Discovering Dependencies in Medical Data by Visualisation

Jacek Dryl[+], Halina Kwasnicka[*], Urszula Markowska-Kaczmar[*], Rafal Matkowski[+], Paweł Mikołajczyk[*], Jacek Tomasiak[*]

[+]Medical University of Wrocław

*Wroclaw University of Technology, Wyb.Wyspianskiego 27, 50-270 Wrocław, Poland

Kaczmar@ci.pwr.wroc.pl

Abstract In the paper the visualization of medical data sets by Samonn's mapping is tested. This mapping performs the projection of multidimentional data into the small size space. For the purpose of visualization the size is equal 2. The obtained results were verified with the results of statistical approach. The application of Samonn's mapping seems to be promising results for discovering the new regularities in the data and for producing the prognosis of the disease free survival for new patients, as well. In the future development of application the implementation of neural network is planned, which will perform the transformation of new data in an automatically way and it will help to find the values of missing data.

1. Introduction

Breast cancer and carcinoma of the cervix uteri are the most frequently diagnosed women's cancer in Poland. Despite of advances in diagnosis and treatment they are also the leading cause of cancer death.

Observational studies show that the disease diagnosed as breast cancer includes at least two entities that are, as yet, not reliably distinguished – one with a rapidly fatal outcome and the other with an outcome only slightly different from that of a group of women of similar ages without evidence of disease. If we could identify, which tumors were particularly aggressive, those patients might want to consider more intensive therapies.

This is a reason why many experimental studies are focused on development of new prognostic and predictive factors.
Owing to the interdisciplinary of our research group, which consists of medical experts and people from computer science discipline, we are able to understand more easily the problem and the profits offered by some heuristic and computational methods.
In the presented research we have tried to investigate whether visualization can be helpful in this problem. If yes, can it be applied alone or only as complementary tool for other methods. In the previous studies data collected in Lower Silesian Oncology Center about the patients with breast cancer were studied with statistical methods offered by package Statistica (StatSoft, Inc) and R package (Bell Laboratories). [9],[10]. Some of them will be presented here in order to compare the results obtained by statistical approach and visualization on the same data. Some experiments with visualization were also made with carcinoma of the cervix uteri patients because this group of patients was much larger.

This paper is organized as follows. In section 2 two experimental data set are presented, on the base which we evaluated Samonn's mapping as the method of visualization. It will be described in details in section 4.
In section 5 the experimental results are shown and compared with those obtained by statistical approach. Future plans and conclusion are presented at the end of paper.

2. The experimental data

In our study we have used two different data files. The first comes from 5-year observation of 527 patients with primary cancer of the cervix uteri treated in Lover Silesian Cancer Center in 1996, 1997 and 1998 [10]. The clinical and pathological data available on these patients include: date of birth and patients age, FIGO stage of the disease (according to FIGO Staging, 1994), tumor size, histological type of the tumor, degree of differentiation of the tumor, interval between diagnosis and first treatment (both dates), type of surgical treatment, type of performed radiotherapy, duration of radiotherapy, assessment of response to treatment, date of end of

hospitalization, last known vital status or date of death, relapse-free survival, overall survival.

The second data contain 5-year observation of 101 patients with primary ductal breast cancer (stage II) treated in Lower Silesian Cancer Center in 1993 and 1994. ER and nm23 expression was analyzed by immunohistochemical procedures in formalin-fixed, paraffin-embedded sections of primary tumors. The other clinical and pathological data available on these patients included: Bloom and Richardson's grade, tumor size, status of axillary lymph nodes, relapse-free survival, overall survival, body mass index, hormonal status and several other data from anamnesis and family history. Only the breast cancer data file was analyzed by statistical approach. [9]

3. Statistical analysis

The main goal of the statistical study with the data containing information about patients with breast cancer was to evaluate the prognostic value of two markers: estrogen receptor (ER) and nm23 protein [9]. The effect of multiple factors on survival was tested with Cox multivariate proportional hazards models modified by Akaike method [6]. Two separate models were investigated: one for overall survival and the second for disease free survival. The relation between two markers and other data were assessed using Kruskal-Wallis, Kendall and Spearman tests [7]. In the study the nm23 expression was not an independent prognostic indicator in breast carcinoma. Moreover, no correlation was observed between its activity and axillary lymph node status. In fact, there was a significant correlation between high level of nm23 protein expression and early distant recurrence of disease and shorter disease free survival.

