10

Neural-Network Enhanced Visualization of High-Dimensional Data

Urska Cvek[1], Marjan Trutschl[1] and John Clifford[2]
[1]*Louisiana State University, Shreveport, LA*
[2]*Louisiana State University Health Sciences Center, Shreveport LA*
USA

1. Introduction

Large quantities of multivariate data generated in scientific, engineering, business and other fields triggered exciting developments in information visualization, data mining and knowledge discovery with the objective to identify and describe properties of data sets. Self-organizing map (SOM) is an unsupervised neural network technique capable of analyzing large and complex multivariate data. It attempts to address the problems of high-dimensional data and identify the underlying patterns by reducing the dimensionality achieved through grouping of similar objects and mapping them to a low-dimensional space, usually to a two-dimensional surface also known as a topological map. The results of a SOM could be misinterpreted if taken out of context. For example, the distance between neighbouring weight vectors does not correspond to the physical location of those vectors on the matrix of output nodes as described by Ultsch (Ultsch & Vetter, 1994). The widespread use of the algorithm is attributed to its simplicity. The analytic and graphical Kohonen SOMs and its variations have been successfully applied to the analysis of complex, large-dimensional data sets from diverse sources, including biomedical data, such as by (Durbin & Mitchison, 1990; Tamayo et al., 1999; Van Osdol et al., 1994), just to name a few. The quest for effective and efficient visualization techniques capable of displaying large numbers of high-dimensional records has been formally underway since 1987 when the National Science Foundation sponsored a workshop on Visualization in Scientific Computing as Wong and Bergeron pointed out (Wong & Bergeron, 1997). The problem dates back to the first graphical representation of various types of data sets. Information visualization, as the field is often named, is summarized by the transformation of data – in whatever form – into pictures, with pictures being interpreted by a human being (Spence, 2007). The main advantage of visual displays is the ability to harness the human perceptual system, improving over tabular or other data representation forms.

We utilize the benefits of the SOM and visual displays through a linked technique that integrates the SOM with two- and three-dimensional information visualization techniques (a.k.a., iNNfovis), which serves as a model for constrained self-organization. Properties of iNNfovis environments are harnessed through interactive analysis of large data sets for non-trivial feature extraction. Various iNNfovis configurations provide unique environments for

clustering and exploration of high-dimensional data. Additionally, large-dimensional data mapped to a low-dimensional surface is poised to result in perceptual ambiguities, such as overlap or occlusion, which occurs when the number of records exceeds the number of physical points in a visualization of a certain size. iNNfovis techniques were successfully applied to overcome the effect of occlusion or overplotting by harnessing dimensional information that provides local spatial organization while maintaining a relatively accurate location in a low-dimensional space.

This chapter is organized as follows: Section 2 first provides an overview of a couple of classic visualization techniques used for visualizing multidimensional and multivariate data: scatter plots and RadViz (lossless projection). In Section 3 we discuss the problem of overlap or occlusion in visualizations and how we address it with our technique. Section 4 presents the constrained self-organization technique applied to scatter the plot and RadViz displays. In Section 5 we analyze the transitional cell carcinoma of the bladder using these techniques. We conclude with Section 6.

2. Graphical Representation of Data

Visualization is increasingly used in the data exploration process. In its early years it was mostly, if not only, used to convey the results of statistical computation or data mining algorithms (Chambers et al., 1983; Tukey, 1977; Cleveland & McGill, 1984). Over the last twenty years, its use spread from the exploratory and confirmatory role to the data cleaning process, certain aspects of the data management process and computational steering processes within the data exploration pipeline. There exist numerous data visualization techniques and taxonomies, but the most common data visualization techniques used today are scatter plots, pie charts and line plots. RadViz, star coordinates, polar charts and parallel coordinates are a few of the high-dimensional techniques commonly used by data analysis specialist. Scatter plots, scatter plot matrices, RadViz and the SOM are all projections onto a two-dimensional surface that suffers from occlusion or overlap problem.

2.1 Scatter plot

Scatter plot is a point projection of the data onto a two-dimensional surface or into a three-dimensional space represented on the screen in a classic (x, y) or (x, y, z) format, respectively. Scatter plots can display a large number of points, according to some analysis up to 100,000 points or more, depending on the data pattern (Eick, 2000). Displayed points can have numerous attributes such as color, size, shape, texture, motion and even sound (when interacted with). To interpret the three-dimensional projection of data, it is necessary to resolve ambiguities, although other techniques such as jitter and animation have been used (Chambers, at el., 1983; Chambers & Hastie, 1992). In its most general form this method is related to iconographic and pixel displays. Figure 1 displays the type of Iris flower (Fisher, 1936) as a two-dimensional scatter plot. We project the sepal length dimension as the x axis, and sepal width dimension as the y axis.

Scatter plots reveal a lot of information about the relationship of two variables (distribution of points, outliers, modes, association), but the aspect ratio of the axes, the size and color of points affect their appearance (Ultsch & Vetter, 1994). Scatter plots provide for limited visual scalability, or capability to effectively display large data sets, in terms of the number of elements of the dimension of individual data elements (Eick, 2000). The primary limiting

factor is point overlap: as the number of data instances increases, point over plotting causes increased occlusion of trends and concentration of points, as well as limited access to and hiding of individual instances, causing possible misinterpretation. Approaches that attempt to solve the visual scalability problem include interactivity (focus+context method, panning and zooming, identification and selection, automatic aggregation and brushing), jittering and density estimation. Jittering increases visual scalability by providing access to individual instances, but lacks an insight into the exact relationship among the instances that have been overplotted (Eick, 2000).

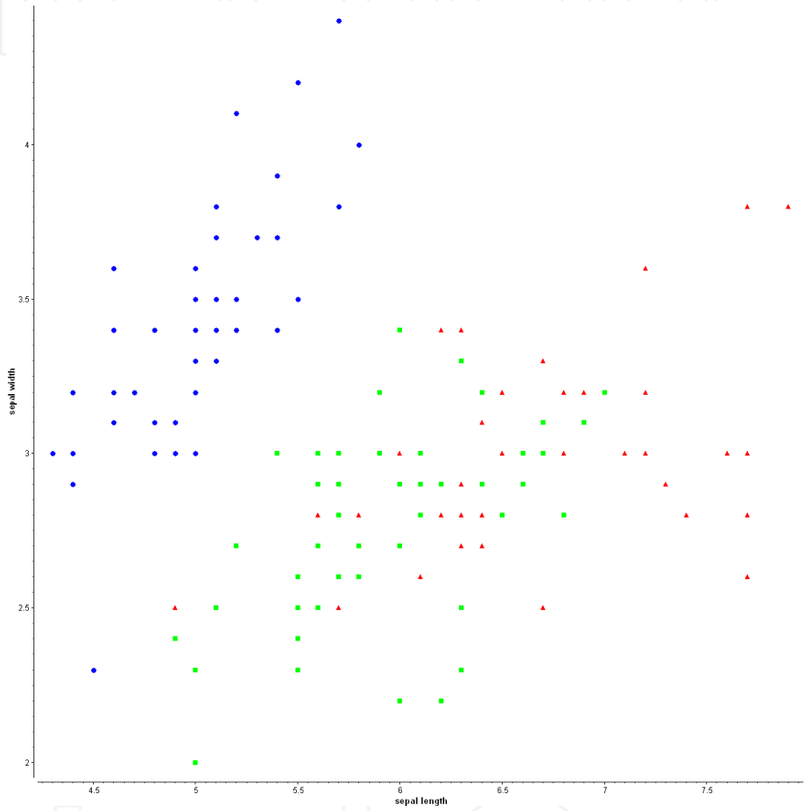


Fig. 1. Two-dimensional scatter plot maps sepal length and sepal width dimensions of the Iris data set

2.2 RadViz

RadViz (Hoffman et al., 1997) is a visualization technique that places dimensional anchors (dimensions) around the perimeter of a circle. Each record is represented as a vector $x_{i1},...,x_{im}$ on these m dimensions and its location in the visualization is determined by the pull of the position vectors (dimensions) $\overline{S}_{1,...,}\overline{S}_m$. Each position vector points from (*0*, *0*) to the corresponding fixed point (dimensional anchor) on the perimeter of the unit circle. Each data point is displayed at the point where the sum of all spring forces equals zero (1).

$$\sum_{j=1,m}\left(S_j - u_i\right)x_{ij} = 0 \tag{1}$$

The position of the data point depends largely on the arrangement of dimensions around the circle; however, data items with similar dimensional values are always placed closer together. The technique has been used in several data domains including biomedical and complemented with dimension ordering techniques, where the dimension order is determined by the structure of the data, or the inherent class separation (Leban et al., 2005; Au et al., 2000; Grinstein et al, 2001; Bertini et al., 2005). We include an example of the technique as applied to the Iris data set, displaying the four dimensions of the data on the circle, and colouring the records by the flower type (Figure 2).

