ON VISUAL EXPLORATION OF BREAST CANCER DATA USING THE SELF-ORGANIZING MAP

Dr. Tomas Eklund
Institute for Advanced Management Systems Research
Åbo Akademi University
Turku. Finland
tomas.eklund@abo.fi

Dr. Mikael Collan
Institute for Advanced Management Systems Research
Åbo Akademi University
Turku. Finland
mikael.collan@abo.fi

Päivi Jalava, M.D.
Department of Pathology
University of Turku
Turku. Finland
paijal@utu.fi

Teijo Kuopio, M.D., Dr.Med.Sci.
Department of Pathology
Jyväskylä Central Hospital
Jyväskylä. Finland
teijo.kuopio@ksshp.fi

Prof. Dr. Yrjö Collan, M.D., Dr.Med.Sci., FRCPath
Department of Pathology
University of Turku
Turku. Finland
yrjo.collan@utu.fi

Abstract – *This paper explores the use of self-organizing maps (SOM) for exploratory data analysis of breast cancer data. We were able to visualize the data with the SOM in a way that makes it possible to, rather easily, identify possible connections between variables. The possible connections found can then be further tested for statistical significance, using well-known statistical methods. We report preliminary results on using SOM for identifying possible relationships in breast cancer data. The results are consistent with the existing scientific medical literature on breast cancer. We discuss the usability of the SOM for finding patterns and connections between variables in medical data.*

Key words – Self-Organizing map, breast cancer, exploratory data analysis

1 Introduction

Medicine is an area characterized by large amounts of multidimensional data and rigorous application of statistical methods and tests. Nearly all statistics-based medical research is based on the identification of patterns and relationships in large amounts of data, making it a suitable area for the application of data mining approaches. In addition, statistical medical research can benefit from methods that have a low sensitivity to erroneous data and outliers, which traditionally have been problematic for statistical methods. Indeed, data mining has been frequently applied to medical datasets since, as in engineering, the availability of large amounts of complex data provides an ideal testing ground for data mining approaches. Neural networks, in particular those based upon supervised learning, have been one of the most frequently applied data mining tools in medicine, and have proven to be good complements to traditional statistical approaches.

Although visualization approaches, such as the self-organizing map (SOM) have also been applied to medical data, the approaches have often been diagnostic or classification based, and the use of the SOM for exploratory data analysis has not been, to our knowledge, strongly emphasized.

The self-organizing map has been used in a wide range of different applications. Primarily, the SOM has been used in engineering applications [2,27], but it has also been applied in other fields, such as financial analysis [19,9,10], macro economic analysis [12,7,16], and text analysis [14,15,31].

The SOM has also been applied in medicine. For example, Oja et al. [25] mention that 268 papers on using the SOM in engineering in biology and medicine were written between 1998 and 2001. Of the papers in medicine, cancer research has received some attention, particularly for diagnosis and classification. Several papers have been written on breast cancer analysis [for example, 6,32,3,18,24]. Most of these studies use the SOM for clustering or classification, focusing on the technical capabilities of the SOM versus other available (usually statistical) tools. Very few studies explore the use of the SOM purely for exploratory analysis, in order to form new hypotheses based on visual analysis of the data.

In this paper, we use the SOM for visual exploration and hypothesis formulation on a dataset, by searching for possible dependencies. Preliminary correlation hunting using the SOM has been explored in a similar manner, for example, in Vesanto and Ahola [28]. This research also builds on the findings of Vesanto [27], who has studied the use of the SOM in exploratory data analysis.

In the breast cancer literature, certain dependencies in the type of dataset we have are known and accepted. Using the SOM for preliminary analysis of the data, we identified a number of relationships and could compare and verify the found relationships by using sources from the medical literature on breast cancer.

2 Methodology

2.1 The Self-Organizing Map

The SOM is a two-layer unsupervised neural network that maps multidimensional data onto a two dimensional topological grid [13]. The data are grouped according to similarities and patterns found in the dataset, using some form of distance measure, usually the Euclidean distance. The result is displayed as a series of nodes on the map, which can be divided into a number of clusters based upon the distances between the clusters. As the SOM is unsupervised, no target outcomes are provided, and the SOM is allowed to freely organize itself, based the patterns identified, making the SOM an ideal tool for exploratory data analysis. "Exploratory data analysis methods, like SOM, are like general-purpose instruments that illustrate the essential features of a data set, like its clustering structure and the relations between its data items" [12]. Thus, the SOM can be said to perform visual clustering of data.

