THE DIPLOMA'S PRACTICE REPORT

By Ledak Eugene

I. Introduction

During the diploma's practice the methods of data visualization were investigated. These approach could help someone to learn more about statistical peculiarities of data set. This might be useful during the future elaborations of data mining approaches.

The diploma project must follow the criterion of practical usefulness. In our country the cancer question in general is very sharp, because the current technological level of our medicine doesn't allow to cure it. And the cancer diagnosis is sounds like a sentence to death

II. Used approaches

During the diploma's practice five approaches of data visualization were used. They are:

1) parallel coordinates;
2) scatter plot matrix;
3) survey plot;
4) circle segments;
5) radviz.

Later the short descriptions of these approaches would be given.

Parallel coordinates — parallel Coordinates are a simple, but powerful way to represent multidimensional data. Each dimension or attribute is represented by a vertical line. The maximum and minimum value of that dimension is usually scaled to the upper and lower points on these vertical lines. An N-dimensional point is represented by N - 1 line segments connected to each vertical line at the appropriate dimensional value.

Scatter plot matrix — grids of two-dimensional scatter plots are the standard way of extending the scatter plot to higher dimensions. For example, if one has 10 dimensional data, a 10 X 10 array of scatter plots is used to look at each dimension versus every other dimension. This is useful for looking at all possible two-way interactions or correlations between dimensions.

The decision to show only one scatter plot at a time was taken, because the resolution of display is too low, too show the scatter plot with proper quality.

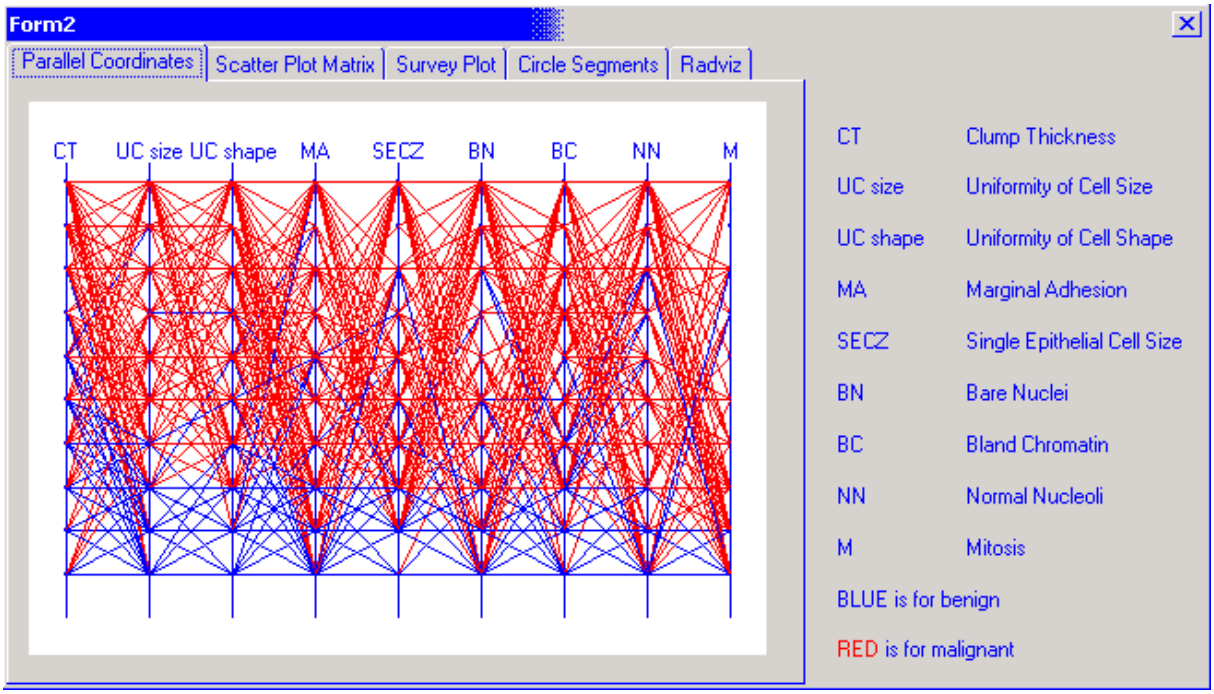
Survey plot — a simple variation of this extends a line around a center point, where the length of the line corresponds to the dimensional value. It is a visualization of N-dimensional data that allows one to quickly see correlations between any two variables especially when the data are sorted on a particular dimension. When color is used for different classifications, a sort can sometimes make it easy to see which dimensions are best at classifying the data.