A drawback of the method is overlap of records, that can not only be caused by records that have identical values on the dimensions, but also by records whose sum of all spring forces equals zero (such records are placed in the center of the RadViz unit circle). Overlap of points also occurs when the records are scaled. For example, records (1,1,1,1,1,1) and (10,10,10,10,10,10) would appear at the same location in the center of the circle (they are pulled by all dimensions equally). Records (1,10,1,1,1,1) and (10,100,10,10,10,10) would also appear at the same location. We can encounter instances whose n dimensions have different values, but the sum of their n dimensions is equal to 0. The interpretation of these types of overlaps is more difficult. Dimension ordering and placement of dimensions away from the radial layout minimizes this problem, but does not completely solve it. We developed an approach that utilizes the third dimension to organize the data when overlap occurs to aid occlusion.

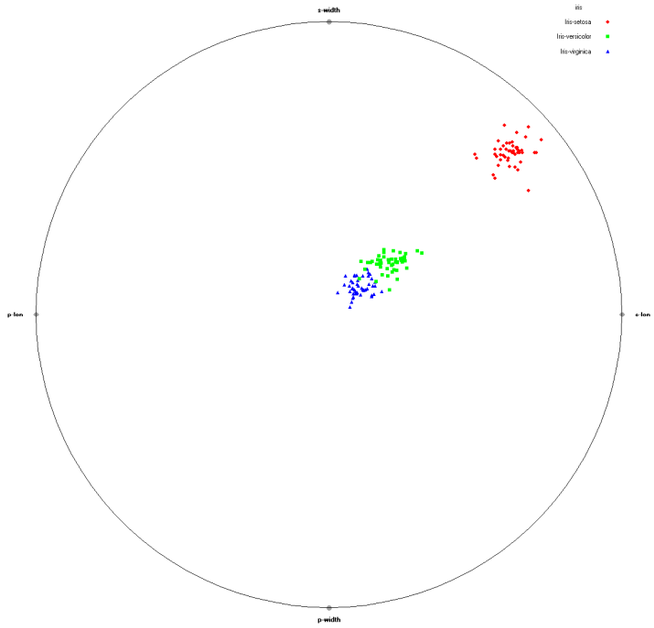


Fig. 2. RadViz visualization of the Iris data set

3. Occlusion or Overplotting in Information Visualization

Large and high-dimensional data sets mapped to low-dimensional visual spaces often result in perceptual ambiguities. One such ambiguity is overlap or occlusion that occurs when the number of records exceeds the number of unique locations in the presentation or when there

exist two or more records that map to the same location. Occlusion in three-dimensional spaces has been extensively studied and approached similarly in two-dimensions (Wong & Bergeron, 1997; Eick, 2000). In these spaces, occlusion is largely due to the placement of records behind the object the user is viewing. The most common approach to resolve it is semi-transparency, where the object is made translucent and additional objects are seen through (Zhai et al., 1996). Overplotted records may also be represented using glyphs where the size of a glyph indicates the number of records that map to the same location (Carr, 1991). In the physical space, where the projection is mostly two-dimensional, transparency of a single object provides only limited occlusion resolution. Another approach is to allow the user to manipulate data, in order to discover additional relationships, such as in the Selective Dynamic Manipulation system (Chuah et al., 1995). Excentric labelling was devised to provide interactive labels for a selection of records by moving the cursor over the records, providing an insight into the values behind these records (Fekete & Plaisant, 1999). Additional techniques were introduced by Manson (Manson, 1999), who utilized retinal properties (size, color and shape), secondary point properties (point border) and animation. The most common approach, used in several commercial packages, is to resolve occlusion using "jittering," displacing overlapping records over a wider area.

Therefore, overlap or occlusion occurs when a data set of n-dimensional records is mapped to an m-dimensional visualization space, where $m < n$ or even $m << n$. Overlap Ω occurs when the number of records r in a data set exceeds the number of physical points in a visualization of size v_x by v_y (in case of a rectangular two-dimensional visualization).

$$\Omega : if\ r > v_x \cdot v_y \tag{2}$$

Overlap may also occur when there are at least two records r_i and r_j that are not unique with respect to their dimensional values x and y or their non-linearly projected x and y positions.

$$\Omega : if\ \exists \left(r_{i_x} = r_{j_x} \wedge r_{i_y} = r_{j_y} \right) \tag{3}$$

Both identities may be modified to reflect the properties of a 3-dimensional overlap.

$$\Omega : if\ r > v_x \cdot v_y \cdot v_z \text{ and } \Omega : if\ \exists \left(r_{i_x} = r_{j_x} \wedge r_{i_y} = r_{j_y} \wedge r_{i_z} = r_{j_z} \right) \tag{4}$$

Overlap can also occur in scatter plots when the two presented dimensions have identical values. Moreover, visualizations based on non-linear projections from an n-dimensional space to an m-dimensional display (such as RadViz and Star Coordinates) most often result in multiple overlapping records regardless of the fact that their dimensional values are unique. All visualizations exhibit similar behavior when analyzing large, high-dimensional data sets. In general, they fail to handle overlapping records and crowding, when mapping numerous instances to a limited display space.

If we assume that we have a two-dimensional visualization, a record maps to position (x, y) on the display. The most common technique to solve the overlap problem is "jittering," which offsets the record from its mapped location (x, y) on the two-dimensional output surface to location (x', y') with $x' = x \pm \Delta x$ and $y' = y \pm \Delta y$, Δx and Δy representing randomly generated offset distances. If more than two points map to the same (x, y) location, the jitter

algorithm randomly generates Δx and Δy for each of the overlapping records, keeping the Δx and Δy within a predefined range. Chambers (Chambers et al., 1983) described jitter in scatter plots that adds random noise to one or both of the variables using two sets of equally spaced values from -1 to 1 and utilizing fractions of the variable range to calculate the offset. Cleveland and McGill (Cleveland & McGill, 1984) and later Cleveland (Cleveland, 1993) described jittering as adding small random variables, in addition to moving points, transformations, open circles and sunflowers as approaches to address overlap.

A more advanced jittering algorithm may also keep track of each jittered point, minimizing possible new overlaps, since it is statistically possible for a random number generator to produce identical offsets Δx and Δy for two or more overlapping points. The amount of jitter can follow a distribution, such as a normal distribution. However, if the number of records is greater than the number of unique locations within the predefined jittering surface of size $\Delta x_{range} \cdot \Delta y_{range}$, the jittering process unavoidably results in new overlaps. Such handling of overlapping records is considered efficient, if the goal is strictly to show the number of records mapping to a particular area, but fails to provide an insight into the exact relationships among the instances that have been overplotted. The main weakness of such jitter technique is spatial displacement that is not driven by the data, causing difficulties in interpretation.