The SOM differs from statistical clustering methods in a number of ways, although it is similar to k-means clustering. Firstly, when using the SOM the targeted number of clusters does not have to be defined. Secondly, the SOM is more tolerant towards data that do not follow a normal distribution. Thirdly, the SOM is quite efficient, and is faster than most top-down hierarchical clustering methods [29]. Finally, and most importantly, the SOM is a very visual method, as opposed to many statistical methods.

2.2 The data

The dataset consisted of 497 rows of data, each corresponding to one patient. The variables included are the age of the patient (age), the size of the tumor in centimeters (size), lymph node status (node, presence or absence of cancer foci in the regional lymph nodes), histological grade of the tumor (GR, from grades 1 to 3) and estrogen receptor status (ER, percent of ER positive nuclei). The dataset is from the Turku University Department of Pathology (Finland) and is also used in a paper by Nastac et al. [23].

The histological grade of the tumor refers to the degree of differentiation of the cancer cells compared to normal breast cells, where grade 1 is well differentiated and grade 3 is poorly differentiated. A lower grade implies cells that look more like normal cells and grow slowly, whereas a high grade implies the opposite. High grade cells have a tendency to spread. The ER status reflects how well the cells are receptive to estrogen. A high ER tumor is more likely to grow in a high estrogen environment. The ER status defines which type of treatment the tumor is receptive to.

In earlier research, the dataset was used in an attempt to predict the ER status based upon the other clinical data available, using an adaptive back-propagation neural network model [22]. The ER status of a tumor can be evaluated using laboratory tests, but these may not always be available due to lacking financial assets or insufficient laboratory conditions, for example, in developing countries.

2.3 Training

The map was created using Viscovery SOMine 4.0 (http://www.eudaptics.de). Viscovery is a user-friendly SOM implementation employing a graphical user interface. SOMine also uses the batchoriented training algorithm, as opposed to sequential training used, for example, in SOM_PAK. Other features include automatic cluster identification based on a hierarchical clustering algorithm, Wards method, modified for use with the SOM. Thus, SOMine is a very fast and easy to use SOM package [9].

Choosing the correct size of the map is always a delicate consideration that is based upon the intended purpose of the map. In general, a small map is preferable for visualization purposes, whereas a larger map is preferable for visualization [9, p.208]. In this case, we are more interested in visualization capabilities, however, to a certain degree, compression of the data is preferable when looking for correlations. Therefore, a 50 node map was created. After outliers and erroneous data were removed, the data were scaled according to the variance.

3 Results

The final cluster map was not used in this study. Instead, attention was focused on the feature planes. The feature planes of the map are displayed in Figure 1.

