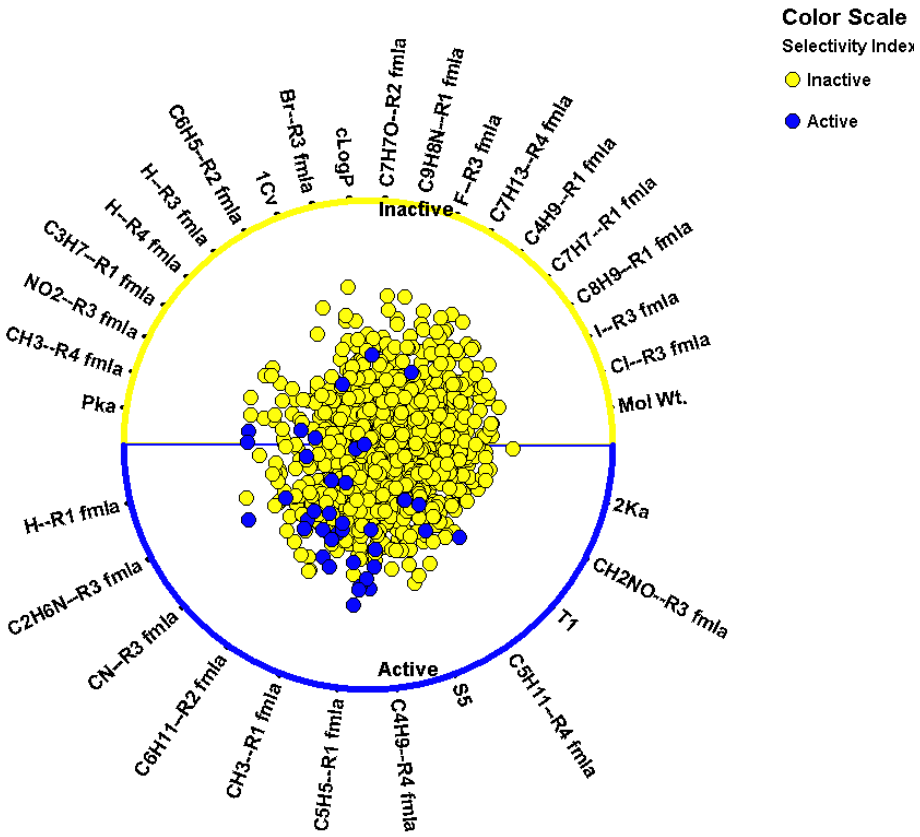


Figure 11 - Layout from increased chemical descriptors after flattening

In RadViz data points where all dimensions are 0 except for one are plotted exactly on the anchor point for that dimension on the circumference of the RadViz circle. If we just use the flattened R3 groups we see this in **Figure 12.** The data points have been jittered (slight random moves) so that overlapping points can be shown. Also a negative zoom factor is used to move the points in slightly from the edge of the circle. It can be seen that there are more "active (blue-dark)" chemicals in the R3 groups CN and C2H6N then in any other R3 groups.