To emphasize the overlap effect, we present overplotting examples of a modified Fisher Iris flower data set that contains four dimensions and the flower type (Fisher, 1936). In this modification, we round the fractional part of sepal length and sepal width dimensional values to the nearest value, and leave petal length and petal width dimensions unchanged. The data set is interesting because in the original data set we cannot find a clear boundary between the three types of flowers when using two pairs of dimensions at a time. Petal width and petal length are very closely related dimensions with only limited correlation existing between those two dimensions and sepal length. In all of our images, we colour Iris *setosa* records red, Iris *versicolor* green and Iris *virginica* blue. Figure 3 is a parallel coordinate display (Inselberg & Dimsdale, 1990) showing the modified set. The result is a reduced number of unique values, leading to multiple overlapping points.

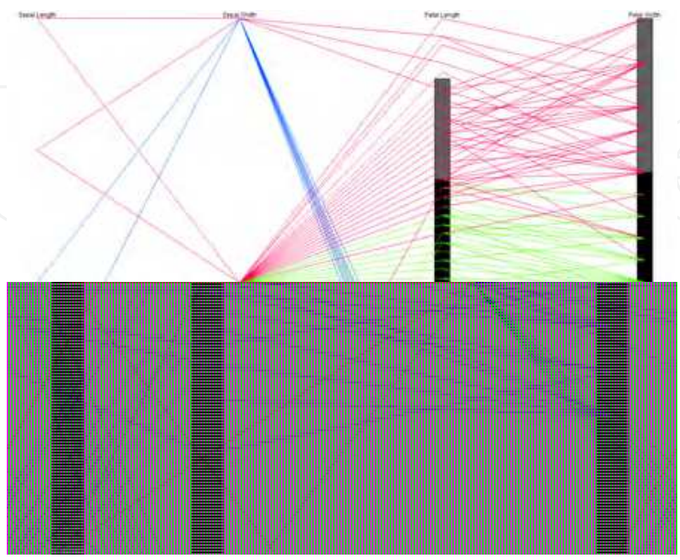


Fig. 3. Parallel coordinates of the modified Iris flower data set

The scatter plot display in Figure 4 uses sepal length and sepal with as the x and y axes, respectively. The visualization contains only eleven points, representing 150 instances, with the remaining 139 instances overlapping at these eleven locations. The overlap is present because the sepal length dimension contains five unique values while the sepal width dimension contains three unique values. This visualization provides little insight into the number of records at each of the locations, or the relationships between individual records. Additionally, the color (type of flower) of each of the points reflects only the last record that was plotted at a location. These records need to be spatially reorganized or jittered in order to address these problems. When the records are randomly jittered, the result displays all or most of the 150 records (Figure 5).

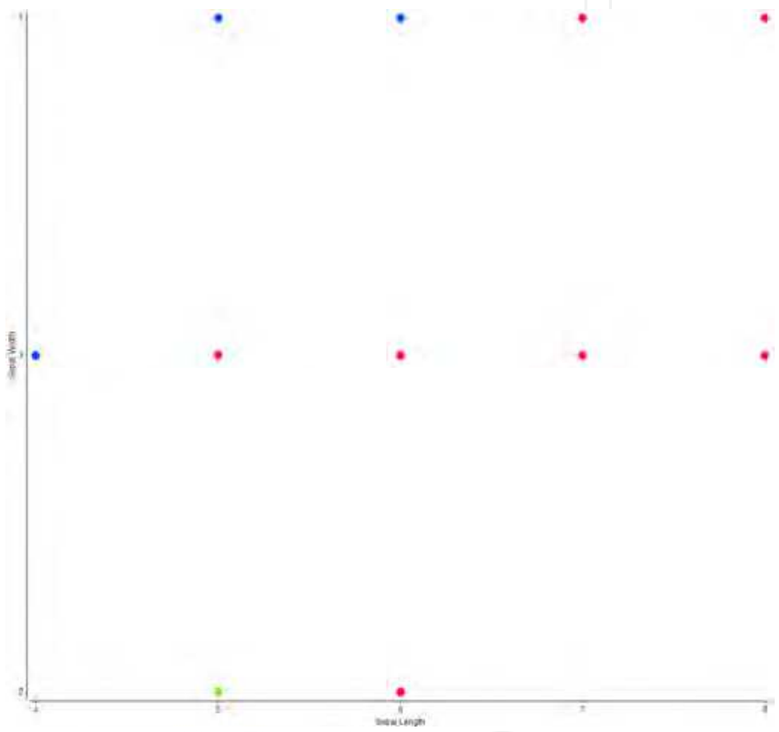


Fig. 4. Scatter plot of the modified Fisher Iris flower; sepal length and sepal width mapped to x and y coordinates. Occlusion results in only one type of flower being displayed at each location.

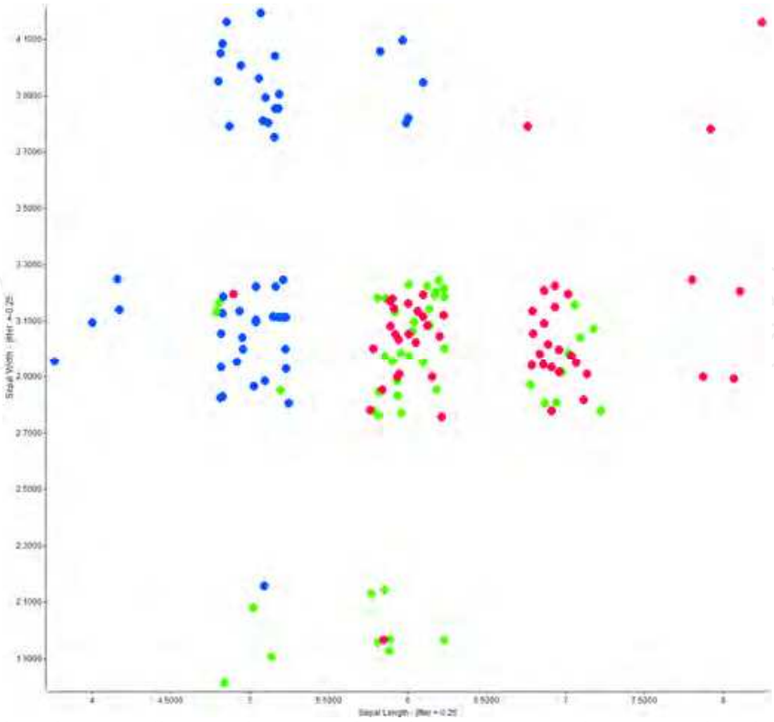


Fig. 5. Modified Fisher Iris flower data set. Overlapping records are randomly jittered to alleviate occlusion. Records jittered between -0.25 and 0.25.

4. Integration of SOM with Classic Data Visualizations: iNNfovis

We utilize the benefits of data visualizations and combine them with the mapping power of the SOM. This combined technique is not a novel way to determine the (x, y) position for a record but rather a refinement method for spatial organization of overplotted input vectors. In general, our approach can be applied to reduce occlusion given any record placement strategy (method to find the x, y position) on a two-dimensional or three-dimensional display. We could use any dimension pair or mapping onto the x and y dimensions, or any other mapping onto a two-dimensional plane or three-dimensional surface. The algorithm harnesses dimensional information of an input vector to provide local spatial organization while maintaining a relatively accurate x and y location on the surface.

The algorithm maps input vectors with similar properties to the same or neighbouring output nodes of the display. Input vectors located in proximity of each other are likely to correlate more than vectors located farther apart. The correlation factor depends on the weight vectors of neighbouring output nodes.

Displacement around a (x, y) position is driven by the dimensions of the data using the SOM algorithm. This creates a constrained clustering environment, a type of unsupervised clustering that is identified by the chosen projection onto the x and y or x, y, z position (three-dimensional) of the output matrix, primary mapping, secondary mapping and the combined SOM-visualization neighbourhood of qualifying output nodes. These constraints differ from constraints in supervised clustering, where the constraints are identified by the outcome of the training process.