Circle segments — it is similar to the Survey Plot. However, the data start from the center of a circle and radiate to the perimeter. A gray scale is used to show the value of a particular dimension, while the class value is colored in pie segments sandwiched around the dimensional values.

Radviz — (the short for "radial visualisation") n-dimensional data points are laid out as points equally spaced around the perimeter of a circle. The ends of each of n springs are attached to these n perimeter points. The other ends of the springs are attached to a data point. The spring constant Ki equals the values of the i-th coordinate of the fixed point. Each data point is then displayed where the sum of the spring forces equals 0. All the data point values are usually normalized to have values between 0 and 1.

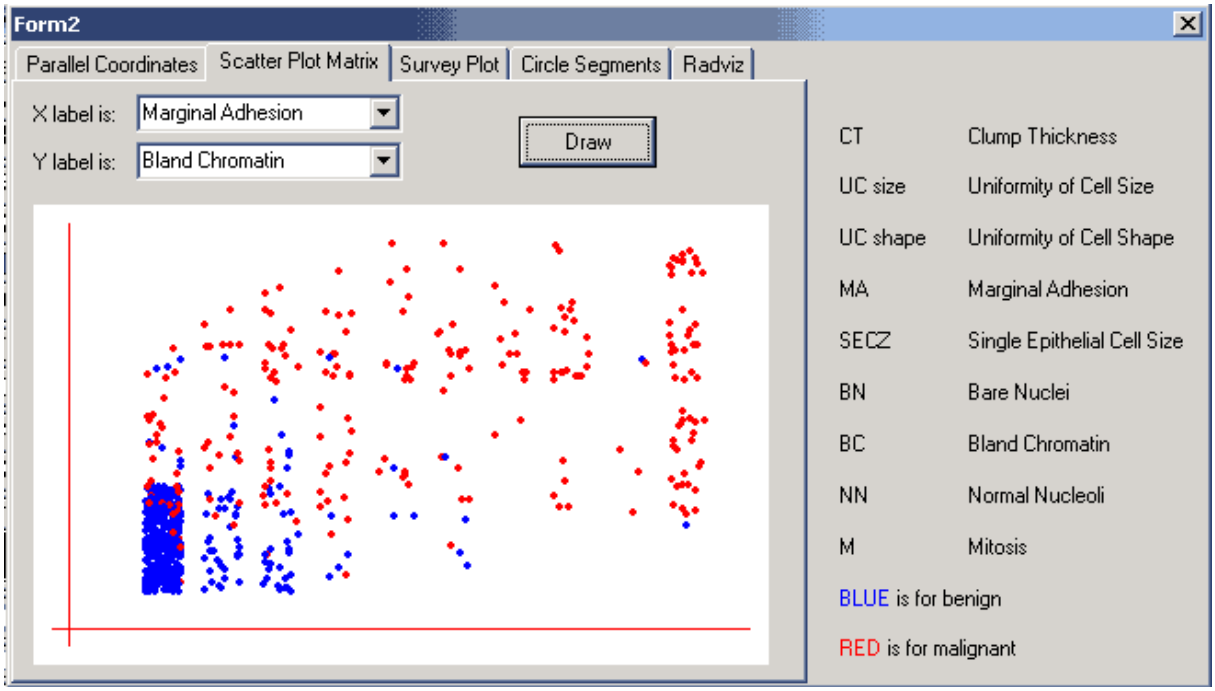
III. Received results

Later the pictures with the images of data set, received after visualization, would be given.