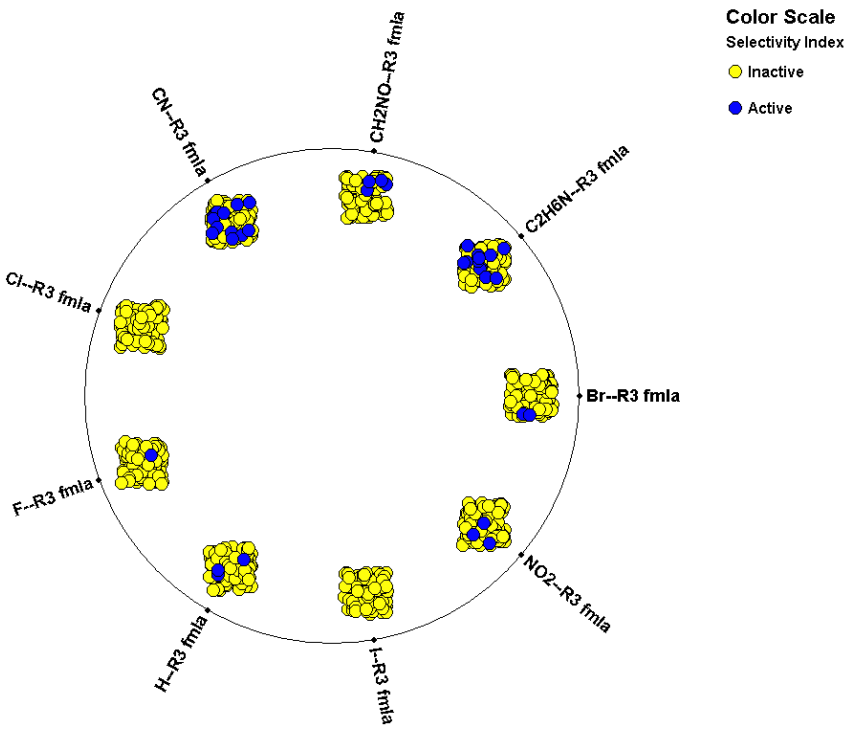


Figure 12 Showing just the R3 group values (Flattenned, jittered, and zoomed)

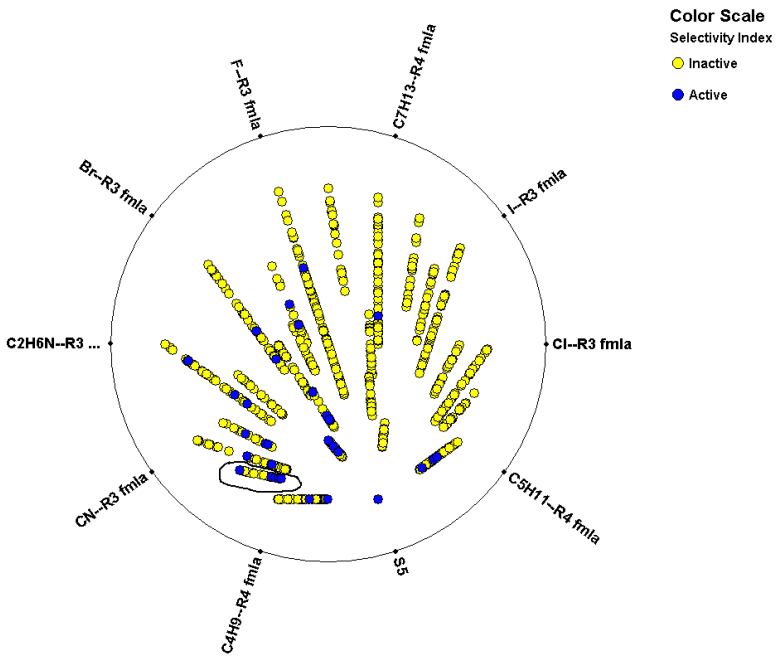


Figure 13 R3, R4 groups and the continuous variable S5

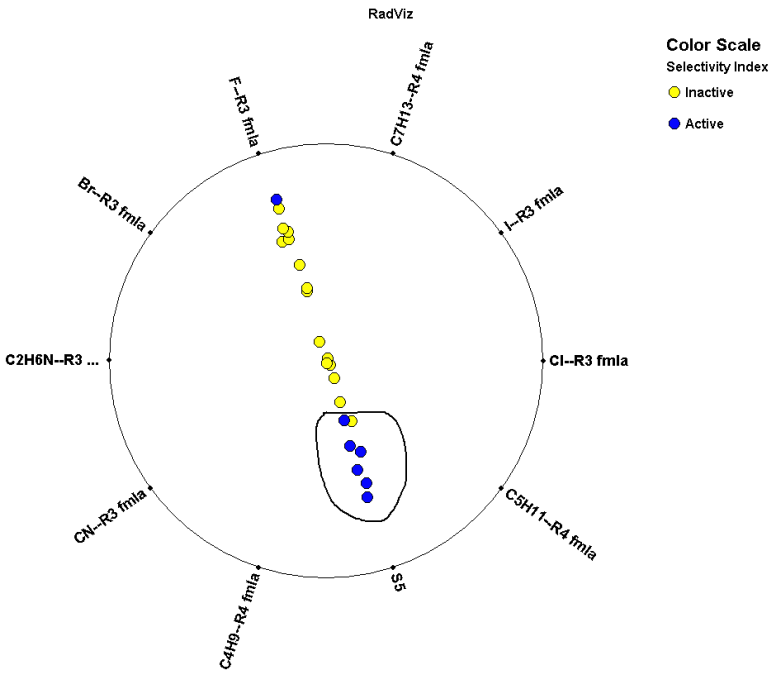


Figure 14 Association rule R3, R4 groups and the continuous variable S5

That S5 > 4.82 AND R3 = CN AND R4 = C4H9 have 5 Active candidates with 100% confidence.