4.1 SOM - Scatter plot

We combine SOM with the scatter plot by binning the *x*- and *y-axes* and applying the SOM to this new output grid. Binning is a summarization technique most applicable to large volumes of data and can reduce the number of records to be plotted. Even a scatter plot, for instance, could be treated as an enormous grid of bins, although it is seldom viewed that way. To reflect the properties of the data set, the actual number of bins needed to uniquely map n two-dimensional or three-dimensional vectors is determined based on the range ρ and the number of decimal places δ of data values shown in Equation 5.

$$Bins_{2D} = \left(\rho_x \cdot 10^{\delta_x}\right)\left(\rho_y \cdot 10^{\delta_y}\right) \tag{5}$$

The display surface is replaced with two grids; a secondary output grid within a primary output grid. The grid sizes (or resolution) are specifiable, or can be data-driven, based on the distribution and number of overplotted points, or on the overall number of displayed records.

In Figure 6 we show how we grid the surface of the scatter plot. At the bottom of Figure 6 are the n dimensions of a sample data set, with dimensions 1 and 4 selected as the x and y dimensions. The top of the Figure is the scatter plot grid. Each primary output node on a primary grid contains a set of secondary output nodes, or a grid within a grid. In this example each primary output node contains a secondary output grid of 25 nodes, shaped as a square of 5x5 secondary nodes. We first perform *primary mapping* onto the primary output grid, followed by the secondary mapping. We map a record onto the primary output node *Wp* as determined by the record's values of dimensions 1 and 4. Self-organization is repeated for every primary output node within the grid. *Secondary mapping* first randomly initializes weight vectors of each secondary output node in the primary output node. The distance between an input vector and each output weight vector in this secondary grid is calculated, and the winning secondary output node W_S is determined based on the smallest Euclidean distance. The record is mapped into that winning node W_S (Figure 7) and the weight vector of the winning node adjusted, in addition to limited functional adjustment of the neighbouring secondary weight vectors, depending on the neighbourhood function. Adjustment of weights and self-organization is not limited to a single primary grid, but is rather driven by the neighbourhood function and the properties of the records. This process repeats for every primary output node, in successive training passes through the input data set. The result is a self-organized scatter plot display that addresses the problem of over plotting.

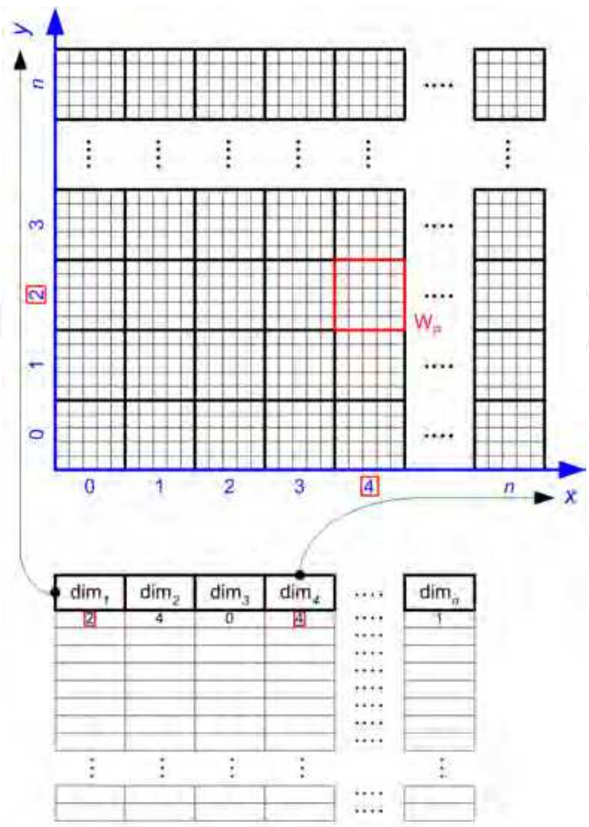


Fig. 6. Secondary output grid within a primary output grid; W_P is one primary output node

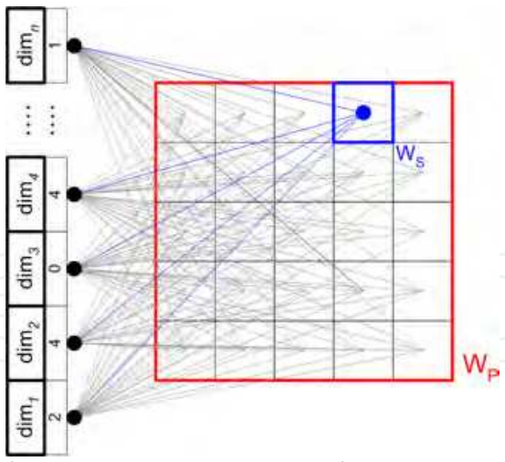


Fig. 7. Mapping of a record into a primary output node

We can similarly extend the two-dimensional scatter plot into the third dimension. We first find the position in the primary grid (*R'*) as we have described, and then select a secondary node in the third dimension into which this record belongs (*R''*). Figure 8 shows the secondary output nodes available for the mapping. The movement of the record is constrained to the vertical stack of secondary output nodes at the primary grid cell. Each of these secondary output nodes is associated with a weight vector. After the winning

secondary output node is found, a neighbourhood function adjusts the neighbourhood nodes not only in the vertical stack of secondary nodes at the current primary output node, but also for the neighbouring output nodes in the three-dimensional cube, effectively utilizing a three-dimensional Gaussian neighbourhood function. This process is repeated for each of the records in the data set, for a selected number of training epochs.

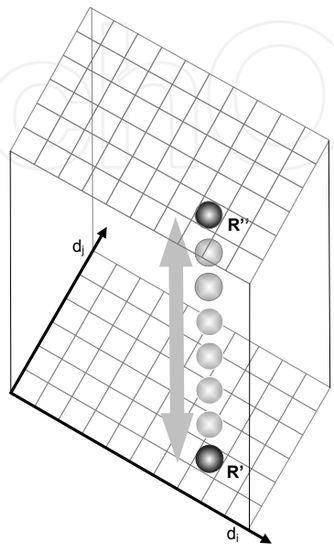


Fig. 8. Secondary mapping into the third dimension (*R''*) from the original position (*R'*) in the primary grid

Figure 9 is a SOM-scatter plot display of the modified Iris data set using a 5x5 primary and 10x10 secondary grid. Each input vector is first placed on the primary grid as determined by the value of sepal length and sepal width dimensions and boundaries of the primary grid of the range of values. Each primary output node contains a 10x10 secondary grid. These 100 secondary output nodes are initialized with random weight vectors, which are adjusted with every input vector mapping. Each input vector is first mapped to the primary output node, and within it into a secondary output node based on the shortest Euclidean distance between an input vector and the weight vectors of the 100 secondary output nodes. The algorithm first maps every input vector until the data set is depleted, and then repeats the training process until the target state is reached. Figure 10 shows a central section of Figure 9 and displaying the lines of separation among the secondary output nodes. These lines are a visual tool for identification of distances among the neighbouring output nodes and extend the U-Matrix (Ultsch, 1994) approach by displaying the records mapping to output nodes combined with inter-nodal distance representations. This feature is a recommended extension for dense primary output grids with large secondary output grid, which take up most of the white space that is otherwise used as an indicator of grid edges. As the lines of separation indicate, it is possible that there is a large difference between two neighbouring secondary output nodes within the same primary output node.