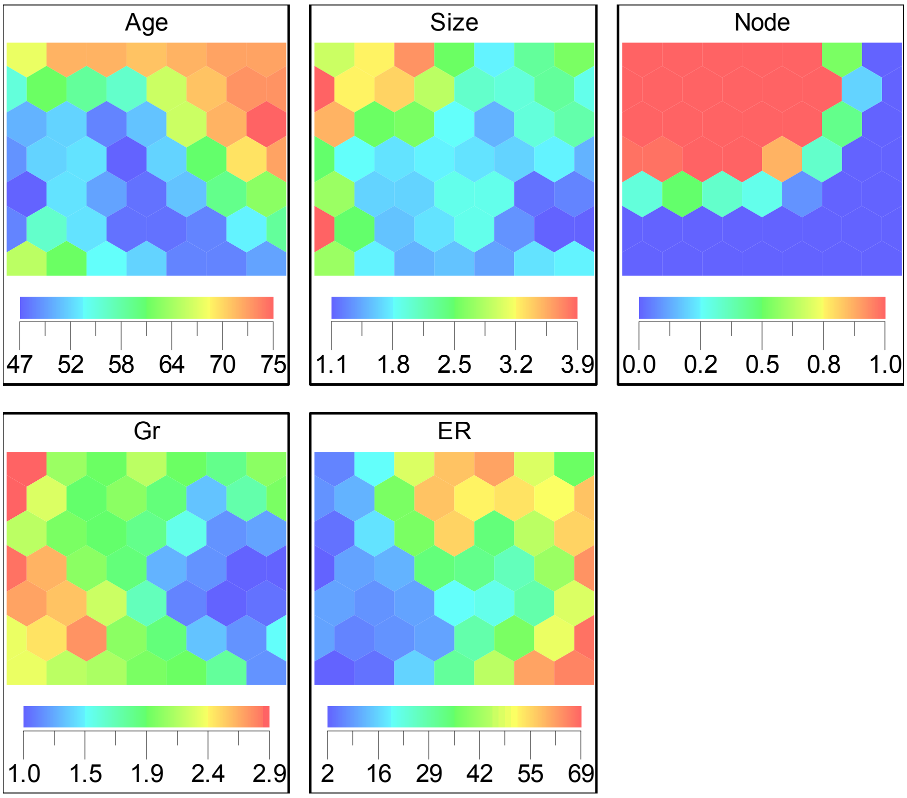

Figure 1. The feature planes of the resulting map.

The resulting map can be divided into a number of areas. Generally speaking, the right hand side of the map contains er positive patients, that is, patients who's ER status is above a specified ER threshold, termed ER-positive. This means that these patients are responsive to anti-estrogen treatment, whereas the patients on the left hand side are not. The patients on the left hand side of the map display a higher histological grade, i.e. their tumors have a higher degree of mutation than those on the right hand side. The patients on the upper left hand side of the map have tumors that have spread (sent metastases) to the lymph nodes (i.e. node is positive), and the size of the tumor is also larger.

The map showed numerous apparent associations. There were many associations with size. Size did not seem to be markedly associated with age, but showed association with lymph node status (large size, positive lymph nodes; small size, negative lymph nodes) and possibly grade (large size, high grade; small size, low grade). Negative lymph node status seemed to be associated with low grade. High grade tumors seemed to be associated with low ER expression.

Thus, the following three correlations in particular are interesting and will be further discussed.

1. The size of the tumor (size) is important for the likelihood of metastases being found in the lymph nodes (node).

2. With increasing size of the tumor (size), the histological grade (GR), i.e. degree of mutation compared to normal breast cells, will increase.

3. A high histological grade of the tumor (GR) reduces the tumor's receptiveness to anti-estrogen treatment (ER), i.e. where GR is high, ER will be low**.**

The size associations that were obvious on the map are biologically and clinically relevant. Size associations are explained by the probability for metastasis (extension outside the primary tumor) to lymph nodes, and distant metastasis [5,26,21]. Larger tumors will have more cells and a higher fraction of the cells will be shed out from the original tumor focus and allowed to be spread through interstitial tissues and along blood capillaries and lymphatics. In addition, larger tumors have had time to produce mutated cell populations which are especially capable of spreading. This is shown also in the size/grade association because the histological appearance reflects the capacity of the cells for spread and proliferation [8].

The grade/lymph node status association has the same type of biological background [11,17]. The association of ER with low grade is also well known [1,20]. High ER values reflect the differentiation of the neoplasm, and high grade tumors (which are poorly differentiated) are often ER-negative or low in ERs [4].

We can thus conclude that the SOM has correctly identified a number of existing relationships in a breast cancer data, and that strong support for these relationships can be found in the medical literature.

4 Conclusions

In this paper, the SOM has been used to form preliminary hypotheses concerning a database of breast cancer data. A SOM was constructed based on a medical dataset of breast cancer patients. Based on the resulting map, a number of connections were observed. The connections which were found using the SOM are well known associations suggesting that SOMs really can be used for scanning descriptive data, and finding associations for further study. Thus, in this case, the SOM has been applied for exploratory data analysis and preliminary hypothesis formulation on a medical dataset of breast cancer data.