In the study the expression of ER was and independent prognostic indicator in breast cancer. The patients with high expression of ER had longer overall survival and disease free survival.

The multivariate analysis revealed also the most important conventional prognostic factors in breast cancer. The lymph node status remains the major prognostic factor. The second one (apart ER expression) was the menopausal status. Women still menstruating had better prognosis comparing to those in menopause.

As for relapse free survival, the factors with significant influence were (aside from ER expression): lymph node status, tumor size and clinical stage, menopausal status, parity and other cancer in patient anamnesis.

4. Visualization

Another technique of data analysis can be analysis-through-visualization. It is proved by psychologists that information in graphical form is more comprehensible for human, so from ages people tried to find the techniques, which could illustrate the relationships hidden in the data in order to better understand them and possibly in order to find new dependences, to discover new categories in the data. In other words we can say it can be helpful in discovering new clusters in data. In presented case study we were interested in verifying usability of Samonn's mapping for discovering regularities in the above described data. The results were verified by medical experts and by statistical analysis. We have chosen Samonn's mapping because comparing it to PCA it better preserves distances between points even for the small ones in the original space.

4.1. Sammon's mapping

The obtained data had a form of table in Microsoft Excel (Microsoft Inc.)., where each row described data about one patient. These source data can be considered as a collection of points in multi dimensional space (one per row). Introducing Euclidean metric in the space, we get *geometrical metaphor* of the table of data. In this space close points correspond to the objects with similar properties [3]. As human brain can't operate points of dimension higher than 3, further analysis of data requires some kind of mapping (MultiDimensional Scaling - MDS) from multidimensional space to 2-3 dimensional one. In described case, Sammon's mapping [2] was used, but there are some other possible techniques (see [1], [4]).

General idea of Sammon's mapping is to minimize error function E, which measures the difference of distances between original high-dimension points and corresponding low-dimension points.
Usually error function E is given by the following formula:

$$E = \frac{1}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\delta_{ij}} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{\left(\delta_{ij} - d_{ij}\right)^2}{\delta_{ij}}$$

where:
d_{ij} – distance between points in n-dimensional space,
δ_{ij} – distance between points in m-dimensional space.
In the above formula usually $m \ll n$.
After a number of iterations it leads to the situation, where distances between calculated points approximate the original ones as good as possible [4]. Method is very flexible and it is possible to use it to map any collection *of* n-dimensional points onto m-dimensional one, where

$m \ll n$. As mentioned before, 2 or 3 dimensional space is the most convenient for visualization tasks.

Almost any minimization algorithm can be used here, for example Pseudo-Newton, Conjugate Gradients or Scaled Conjugate Gradients [4].

4.2. Data preparation

Analyzed source data set is initially represented as a table, where rows correspond to objects (cases) and columns, to properties (diagnosis, treatment, etc.). In most cases, it is necessary to scale each column by subtracting out the means and dividing by standard deviations (or other technique as suggested in [1]). This procedure makes columns equally important, despite their initial values.