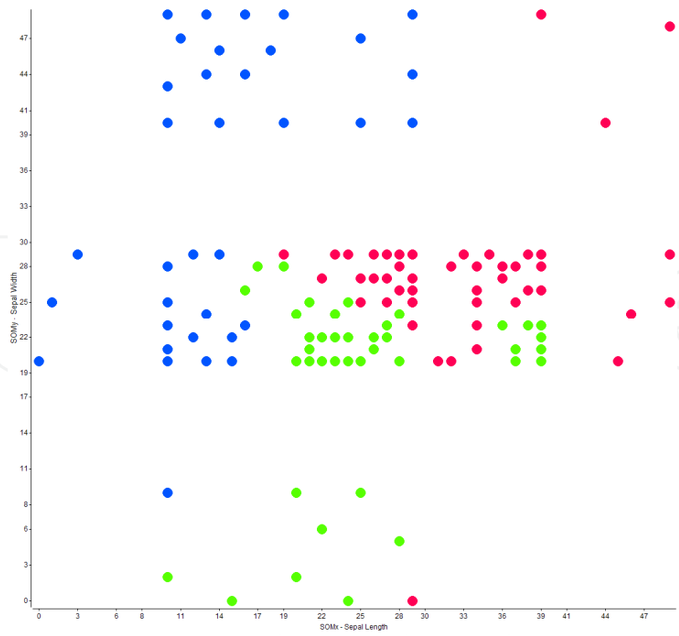


Fig. 9. SOM – Scatter plot visualization of the modified Iris flower data set with 5x5 primary and 10x10 secondary output grid

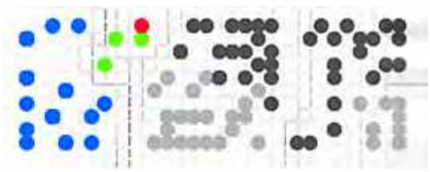
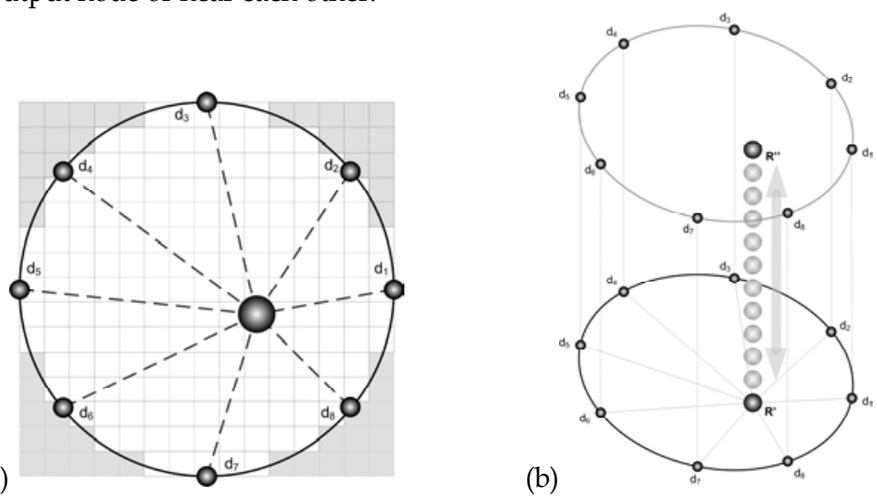


Fig. 10. Distance between neighbouring output nodes as indicated by the lines of separation. The records shown in colour were selected for further analysis (non-selected records are shown in grey colour).

Records in Figure 10 represent all three flower types. We selected them guided by the lines of separation and selected one *setosa*, three *versicolor* and additional *virginica* records (in color). Parallel coordinates would show that these records differ only on petal length and petal width dimensions. Their values on the sepal length and sepal width dimension are the same, placing them in the same primary output node. SOM-scatter plot separated most of the *virginica* (blue) records from the other four records. Two records within the same primary output node are not necessarily more similar than two records that belong to two different primary output nodes. The *setosa* record in red is more similar to the records in the neighbouring primary output node (grey) than to other records in its primary output node. We also demonstrate that pairs of instances spatially equally distant on the output grid are not necessarily equally similar. The lines of separation show that the weight vectors of secondary output nodes of *virginica* (blue) records are much more similar than the weight vectors of secondary output nodes with *versicolor* and *setosa* records.

4.2 SOM - RadViz

We start the incorporation of the SOM technique with RadViz by binning the surface of the unit circle setup for the RadViz visualization. We determine the position on this primary output grid by applying the RadViz algorithm and selecting the primary output node *R'*, which lies at the position of the RadViz mapping. The user can select the number of grid rows and columns for this RadViz display (see Figure 11), which determines the number of grid nodes. Additionally, we associate each primary output node in this unit-circle grid with a stack of secondary output nodes, to project the input vectors into the third dimension. An identical approach is taken in the three-dimensional scatter plot and this technique significantly increases the intrinsic dimensionality of a RadViz visualization. Our algorithm is capaable of handling the input vectors that exhibit the same profiles that differ in magnitude (i.e., records {1,2,1,2} and {5,6,5,6}). The original RadViz algorithm maps such vectors to the same location thus introducing ambiguity and overlap. Although SOM-RadViz algorithm also places such records at the same (x, y) location, their position on the z-axis differs, where only records with the same or very similar profile are placed in the same secondary output node or near each other.



(a) (b)

Fig. 11. (a) RadViz mapping with an overlaid 15x15 primary grid; (b) SOM-RadViz mapping into the third dimension

Figure 11 shows this process of first calculating the position of the record in the RadViz display, followed by the "pull" of the record into the third dimension. In this case, we are using twelve output nodes into which the record can map, based on its relationship to other records. The neighbourhood function has effect into the three dimensions, as well.

The modified Iris data set is shown in Figure 12. In this example, we are using a grided surface that replaces the unit circle drawn in the other RadViz displays (simplification). Using the complete four dimensions of this data set, we confirm that *versicolor* and *virginica* flowers aren't separable in two dimensions, but all the records of *setosa* flowers are different in their dimensional values. Please, note that this image is similar to Figure 2, but due to the different zoom of that image, the records appear more separated than in Figure 12. We try to identify other relationships among these records and proceed with the three-dimensional SOM-RadViz approach (right image in Figure 12). In SOM-RadViz, we use the RadViz visualization as the basis (at the floor) separated into 50x50 grid of cells, and then pull the

records into the third dimension of a stack of 50 output nodes at each of these grid locations. We utilize the fairly large grid on RadViz to approximate the true positions of the records as closely as possible. Again, we present it without the unit circle, and rather use the grid as a simplification. Although we previously confirmed that the *setosa* (red) flowers are fairly coherent, we can now detect that their dimensional values identify a structure within them as well. There are approximately four *setosa* records (at the top) that are separated from the rest. There is a cluster of *setosa* records that are not as similar and form a line of records in the center, followed by another set of records at the bottom. Similarly, we now discover that the *versicolor* (green) flowers are very similar and the *virginica* (blue) flowers are separated from them and dissimilar from each other. We reconfirm the relative similarity for the *versicolor* records on all the dimensional values, with some variability. *Virginica* records scatter and are pulled upwards, reflecting the self-organization that the records exhibit based on their dimensional values. This structure is otherwise not seen by any of the other displays – RadViz, scatterplot or SOM and is an excellent example of the power of the combination of the SOM with two-dimensional visualizations.

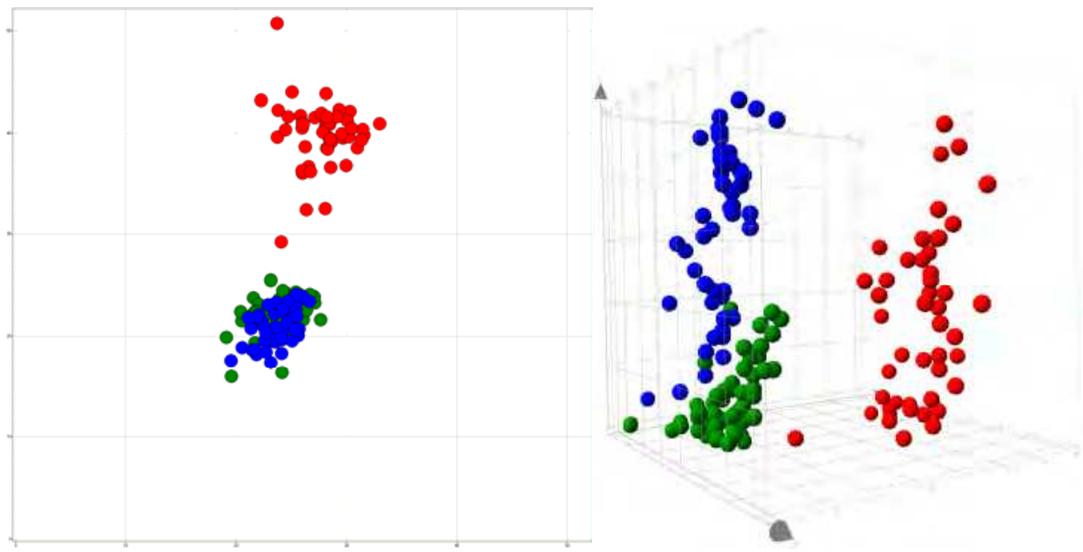


Fig. 12. SOM-RadViz display of the modified Iris data set. The left image is the two-dimensional RadViz projection (as in Fig. 2) and the right image is the three-dimensional SOM-RadViz.