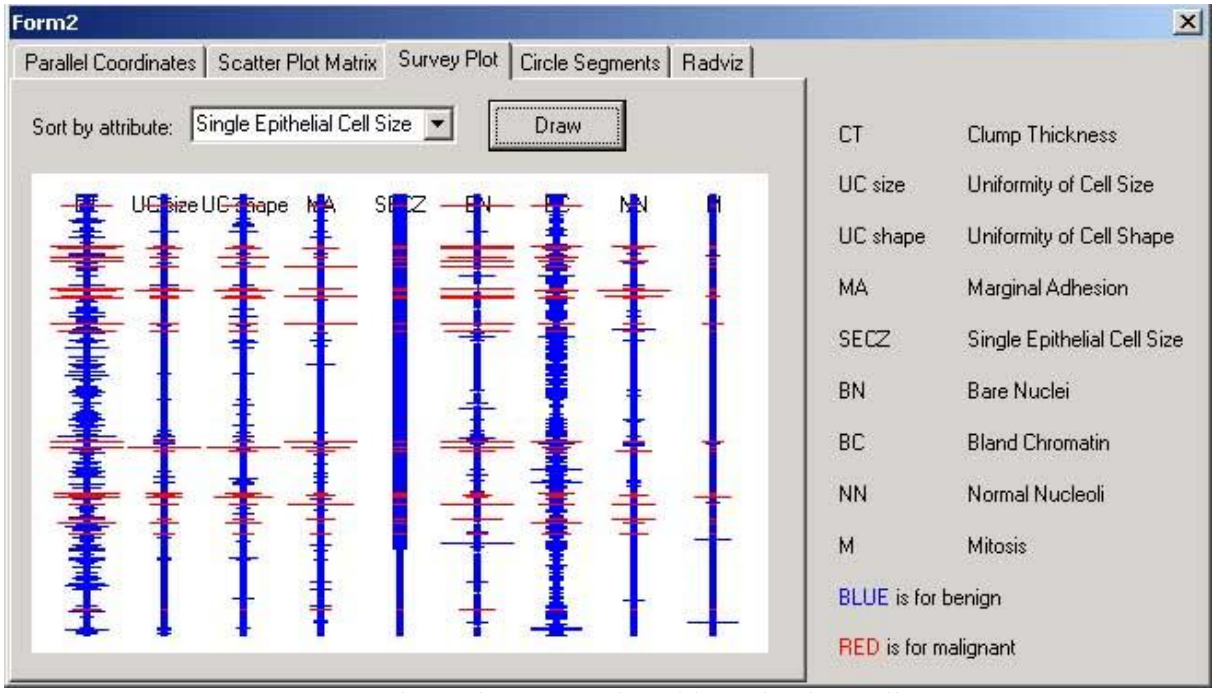
pic 1. The parallel coordinates

From this picture we can see, that the persons with benign cancer have low values of attributes in the most cases, when the persons with malignant cancer have high values of attributes.