Clustering PCA first, then RadViz"

RadViz inherently clusters, but combined with PCA

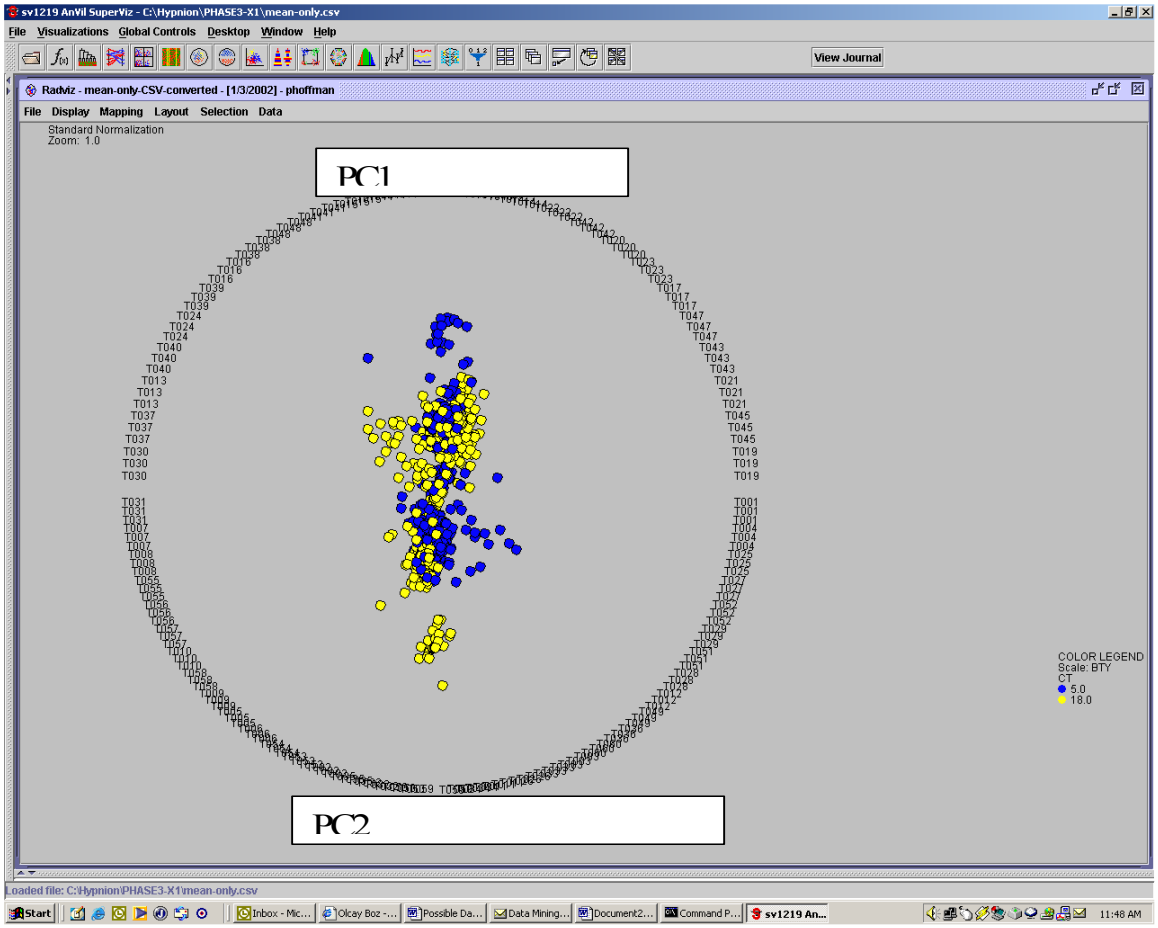
Principal Component – RadViz mapping.