5. Transitional Cell Carcinoma of the Bladder

We are using a data set derived from the analysis of transitional cell carcinoma (TCC) of the bladder generated by the Clifford Lab to demonstrate the application of these techniques (Stone et al.). Transitional cell carcinoma of the bladder ranks 4th in incidence of all cancers in the developed world, yet the mechanisms of its origin and progression remain poorly understood and there are few useful diagnostic or prognostic biomarkers for this disease. In an attempt to generate a mouse model for bladder cancer progression, investigators in the laboratory of Xue-Ru Wu engineered transgenic mice carrying a low copy number of the SV40 large T (SV40T) oncogene, expressed under the control of the bladder urothelium specific murine uroplakin II promoter (Zhang et al., 1999). These mice (called UPII-SV40Tag). develop a condition closely

resembling human carcinoma *in situ* (CIS), a pre-cancerous lesion, starting as early as 6 weeks of age. This condition eventually progresses to invasive TCC from 6 months of age onward. We have combined the UPII-SV40Tag mouse model for bladder cancer progression with Affymetrix DNA microarray technology. With the Mouse GeneChip (Mouse Genome 430 2.0) it is possible to determine a relative expression level of over 39,000 mouse transcripts (45,101 probe sets), representing the majority of the transcribed mouse genome, in a given mRNA sample. Duplicate arrays were performed for UPII-SV40Tag and non-transgenic wild type littermates (WT) at four time points during the course of tumor development, yielding a set of duplicated arrays for two factors: mouse genotype (WT or UPII-SV40Tag) and weeks of age (3, 6, 20, 30) creating eight targets. The WT line at the 6 week time point is the exception; due to the degradation of the RNA we only have one quality array. We followed the recommended analysis techniques (Gentleman & Huber, 2003; Gentleman & Carey, 2005) using R (R Development Core Team, 2008), bioconductor (Gentleman et al., 2004), and used the *limma* (Smyth, 2005) and *affy* packages (Gautier et al., 2004).

The 3, 6, 20 and 30 week time points were chosen to characterize the histologic progression from premalignant urothelium (cells that line the bladder wall), to CIS, to early invasive TCC and finally advanced invasive TCC, respectively, in the UPII-SV40Tag mice. We have identified approximately 1,900 unique genes (as identified by their Affymetrix Probe Set IDs) that are differentially expressed (≥ 3 fold difference at one or more time points) in urothelium between UPII-SVTag mice and their WT age-matched littermates. Preliminary biometric analysis using the Ingenuity Pathways Analysis software package (Ingenuity Systems Inc.) revealed that cell cycle regulatory, DNA replication, and cancer related genes were more strongly expressed in the UPII-SV40Tag urothelium at the highest proportion, even at the 3-week point.

Empirical Bayes method moderated *t*-statistic was used to test each individual contrast equal to zero and compute the moderated F-statistic which combines the *t*-statistic for all the contrasts into an overall test of significance for that gene. The *p*-values were adjusted for multiple testing using the method of Benjamini and Hochberg to control the false discovery rate. Tests were considered to be significant if the adjusted *p* value did not exceed 0.05. We eliminated the control probes from our analysis using the cutoff *p*-value and required at least a one-fold change between the arrays to consider them as differentially expressed.

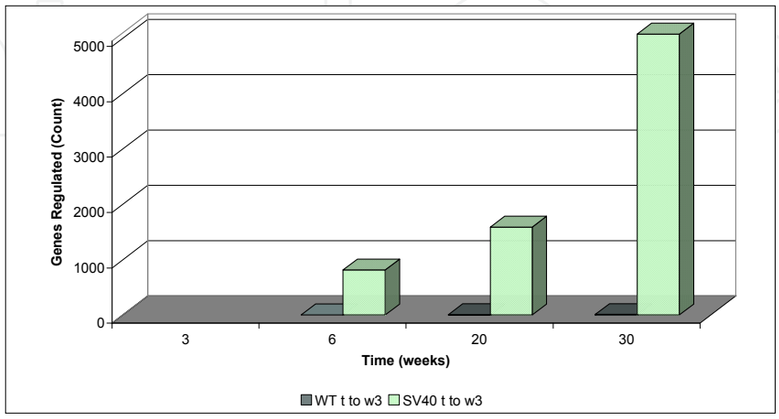


Fig. 13. Gene regulation at time points 6, 20 and 30 for the WT and UPII-SV40Tag lines. Differential expression is observed only in the UPII-SV40Tag line (green).

Figure 13 shows the number of genes that were increased or decreased in expression for each of the lines, when compared to the first time point for the WT. There is virtually no change in expression levels in the WT urothelium, while we observe an exponential increase in the number of differentially expressed genes for the UPII-SV40Tag line from approximately 1,300 to 2,100 and 4,400 at time points 6, 20 and 30, respectively. We next identified the number of genes differentially expressed between the WT and UPII-SV40Tag lines for each time point. Figure 14 shows the number of genes that are differentially expressed at each time point using the F-statistic with an additional requirement of a log fold change of 1.5 or greater. The increase of the number of differentially expressed genes is still apparent, and there are 17 genes that are exponentially changing in expression (either up or down) at every point from 3 weeks to 30 weeks (the intersection).

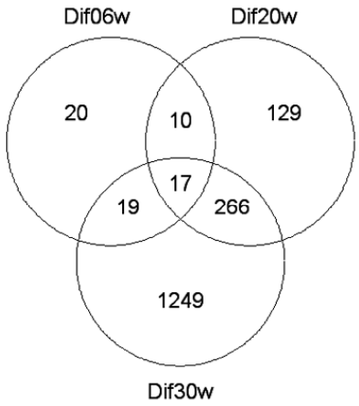


Fig. 14. Two-way analysis confirms the exponential increase in the number of regulated genes with time, comparing UPII-SV40Tag to WT mice at each time point.

A major goal of this project is the identification of biomarkers for early stage bladder TCC. We analyzed the set of 17 genes that change at every one of these time points, hypothesizing that changes at every time point are exacerbated with time progression. At the same time, we are interested in genes that are changing at the early stage of disease progression, namely the time points 6 and 20 (week 20 is classified as the early stage papillary TCC). We then selected a larger set of 585 genes that are differentially expressed at either of the two early stages (Figure 14).

We first analyze the RadViz visualization of this selected set of records. In Figure 15 and following figures we colour the upregulated genes *red* and downregulated genes *green*, following the standard gene expression nomenclature. All the fifteen dimensions of the data set were laid out on the unit circle, starting with the dimensions of the UPII-SV40Tag line followed by the WT line in a counter clockwise fashion. For the case of simplicity, we again use the grided layout. We discern that majority of the genes are upregulated, but the red records prevail in the image, which indicates that a large number of genes are downregulated. These records are pulled into the two directions – to the right side or the left side of the display, depending on the down or up regulation. The positioning of the majority of the records in the centre of the image lead us to believe that most of these records have very similar signal values on the 15 arrays, as the records that pulled away from the centre are more unbalanced and have some values that are out of proportion. We are interested in

the profile of these records and their behaviour over the time course for the two lines (WT and UPII-SV40Tag lines). We utilize the SOM-RadViz visualization as our next step to analyze their behaviour.