The purpose of this paper has not been to gain new knowledge concerning breast cancer research. Instead, the focus has been on studying the use of the SOM for visual exploratory data analysis. The SOM provides a fast, intuitive, and visual way of performing exploratory data analysis and forming preliminary hypotheses, which can then be tested using statistical tools.

The potential application area for such an approach is wide. For example, in Genome research, which is characterized by extreme amounts of multidimensional data, visualization methods such as the SOM could be useful for quick analysis of the descriptive characteristics of the data.

References

[1] H.-O. Adami, S. Graffman, A. Lindgren, et al. (1985) Prognostic implications of estrogen receptor content in breast cancer. *Breast Cancer Research and Treatment,* **Vol. 5**, p. 293.

[2] E. Alhoniemi, J. Hollmén, O. Simula and J. Vesanto (1999) Process Monitoring and Modeling using the Self-Organizing Map. *Integrated Computer Aided Engineering,* **Vol. 6**(1), p. 3-14.

[3] O. Beckonert, J. Monnerjahn, U. Bonk and D. Leibfritz (2003) Visualizing metabolic changes in breast-cancer tissue using 1H-NMR spectroscopy and self-organizing maps. *NMR in Biomedicine,* **Vol. 16**(1), p. 1-11.

[4] G. Blanco, M. Alavaikko, A. Ojala, Y. Collan, M. Heikkinen, T. Hietanen, R. Aine and J. Taskinen (1984) Estrogen and progesterone receptors in breast cancer: relationships to tumour histopathology and survival of patients. *Anticancer Research,* **Vol. 4**(6), p. 383-389.

[5] C. L. Carter, C. Allen and D. E. Henderson (1989) Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer,* **Vol. 63**(1), p. 181-187.

[6] D. Chen, R. Chang and Y. Huang (2000) Breast Cancer Diagnosis Using Self-Organizing Map for Sonography. *Ultrasound in Medicine and Biology,* **Vol. 26**, p. 405-411.

[7] M. Collan and T. Eklund (2004). Transition of the Transition Economies - Using SOM to Map Socio-economic Development. In *Proceedings of the First International Manas University Conference in Economics*, Bishkek, Kyrgyzstan, 23-24 Sept., Kyrgyz-Turkish Manas University.

[8] Y. Collan, M. J. Eskelinen, S. A. Nordling, P. Lipponen, E. Pesonen, L. M. Kumpusalo, P. Pajarinen and K. K. O. (1994) Prognostic studies in breast cancer. Multivariate combination of nodal status, proliferation index, tumor size, and DNA ploidy. *Acta Oncologica,* **Vol. 33**(8), p. 873-878.

[9] G. J. Deboeck and T. Kohonen, Eds. (1998). *Visual Explorations in Finance with Self-Organizing Maps*. Springer finance. Berlin, Springer-Verlag.

[10] T. Eklund, B. Back, H. Vanharanta and A. Visa (2003) Using the self-organizing map as a visualization tool in financial benchmarking. *Information Visualization,* **Vol. 2**(3), p. 171-181.

[11] E. Fisher, C. Redmond and B. Fisher (1980) Histological grading of breast cancer. *Pathol Annal,* **Vol. 15**, p. 239.

[12] S. Kaski and T. Kohonen (1996). Exploratory Data Analysis by the Self-Organizing Map: Structures of Welfare and Poverty in the World. *Neural Networks in Financial Engineering*. P. N. Apostolos, Y. A.-M. Refenes, J. Moody and A. Weigend (Eds). Singapore, World Scientific, p. 498-507.

[13] T. Kohonen (2001) *Self-Organizing Maps*. Berlin, Springer-Verlag.

[14] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero and A. Saarela (2000) Self organization of a massive document collection. *IEEE Transactions on Neural Networks,* **Vol. 11**(3), p. 574-585.

[15] K. Lagus (2000). *Text Mining with WEBSOM*. Department of Computer Science and Engineering. Espoo, Finland, Helsinki University of Technology.

[16] A. Länsiluoto, T. Eklund, B. Back, H. Vanharanta and A. Visa (2004) Industry Specific Cycles and Companies' Financial Performance - Comparison with Self-Organizing Maps. *Benchmarking: An International Journal,* **Vol. 11**(4), p. 267-286.