One of the problems with PCA is that the coefficients of the "real" dimensions making up the PC's are not usually shown, thus being somewhat a "black box" clusterer.

One could do a PCA and then use the coefficients of each PC for each "real"dimension as weights in a Radviz layout. The coefficients-weights would be sorted Hi to Low , with negative coefficients on the opposite side or RadViz enhanced to use negative weights.

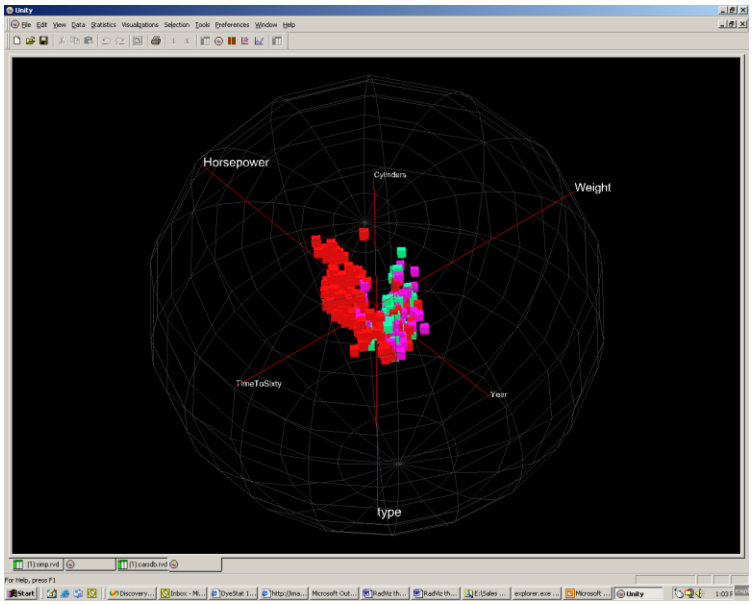
Thus if one wanted to "see" 3 PC's one would have 3 groups of all the real dimensions around the radviz layout, but with spring forces equivalent to the PC coefficients. This should give clustering similar to a 3D PCA analysis, but with the

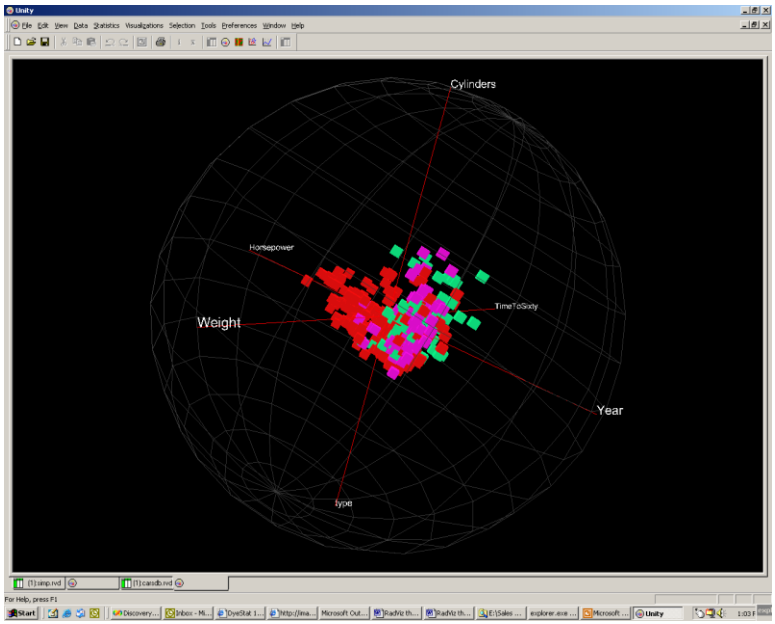
dimensions layed out and shown in the “importance” order. This should help understanding PCA clustering better.



PC layout example dimensions would be duplicated for each PC, and ordered according to size of coefficients. Weights would be equal to coefficients.

3 Dimensional RadViz





Cycles in RadViz

RadViz with Relevance Lines