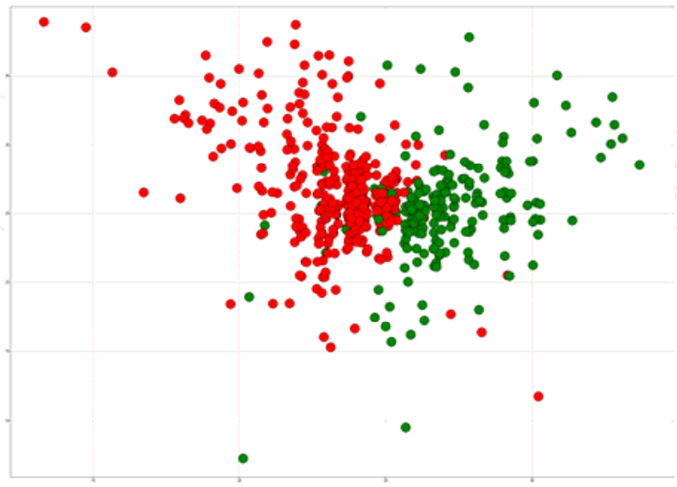


Fig. 15. RadViz visualization of the selected TCC set of 585 Probe Set IDs. Upregulated records are marked red and downregulated green.

We confirm in Figure 16 that a variety of differences exist in the two sets of records. For the SOM-RadViz visualization we use the RadViz approach as described above and overlay the surface of the unit circle using a 50x50 grid of cells. Each of these RadViz cells is provided with a 50-node stack into the third dimension. The neighbourhood function is a three-dimensional Gaussian function, adjusting all the output nodes in the proximity of the mapped location in a three-dimensional space. We used all of the fifteen dimensional values for the RadViz and SOM-RadViz (third dimension) approach. The genes concentrate at the top on the z axis, with much fewer records located at the bottom. We select a set of records from Figure 16 that are located at the three extremes: upper right (7 records) and left (4 records) corners and the bottom (22 records) to showcase the effect of the SOM-RadViz approach and the positioning of the records that is based by their dimensional values.
Figure 17 shows 33 selected records in three colours, depending on their location in Figure 16 (red, blue or purple) and as projected from the camera viewpoint in this figure (side projection). We have to draw a parallel coordinates graph (Figure 18) to show the details of their positioning. The genes in red have low dimensional values on all but the two time points of the UPII-SV40Tag line. The blue genes have higher values on the first two time points of the UPII-SV40Tag line and the purple genes have high dimensional values at all of the time points, although the first two time points of the UPII-SV40Tag line have the least deviation.

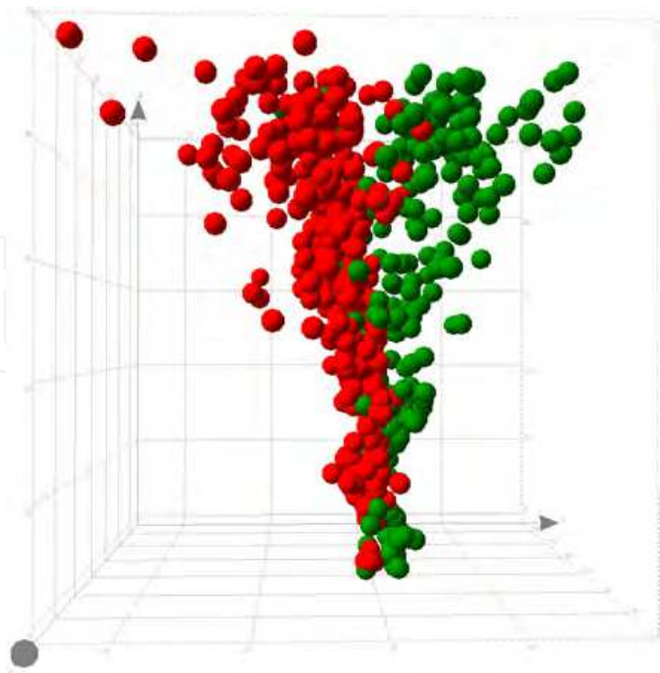


Fig. 16. SOM-RadViz visualization of the selected TCC set of 585 Probe Set IDs

A major goal of this project is the identification of biomarkers for early stage bladder TCC It is hoped that such markers can aid in predicting future developemnt of TCC, particularly in patients that have previously been treated successfully for low grade TCC, but have a high likelihood of recurrence. We have begun testing several of the genes upregulated in UPII-SV40Tag urothelium, including hyaluronan mediated motility receptor (RHAMM), autocrine motility factor receptor (AMFR), proliferating cell nuclear antigen (PCNA) and others as biomarkers for premalignancy, in patient urine samples obtained from a recently completed clinical trial. Our ultimate goal is to establish a panel of markers that can be readily detected in urine that will have high predictive value for TCC, thereby reducing the need for invasive and costly methods such as cystoscopy.

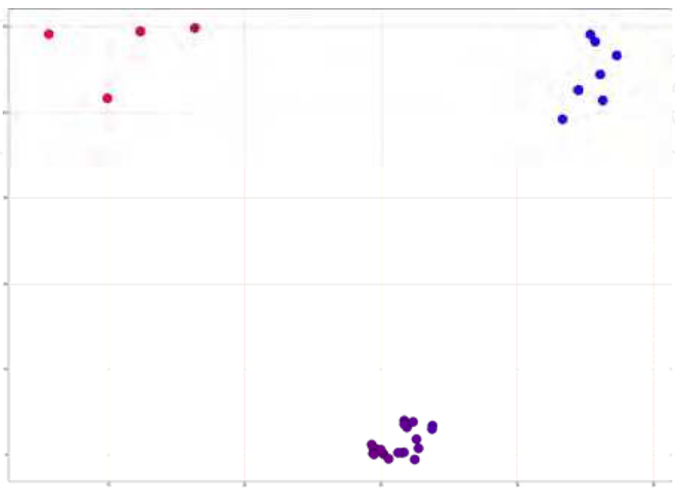


Fig. 17. SOM-RadViz visualization of a subset of 33 records of the TCC set

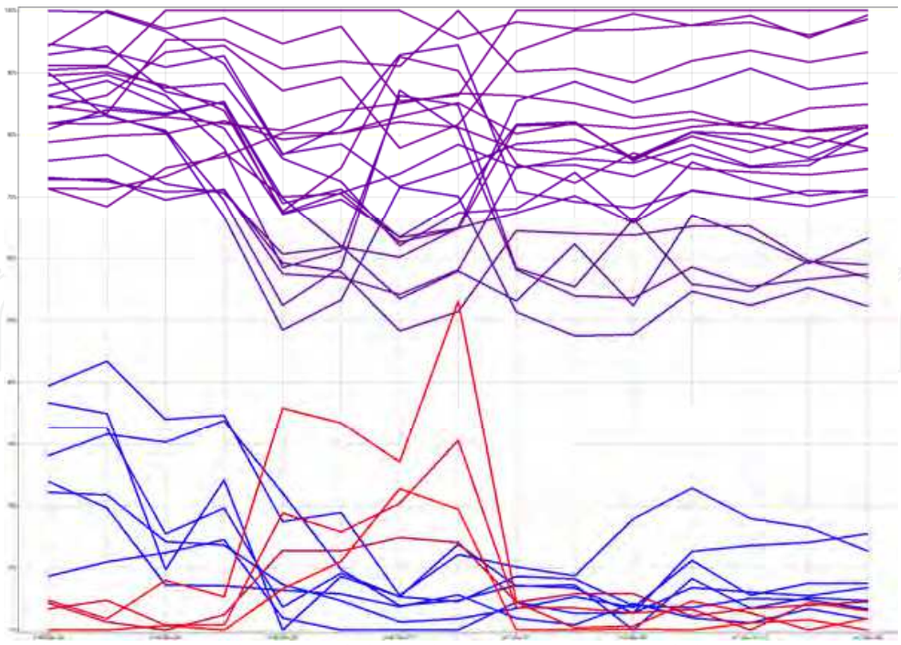


Fig. 18. Parallel coordinates display of the subset of 33 records of the TCC set

6. Conclusion

We provide a description of an approach for intelligent spatial placement of high-dimensional records, based on a modified Kohonen SOM algorithm. SOM-augmented visualizations provide for increased visual scalability and offer higher intrinsic dimension than the classic visualizations. The resulting mapping efficiently alleviates crowding and occlusion, and emphasizes the relationships among the neighbouring multi-dimensional records. Utilizing these techniques, the user can efficiently approach perceptual ambiguities associated with occlusion and gain insight into multi-dimensional data sets using an informative visualization, instead of a series of linked visualizations. These algorithms were tested on high-dimensional biomedical data sets and have provided for meaningful associations in an interactive environment.