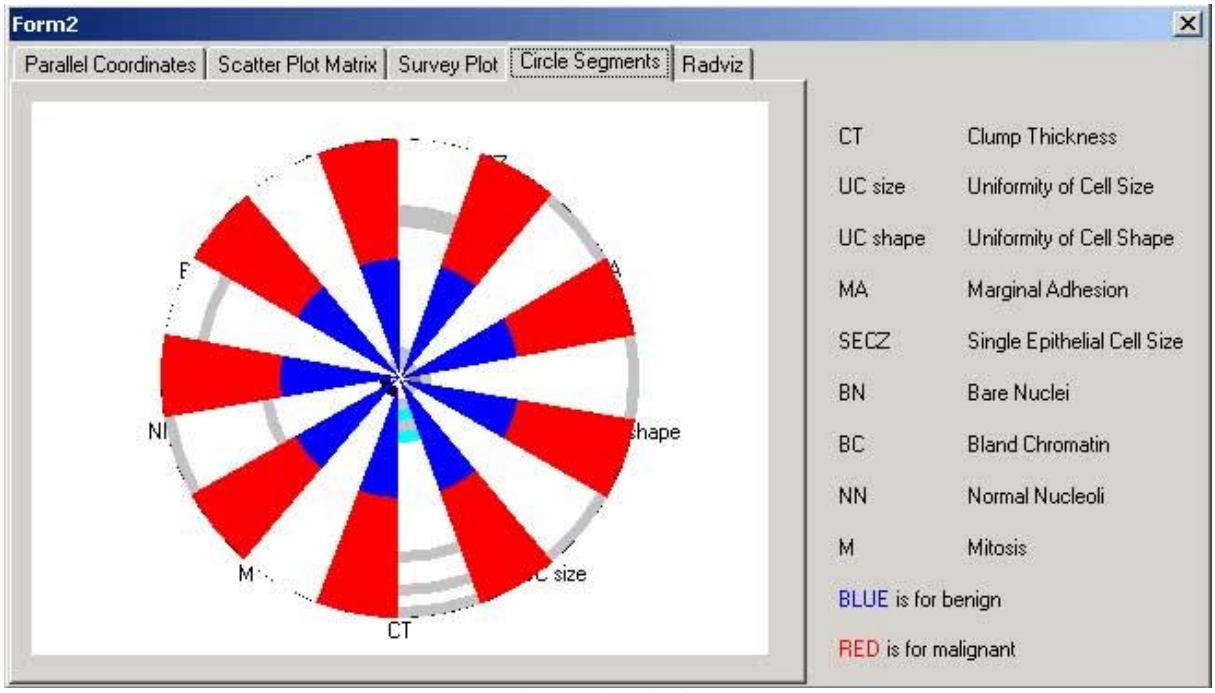
pic 2. The scatter plot

From this picture we can see, that the persons with benign cancer have low values of "Marginal Adhesion" and "Bland Cromatin" attributes in the most cases, while the persons with malignant cancer have values, disperced on the coordinate plane.



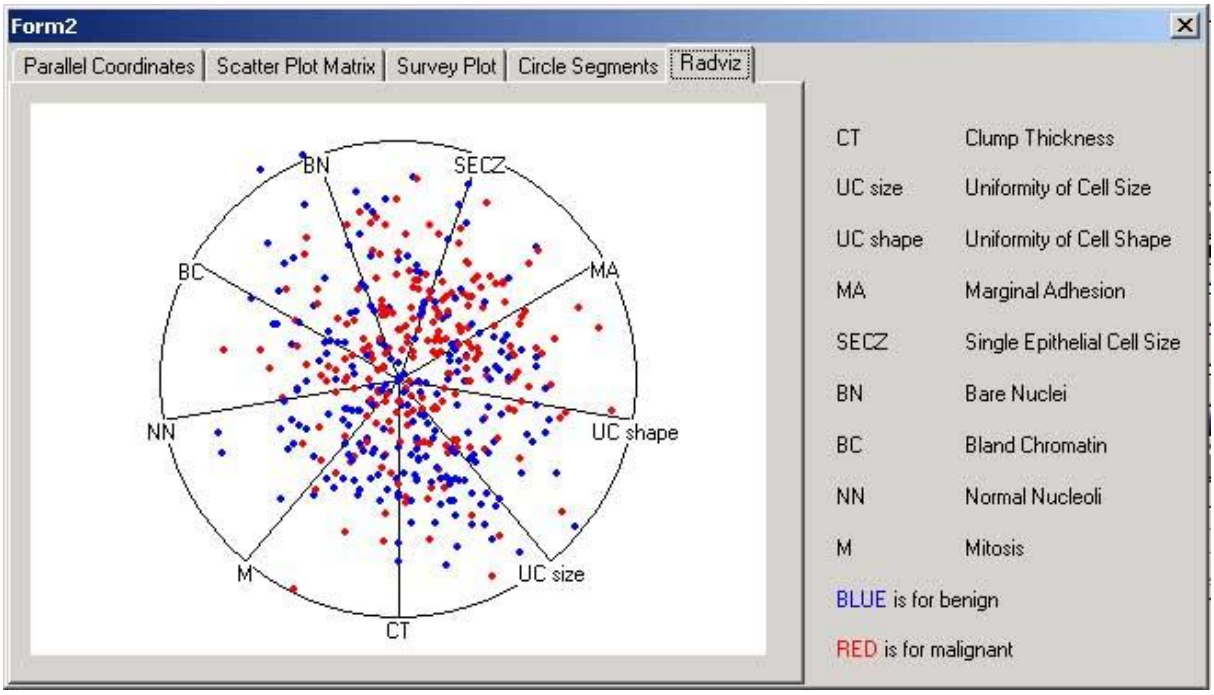
pic 3. The survey plot with sorting by attribute

The survey plots on this picture are sorting with the "Single Epithelial Cell Size". From this picture we can see, that the "Single Epithelial Cell Size" attribute isn't correlated with another attributes.



pic 4. The circle segments

From this picture we can see, that the persons with benign cancer have low values of attributes in the most cases, when the persons with malignant cancer have medium and high values of attributes.



pic 5. The radviz

From this picture we can see, that the values of attributes for both classes of persons (with benign and malignant cancer) have very versatile distributions.

IV. Conclusions

During the diploma's practice the methods of data visualization were investigated. These approach will help us to learn more about statistical peculiarities of data set. This might be useful during the future elaborations of data mining approaches (for testing the algorithms of C4.5 or ID3, for example).