[17] V. Le Doussal, M. Tubiana-Hulin, S. Friedman, K. Hacene, F. Spyratos and M. Brunet (1989) Prognostic value of histologic grade nuclear components of Scarff-Bloom-Richardson (SBR). An improved score modification based on a multivariate analysis of 1262 invasive ductal breast carcinomas. *Cancer,* **Vol. 64**(9), p. 1914-21.

[18] M. K. Markey, J. Y. Lo, G. D. Tourassi and C. E. J. Floyd (2003) Self-organizing map for cluster analysis of a breast cancer database. *Artificial Intelligence in Medicine,* **Vol. 27**(2), p. 113-127.

[19] B. Martín-del-Brío and C. Serrano-Cinca (1993) Self-organizing Neural Networks for the Analysis and Representation of Data: Some Financial Cases. *Neural Computing and Applications,* **Vol. 1**(2), p. 193-206.

[20] B. H. Mason, I. M. Holdaway, P. R. Mullins, L. H. Yee and R. G. Kay (1983) Progesterone and estrogen receptors as prognostic variables in breast cancer. *Cancer Research,* **Vol. 43**(6), p. 2985-2990.

[21] W. McGuire and G. Clark (1992) Prognostic factors and treatment decisions in axillary-node-negative breast cancer. *North England Journal of Medicine,* **Vol. 326**(26), p. 1774-1775.

[22] I. Nastac, Y. Collan, B. Back, M. Collan, P. Jalava and T. Kuopio (2004). *A Neural Network Model for Estrogen Receptor Status Prediction*. TUCS Technical Report No. 610. Turku Centre for Computer Science, Turku.

[23] I. Nastac, P. Jalava, M. Collan, Y. Collan, T. Kuopio and B. Back (2004). Breast cancer prediction using a neural network model. In *Proceedings of the WAC 2004 Conference*, Seville, Spain, 28.6.-1.7.2004.

[24] T. W. Nattkemper, B. Arnrich, O. Lichte, W. Timm, A. Degenhard, L. Pointon, C. Hayes and M. O. Leach (2004) Evaluation of radiological features for breast tumour classification in clinical screening with machine learning methods. *Artificial Intelligence in Medicine.* Article in press.

[25] E. Oja, S. Kaski and T. Kohonen (2003) Bibliography of Self-Organizing Map (SOM) Papers: 1998-2001 Addendum. *Neural Computing Surveys*, **Vol. 3**, p. 1-156.

[26] P. Rosen, S. Groshen, P. Saigo, D. W. Kinne and S. Hellman (1989) A long-term follow-up study of survival in stage I (T1N0M0) and stage II (T1N1M0) breast carcinoma. *Journal of Clinical Oncology,* **Vol. 7**, p. 355-366.

[27] O. Simula, P. Vasara, J. Vesanto and R.-R. Helminen (1999). The Self-Organizing Map in Industry Analysis. In *Industrial Application of Neural Networks*. L. C. Lain and V. R. Vemuri (Eds). London, CRC Press, p. 87-112.

[28] J. Vesanto (2002). *Data Exploration Process Based on the Self-Organizing Map*. Doctoral dissertation, Acta Polytechnica Scandinavia, Mathematics and Computing Series No. 115. Helsinki University of Technology, Espoo.

[29] J. Vesanto and J. Ahola (1999). Hunting for Correlations in Data Using the Self-Organizing Map. In *Proceedings of the International ICSC Congress on Computational Intelligence Methods and Applications* (CIMA'99), ICSC Academic Press.

[30] J. Vesanto and E. Alhoniemi (2000) Clustering of the Self-Organizing Map. *IEEE Transactions on Neural Networks,* **Vol. 11**(3), p. 586-600.

[31] A. Visa, J. Toivonen, B. Back and H. Vanharanta (2000). A New Methodology for Knowledge Retrieval from Text Documents. *TOOLMET2000 Symposium - Tool Environments and Development Methods for Intelligent Systems*.

[32] D. West and V. West (2000) Model selection for a medical diagnostic decision support system: a breast cancer detection case. *Artificial Intelligence in Medicine,* **Vol. 20**(3), p. 183-204.