4.3. Colors and interaction

To make visual analysis easier, in the computer application points obtained from Sammon's mapping

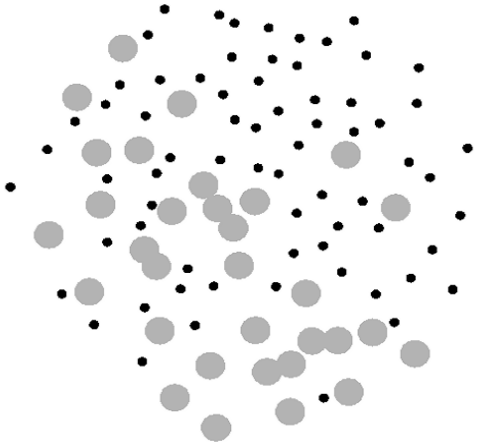

Fig. 1. The distribution of patients with breast cancer depending on the grade of tumor differentiation (B II or B III).Small points represent small size of tumor while the big points correspond to the big size of tumor

were plotted with different colors and sizes. Temperature color scale (blue-green-yellow-orange-red) was used. Parameters of individual points were calculated from normalized values of selected column in source data set. Researcher can easily change coloring (size) criterion and look for some similarities in points distribution. As there are only points visible in main (plot) view, it is always possible to select (highlight) one of them and read all information about the case in table view.

Comparing the results shown on Figures 1. and 2. we can see that with bigger number of points (larger data files) it is easier to see clusters in the picture. However taking into account points representing patients with low degree of differentiation – B III (Fig.1. – big red

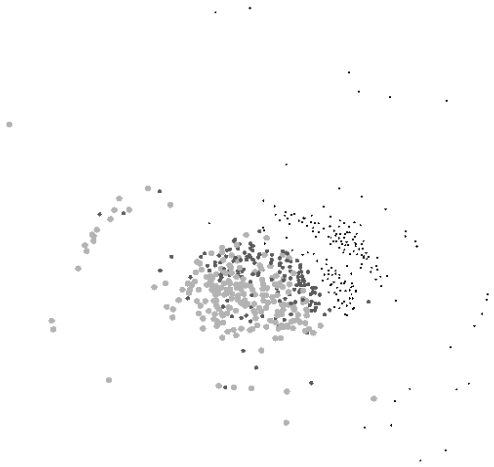

Fig. 2. The distribution of patients with primary cancer of the cervix uteri depending on the tumor size. Small points represent small size of tumor while the big points correspond to the big size of tumor

points), which are close to each other, we can conclude that they create a cluster.

5. Results

Based on statistical study it was concluded that the nm23 expression was not an independent prognostic factor in breast carcinoma. Moreover, no correlation was observed between its activity and axillary lymph node

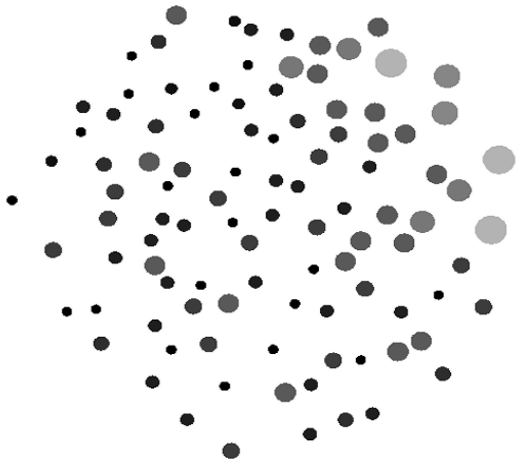

Fig. 3. The distribution of patients depending on nm23 expression in patients with breast cancer. The bigger points correspond to the bigger expression

status. In fact, there was a significant correlation between high level of nm23 protein expression and early distant recurrence of disease and shorter disease free survival.

The lack of correlation between expression of nm23 and overall survival can be easily noticed on Fig. 3, which shows the distribution of points representing breast cancer patients and colored depending on nm 23 expression.

In the statistical study the expression of ER was an

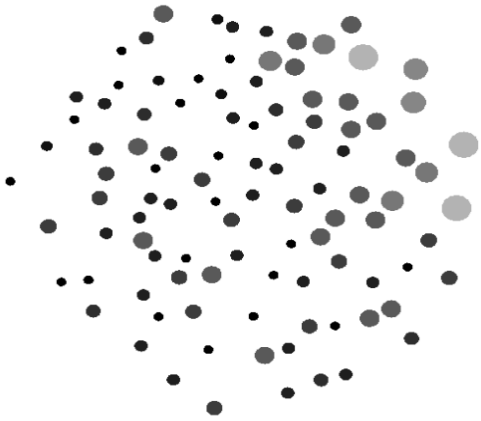

Fig. 4. The distribution of patients with breast cancer according to the ER expression. Small points represent short time while big ones correspond to longer time

independent prognostic indicator in breast cancer. The patients with high expression of ER had longer overall survival and disease free survival. This statistically proved hypothesis is not confirmed by the visualization analysis. When we compare Figure 4. with Figures 5.

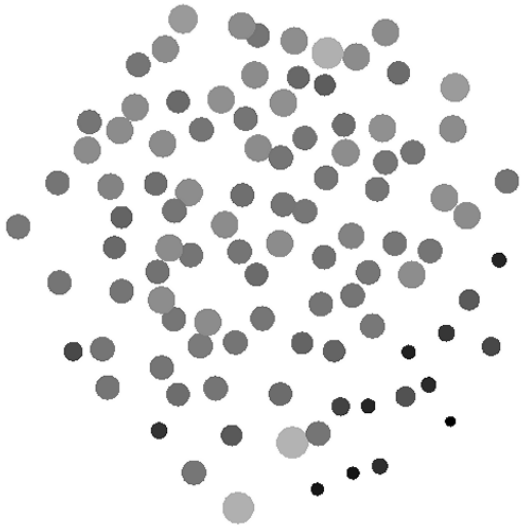

Fig. 5. The overall survival time for patients with breast cancer. Small points represent short time while big ones correspond to longer time

and 6. it is rather difficult to find this relationship. The possible explanation is the number of patients which were examined by statistical method.

How can it be analyzed using visualization technique?

At the beginning let us see Fig. 5., which presents the distribution of patients with breast cancer according to the overall survival. Bigger red points correspond to the longer survival time.

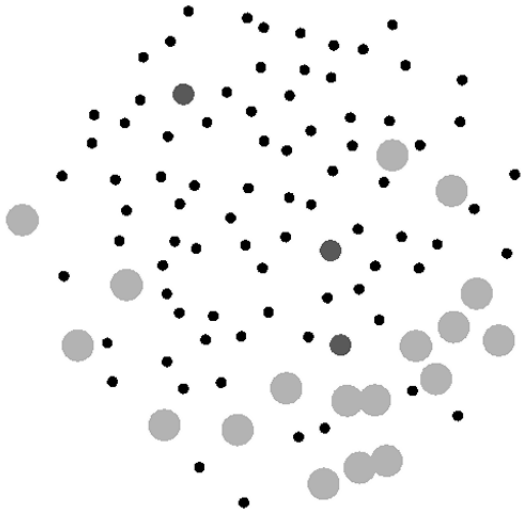

Fig.6. The distribution of points representing patients with breast cancer depending on presence and site of relapse. Small points indicate patients with lack of relapse while bigger ones represent patients with relapse

Figure 6 shows points representing patients with breast cancer according to the relapse. Small blue points indicate lack of relapse. As more points are in warm colour and as bigger they are they represent patients with worse relapse. Conclusion seems obvious: patients without relapse lived longer. It is truism, but it confirms correctness of the method itself.

The multivariate statistical analysis revealed also the most important conventional prognostic factors in breast cancer. The lymph node status remains a major prognostic factor. Figure 7. bears this relation out.

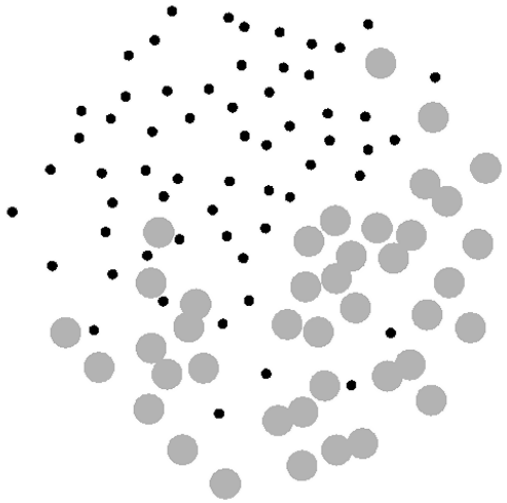

Fig.7. The distribution of patients depending on lymph node status. Small points represent better state then bigger ones

The next important prognostic factors according to the statistical analysis was the menopausal status. Women still menstruating had better prognosis comparing to

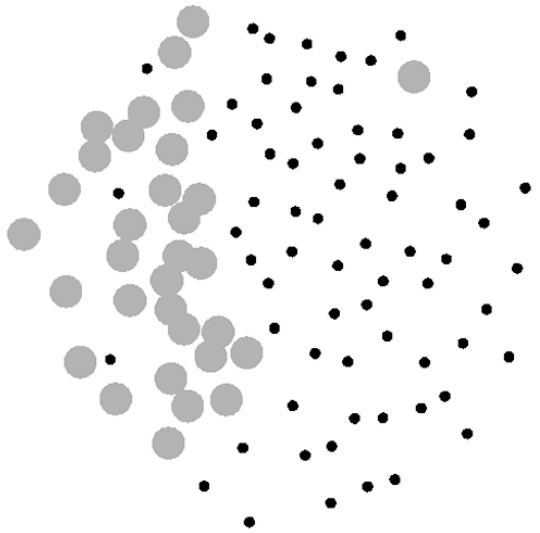

Fig. 8. The distribution of patients with breast cancer depending on menopausal status. The big points correspond to the menstruating woman

those in menopause. This statement can not be verified

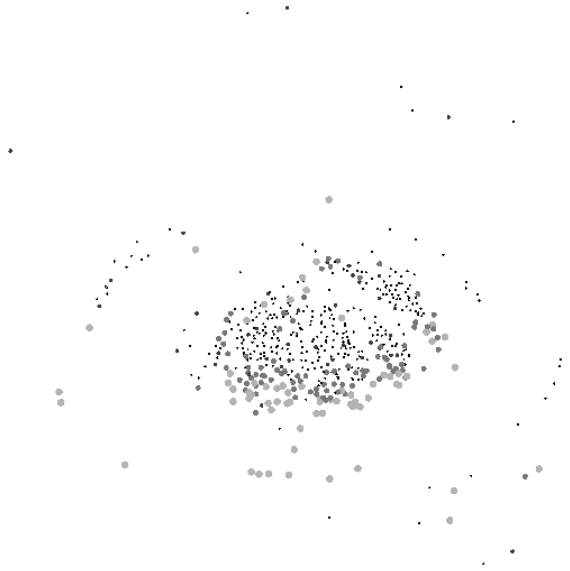

Fig.9. The distribution of patients with patients with primary cancer of the cervix uteri according to the degree of tumor differentiation. Bigger points represent patients with less differentiate tumors

by the visualization analysis to the whole extent. Let us compare Figure 5. showing the overall survival and Fig. 8. where menopausal status of patients is presented. The red points correspond to the menstruating women. There are some points in the left lower part of the plot, which can confirm this hypothesis, but this dependency is not so strong as for lymph node status.

As for relapse free survival the statistical analysis has shown that factors with significant influence were: ER

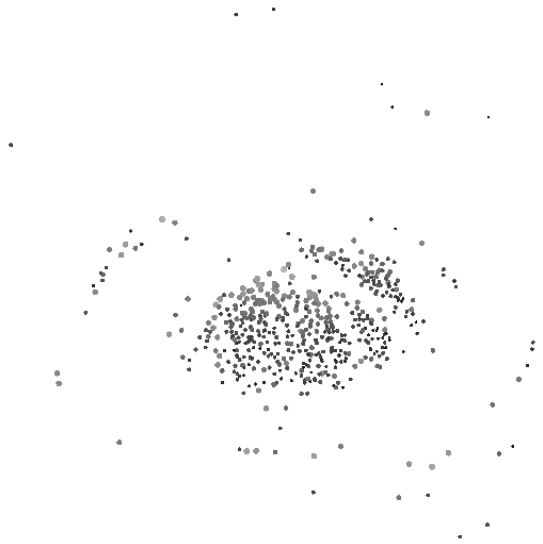

Fig. 10.The distribution of patients with primary cancer of the cervix uteri according to their age. As much lighter is the point as younger patient it corresponds

expression, lymph node status, tumor size and clinical stage, menopausal status, parity and other cancer in patient anamnesis. The visualization analyses has confirmed most of these relationships. In our opinion some discrepancies can occur because of the small amount of data in this case.

Unfortunately, this data set was composed of only about 100 cases, so it was difficult to find some interesting, new relations. But it was good example to make preliminary experiments in order to compare the results obtained by statistical approach and by visualization.

We also experimented with data concerning patients with primary cancer of the cervix uteri. One interesting relationship can be found when comparing Fig. 9 and 10. The first one presents the distribution of patients depending on degree of differentiation of the tumor, whereas the second shows the plot of patients according to their age. On the basis of these two plots we can see that younger patients have less differentiated tumors.

6. Future works

Presented experiments proof this visualization method to be a very useful tool for predicting a prognosis (disease free survival). In the future development of application

each point will be labeled by the data of patient. Next, when we introduce the data of new patient by localizing this new point and its neighbors in the plot, we can simply conclude the disease free survival time.
There are some issues that have to be solved in future studies. In present application, there is no simple way to add new points into the set generated with Sammon's mapping. We consider using a neural network to let the user add and/or modify points in previously transformed data set. Experiments with ANNs have been previously performed as described in [4], but we hope that generalization abilities of ANN could be used not only to add/modify points, but also to fill some missing values in existing data.

7. Conclusions

Presented results confirmed that data visualization with Samonn's mapping can be seen as good solution in medical data analyses. Their effects can be compared with statistical approach, but they can be easier interpreted by the end user. Described results are preliminary but in the nearest future with developing neural network to support Sammon's mapping we plan to verify it on the benchmark files.

References

[1] McLeod A.I., Provost S.B. (2001). *Multivariate Data Visualization*. http://www.stats.uwo.ca/faculty/aim/mviz

[2] Sammon, J.W. (1969). *A nonlinear mapping for data structure analysis*, IEEE Transactions on Computers, Series C 18, 401-409.

[3] Gorban A.N., Zinovyev A.Yu. (2001). *Visualization of Data by Method of Elastic Maps and Its Applications in Genomics, Economics and Sociology*. http://www.ihes.fr/PREPRINTS/M01/Resu/resu-M01-36.html

[4] Pękalska E., de Ridder D., Duin R.P.W., Kraaijveld M.A., (1999) *A new method of generalizing Sammon mapping with application to algorithm speed-u.* ASCI'99, Proc. 5th Annual Conference of the Advanced School for Computing and Imaging (Heijen, NL, June 15-17), ASCI, Delft, 221-228.

[5] de Ridder, D., Duin, R.P.W., (1997) *Sammon's mapping using neural networks: a comparison*, Pattern Recognition Letters, vol. 18, no. 11-13, pp. 1307-1316

[6] Sakamoto Y., Ishiguro M., and Kitagawa G., (1986) *Akaike Information Criterion Statistics*. D. Reidel Publishing Company

[7] Hollander M., Wolfe D.A. (1973) *Nonparametric statistical inference*. John Wiley & Sons, New York

[8] Kalbfleisch J., Prentice R. L. (1980) *The Statistical Analysis of Failure Time Dat*a. John Wiley & Sons, New York

[9] Matkowski R. (2002) *Prognostic value of estrogen receptor and nm23 protein expression in ductal breast cancer cells and their relations with clinical factors*. Doctoral thesis. Medical University of Wrocław, Wrocław

[10] Kornafel J., Dryl J., Matkowski R. *5-year observation of 527 patients with primary cancer of the cervix uteri treated in Lover Silesian Cancer Center in 1996, 1997 and 1998*. Unpublished data