7. Acknowledgment

The project described was supported by Grant Numbers P20RR016456 and P20RR018724 from the National Center for Research Resources, and from CA116324 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center For Research Resources or the National Institutes of Health.

8. References

Au, P.; Carey, M.; Sewraz, S.; Guo, Y. & Ruger, S. (2000). New Paradigms in Information Visualization (poster session), Presented at 23rd International ACM SIGIR Conference, Athens, Greece.

Bertini, E.; Aquila, L. D. & Santucci, G. (2005). SpringView: Cooperation of RadViz and Parallel Coordinates for View Optimization and Clutter Reduction, *Proceedings of Coordinated and Multiple Views in Exploratory Visualization, Third International Conference*, pp. 22-29.

Carr, D. (1991). *Looking at Large Data Sets Using Binned Data Plots: Computing and Graphics in Statistics*. Springer, New York.

Chambers, J.M.; Cleveland, W.S.; Kleiner, B. & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*, Wadsworth.

Chambers, J.M. ; Hastie, T.J. (1992). *Statistical Models is S*, Wadsworth & Brooks/Cole, Pacific Grove, California.

Chuah, M.; Roth, S.; Mattis, J. & Kolojejchick, J. (1995). SDM: Selective Dynamic Manipulation of Visualizations, *Proceedings of ACM Symposium on User Interface Software and Technology*, 61-70.

Cleveland, W.S.; McGill, R. (1984). Graphical Perception : Theory, Experimentation and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79, 387, 531-554.

Cleveland, W.S. (1993). Visualizing Data, Hobart Press, Summit, NJ.

Durbin, R. & Mitchison, G. (1990). A dimension reduction framework for understanding cortical maps. *Nature*, 346, 6259, 644-647.

Eick, S.G. (2000). Visual discovery and Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 6, 1, 44-58.

Fekete, J.-D. & Plaisant, C. (1999). Excentric Labeling: Dynamic Neighborhood Labeling for Data Visualization, *Proceedings of CHI'99*, pp. 512-519, ACM, New York.

Fisher, R.A. (1936). The use of Multiple Measurements on Taxonomic Problems, *Annals of Eugenics*, 7, 179-188.

Gautier, L.; Cope, L.; Bolstad, B.M. & Irizarry, R.A. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 12, 3, 307-315.

Gentleman, R. & Huber, W. (2003). Working with Affymetrix data: estrogen, a 2x2 factorial design example. Practical Microarray Course, Heidelberg.

Gentleman, R.; Carey, V.J.; Bates, D.M.; Bolstad, B.; Dettling, M.; Dudoit, S., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5, 10, R80.

Gentleman, R.; Carey, V.; Huber, W.; Irizarry, R. & Dudoit, S. (Eds.) (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer.

Grinstein, G.; Jessee, B.; Hoffman, P.; Gee, A. & Oneil, P. (2001). High Dimensional Visualization Support for Data Mining Gene Expression Data, In : *DNA Arrays: Technologies and Experimental Strategies*, CRC Press.

Hoffman, P.E.; Grinstein, G.; Marx, K.; Grosse, I. & Stanley, E. (1997). DNA visual and analytic data mining. *Proceedings of IEEE Visualization 1997*. pp. 437–441, Phoenix, AZ, IEEE Computer Society Press.

Inselberg, A. & Dimsdale, B. (1990). Parallel coordinates: a Tool for Visualizing Multidimensional Geometry, *Proceedings of IEEE Visualization 1990*, pp. 361-378, San Francisco, CA, IEEE Computer Society Press.

Leban, G.; Bratko, I.; Petrovic, U.; Curk, T. & Zupan, B. (2005) VizRank: finding informative data projections in functional genomics by machine learning. *Bioinformatics*, 21, 3, 413-414.

Manson, J. (1999). Occlusion in Two-Dimensional Displays : Visualization of Meta-Data. University of Maryland, College Park, MD.

R Development Core Team. R (2008). *A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna Austria.

Smyth, G.K. (2005). Limma: Linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Gentleman, R.; Carey, V.; Huber, W.; Irizarry, R. & Dudoit, S. (Eds.), 397-420, Springer.

Spence, R. (2007). *Information Visualization: Design for Interaction* (2nd Edition), Prentice Hall, Harlow, England.

Stone, R. II.; Sabichi, A.L.; Gill, J., Lee, I.; Loganatharaj, R.; Trutschl, M.; Cvek, U. & Clifford, J.L. Identification of genes involved in early stage bladder cancer progression. *Unpublished*.

Tamayo, P.; Slonim, D.; Mesirov, J.; Zhu, Q.; Kitareewan, S.; Dmitrovsky, E.; Lander, E. S. & Golub, T.R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *PNAS*, 96, 6, 2907-2912.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Ultsch, A. (1994). The Integration of Neural Networks with Symbolic Knowledge Processing. In : *New Approaches in Classification and Data Analysis*, Diday et al, (Ed.), 445-454, Springer Verlag.

Ultsch, A. & Vetter, C. (1994). Self-Organizing-Feature_maps versus Statistical Clustering : A Benchmark. *Technical Report*, 9, Department of Mathematics, University of Marburg, Marburg, Germany.

Van Osdol, W. W.; Myers, T. G.; Paull, K. D.; Kohn, K. W. & Weinstein, J. N. (1994). Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents. *J. Natl. Cancer Inst*, 86, 24, 1853-1859.

Wong, P. & Bergeron, R. (1997). Scientific Visualization – Overviews, Methodologies and Techniques : 30 Years of Multidimensional Multivariate Visualization, In : *Scientific Visualization, Overviews, Methodologies, and Techniques*, 3-33, IEEE Computer Society Press, 0-8186-7777-5, Los Alamitos, CA, USA.

Zhai, S.; Buxton, W.; Milgram, P. (1996). The Partial-Occlusion Effect: Utilizing Semitransparency in 3D Human-Computer Interaction, *ACM Transactions on Computer-Human Interaction*, 3, 3, 254-284.

Zhang, Z.T.; Pak, J.; Shapiro, E.; Sun, T.T. & Wu, X.R. (1999). Urothelium-specific expression of an oncogene in transgenic mice induced the formation of carcinoma in situ and invasive transitional cell carcinoma. *Cancer Res.*, 59, 14, 3512-3517.

IntechOpen

Self-Organizing Maps

Edited by George K Matsopoulos

ISBN 978-953-307-074-2
Hard cover, 430 pages
Publisher InTech
Published online 01, April, 2010
Published in print edition April, 2010

The Self-Organizing Map (SOM) is a neural network algorithm, which uses a competitive learning technique to train itself in an unsupervised manner. SOMs are different from other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space and they have been used to create an ordered representation of multi-dimensional data which simplifies complexity and reveals meaningful relationships. Prof. T. Kohonen in the early 1980s first established the relevant theory and explored possible applications of SOMs. Since then, a number of theoretical and practical applications of SOMs have been reported including clustering, prediction, data representation, classification, visualization, etc. This book was prompted by the desire to bring together some of the more recent theoretical and practical developments on SOMs and to provide the background for future developments in promising directions. The book comprises of 25 Chapters which can be categorized into three broad areas: methodology, visualization and practical applications.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen