Journal of Applied Mathematics & Bioinformatics, vol.1, no.1, 2011, 159-174
ISSN: 1792-6602 (print), 1792-6939 (online)

Microarray Data Mining with Fuzzy Self-Organising Maps

Chun Ho Yi[1] and Yahya Abu Hasan[2]

Abstract

The main problem of analyzing microarray datasets is that while it can measure genetic expression by the thousands, it has very little samples by comparison. For such data, a "big picture" method was employed here to cluster and visualize the data. Self-Organising Maps (SOM) organises a dataset based on distance measure and subsequently projects the clustered data onto a 2-dimensional plane for visual analysis. In this paper, we look at a modified SOM that hybridises with fuzzy c-means rules, and subsequently used to cluster and visualise leukemia and brain tumour microarray datasets. The resulting clustering and visualisation produced better visual projections, accuracy and significantly less mapping error than ordinary SOM. Interpretation of results in SOM is subjective due to its visual nature, and is dependent on the knowledge and expertise of the individual.

Mathematics Subject Classification : 62P10, 92B05

Keywords: Microarray datasets, Fuzzy-SOM, Clustering, Visualisation

---

[1] Universiti Sains Malaysia, e-mail: chunhoyi@gmail.com

[2] Universiti Sains Malaysia, e-mail: ahyahya@cs.usm.my

Article Info: *Revised* : March 16, 2011. *Published online* : May 31, 2011

1 Introduction

Microarray technology has hastened oncological research by leaps and bounds with the ability to measure the expression of thousands of genes in a single slide or chip. While microarray experiments are costly and time consuming, this has opened up many possibilities in oncological research with its high output capabilities. The production of a single massive dataset containing thousands of genes for a single sample, however, has brought up many statistical issues. One sample with thousands of attributes is hardly “ideal” statistically. Typically, the type of data used for statistical analysis consists of a few attributes with many samples. While a microarray research involves many samples, it still does not match the number of genes. It is a known fact that most gene expression datasets contain fewer than 100 samples, while genetic attributes number in the thousands. It would be a major task for a microarray experiment to be conducted on 100 samples, much less a few thousand samples.

Self-organising maps (SOM) has the characteristic of organising the data on its own with every iteration of its algorithm, based on the concept of neighbouring association. SOM has the advantage of clustering and projecting a large dataset onto a simple 2-dimensional projection, allowing for a visual analysis as opposed to a numerical one. This also separates SOM with conventional neural network methods, as SOM is an unsupervised clustering tool which produces distance-based visualisations of data, rather than produces a specific numerical measurement or output.

The analysis and interpretation of SOM projections is further aided by the use of component planes, which are an extension of SOM, where every attribute is visualised on its own projection [15]. Through this, one would be able to find patterns in the data and correlations between attributes. Conclusions and interpretations are drawn from individual knowledge and expertise based on the visualisations.

The ability of SOM to visualise massive datasets onto 2-dimensional projections

allows for the simplification of massively-sized datasets such as microarray datasets. Its application in analysing microarray data is not new, with some examples in [15], [11], [9], [10], [4], and [3]. These works have explored the use of SOM in microarray data mining involving breast cancer, prostrate cancer, yeast and macrophages. In those works, SOM has proven to be efficient in visualising the massive datasets containing thousands of genes but with only few samples. Furthermore, in [15] the component planes visualisation was fully utilised in looking for biological significance of gene clusters as found in the SOM training. Similarly, variations of SOM have also been explored and used in the analysis of microarray data. Examples of SOM variations include hybridising with partitive k-means [14] and multidimensional scaling [6].

Fuzzy rules are non-discrete rules that involve assigning membership of attributes to more than just one characteristic. Like its name, it allows for an attribute to be “fuzzy”, having some of each characteristic (at varying degrees), rather than being confined to a single one. In the case of microarray data analysis, traditional clustering methods would assign one gene to only one cluster. This can be very limiting, especially in the field of genetics. A single gene produces a single type of protein, but multiple proteins are involved in various bodily functions. Proteins are the primary components of enzymes, most hormones and many cellular components [2]. This concept of incorporating fuzzy rules into SOM has been explored as well, such as in [2], [13], and [5], but little is seen of its use on massive datasets.

A fuzzy SOM (FSOM) algorithm was proposed in [5], which was then used for learning activity patterns. We implemented the FSOM algorithm and tested it on a microarray dataset. The result was a very homogenous SOM that produced a meaningless visualization, making any interpretation impossible. In this work, a modified FSOM (MFSOM) is proposed and tested on a large scale microarray dataset. Fuzzy c-means clustering require some parameters prior to processing, such as the number of centres and “fuzziness factor” (or fuzzy variable). Other

fuzzy SOM algorithms would fix these parameters (such as in [2] and [1]). The FSOM algorithm did not require such parameters, but have omitted the neighborhood function from the SOM algorithm. The MFSOM algorithm proposed here requires few parameters as well and retains the characteristics of SOM while incorporating fuzzy rules, namely membership functions, into SOM.

There are 2 different microarray datasets that are used in this paper, and they are a brain tumor dataset and a breast cancer dataset. The publicly available brain tumor dataset was trimmed to a very small number of genes and processed with MFSOM as a test to see how well the proposed algorithm fared on a smaller microarray dataset. Subsequently, the proposed algorithm was tested on a much large scale dataset, which is the breast cancer dataset. This dataset was obtained from the supplementary of [7], and will be used to evaluate the MFSOM algorithm by comparing its results with the results as found in [7].

2 Self-Organising Maps (SOM)

2.1 Standard SOM

SOM is one of the most commonly used artificial neural network with respect to the unsupervised learning. It was shown to be an excellent tool in exploratory phase of data mining. It provides an easily-implemented algorithm for dimension reduction and visualization. The data vectors are projected onto positions on a two-dimensional grid. The grid consists of a set of regular ordered nodes, each associated with a prototype vector that is of the same dimension as the data points.

Since SOM is an unsupervised learning system, there is no supervisor to say what the output should be and the output vector must be coded by weight patterns and the input data. The objective of SOM is to competitively find the best matching winner and adapt the weights between input and output.

The most widely used technique to solve the cluster boundary identification

problem is unified distance matrix (U-matrix). The U-matrix makes a 2-d visualization of multi-dimensional data possible using the SOM code-vectors as data source.

2.2 Fuzzy SOM

Let X_{PN} denote the input space (matrix), M_{LN} the weight vectors (codebook), d_{PL} the Euclidean distance measure, R_{PL} the fuzzy membership function, and h_j the neighborhood function, where P, N and L are positive integers.

Step 1: Initialize the weight vectors (codebook).

Step 2: Compute *d*, the Euclidean distance measure:

$$d_{lj}(t) = \sqrt{\sum_{i=1}^{N}(X_{li} - M_{ji}(t))} \qquad \begin{matrix} l = 1,2,3,...,P \\ j = 1,2,3,...,L \end{matrix}$$

Step 3: Compute the membership of every map unit:

$$R_{lj} = \frac{(\frac{1}{d_{lj}^2(t)})^p}{\sum_{m=1}^{L}(\frac{1}{d_{lm}^2(t)})^p} \qquad \begin{matrix} l = 1,2,3,...,P \\ j = 1,2,3,...,L \end{matrix}$$

where *p* is a constant integer whose value is arbitrary and dependant on the size of the dataset.

Step 4: Update the weights vector:

$$M_{ji}(t+1) = M_{ji}(t) + h_j(t) \cdot \frac{\sum_{l=1}^{M} R_{lj}(t) \cdot (X_{li} - M_{ji}(t))}{\sum_{l=1}^{M} R_{lj}(t)} \qquad \begin{matrix} j = 1,2,3,...,L \\ i = 1,2,3,...,N \end{matrix}$$

where, in all cases, *t* is an indication of time or iteration step, instead of a variable. This algorithm runs iteratively for a set number of iterations. The stopping criteria used in this work is a fixed training length, where the number of training epochs is based on the dimensions of the dataset and number of map units used. The original SOM toolbox uses this stopping criterion, and it remains unchanged for the

implementation of the MFSOM algorithm.

At beginning of every iteration, one random input vector is chosen. Subsequently, the smallest Euclidean distance of the selected input vector to the codebook is chosen as the best matching unit (BMU). This is used for the calculation of the simple Gaussian neighborhood function used in the neighborhood function, h, where it is given as:

$$h_{ci}(t) = a(t) \cdot e^{\frac{d_{ci}^2}{2\sigma^2(t)}}$$

where d_{ci} is the distance between the BMU and the current map unit, $\sigma(t)$ is the current neighborhood radius and $\alpha(t)$ is the training rate (non-increasing function).

The training rate $\alpha(t)$ used is defined as:

$$a(t) = a_0 \left(\frac{a_T}{a_0} \right)^{\frac{1}{T}}$$

where α_0 is the initial learning rate, α_T the final learning rate, t is the iteration step and T is the learning length. The learning rate is determined based on the size of the dataset.

3 Data Sets

Prior to data processing, both datasets are first pre-processed. In the first step of data preparation, both datasets are edited to retain only the gene labels and sample names.

The brain tumour dataset, comprising of 7070 genes with 69 samples and 5 classes, was subjected to thresholding, i.e. forcing the entire data range to within [20, 16000]. A fold difference filtering was then applied and subsequently, the top 30 genes with the highest T-Value based upon gene classes for every gene is selected. The final dataset contained 145 genes and 69 samples.

The breast cancer dataset, comprising of 7650 genes with 99 samples, was

prepared using the standard deviation, a simple measure of variability, where genes with low variability were removed. From the patient information data, 4 more variables were added, namely tumour grade, estrogen receptor (ER), nodal status and relapse. The final dataset contained 2085 attributes with 99 samples.

Both datasets were processed using the normal SOM batch algorithm, FSOM, and MFSOM. For the MFSOM algorithm, the "fuzziness factor" variable, *p*, is arbitrary. Multiple values of *p* were attempted, and the value which produced the best result was used to compare with the other algorithms. For both datasets, the quantisation and topographic error were used as the measure of quality of the algorithm. Quantisation error refers to the mapping precision of the algorithm while topographic error refers to how well the topology of the input data has been preserved during training and how smooth the map is.

4 Experimental Results

In this section, the projections for the normal SOM algorithm and the MFSOM algorithm are compared. The component planes of the genes are not included here as they are irrelevant for this paper's work. For the breast cancer dataset, however, we include the component planes of the clinical data as it is of interest, due to the fact that the analysis can be compared to a previous work. While the quantisation and topographic errors of FSOM algorithm are included here for comparison, the U-matrices of the FSOM are not, as the visualisations produced are homogeneous and impossible to interpret.

In the initial phase, multiple *p* values for the MFSOM algorithm were tested to determine an optimal value for use in analysis. Table 1 lists the *p* values attempted for both datasets and the resulting quantisation and topographic errors. Figure 1 and Figure 2 are the plots of the information as found in Table 1. From the graph in Figure 1, it is observed that a *p* value of 25 produces the least quantisation error for the brain tumour dataset. In Figure 2, a *p* value of 45 produces the least quantisation error for the breast cancer dataset. Thus, we take the *p* values of 25

and 45 for the brain tumour and breast cancer dataset respectively as optimal, and use those values for the MFSOM algorithm when for performing comparisons with the other SOM algorithms.

In Table 2, based on the brain tumour dataset, the quantisation and topographic errors for the normal SOM, FSOM and MFSOM algorithms are listed. In Table 3 is based on the breast cancer dataset, lists the errors for the 3 different SOM algorithms as well. The MFSOM algorithm has shown better quantisation error in its training of the datasets. While the topographic errors are higher than the normal SOM and FSOM algorithms, nonetheless they are still within a reasonable range and are acceptable.

Table 1: Error table of varying *p* values for MFSOM.

p	brain tumour (145 genes)		breast cancer (2180 genes)	
	quantisation error	topographic error	quantisation error	topographic error
0.1	2.218	0.667	9.198	0.606
0.5	2.124	0.377	9.171	0.343
0.9	1.977	0.319	9.141	0.182
1	1.952	0.217	9.149	0.283
5	1.123	0.145	8.740	0.040
10	0.922	0.246	8.127	0.111
15	0.879	0.362	6.614	0.354
20	0.850	0.391	6.155	0.424
25	0.797	0.304	6.010	0.535
30	0.863	0.203	5.921	0.525
35	0.872	0.449	5.868	0.495
40	0.892	0.319	5.900	0.545
45	0.909	0.377	5.686	0.495
50	0.871	0.319	5.958	0.414
55	0.905	0.333	6.018	0.404
60	0.893	0.246	5.849	0.465
65	0.884	0.348	5.904	0.606
70	0.856	0.420	5.865	0.475

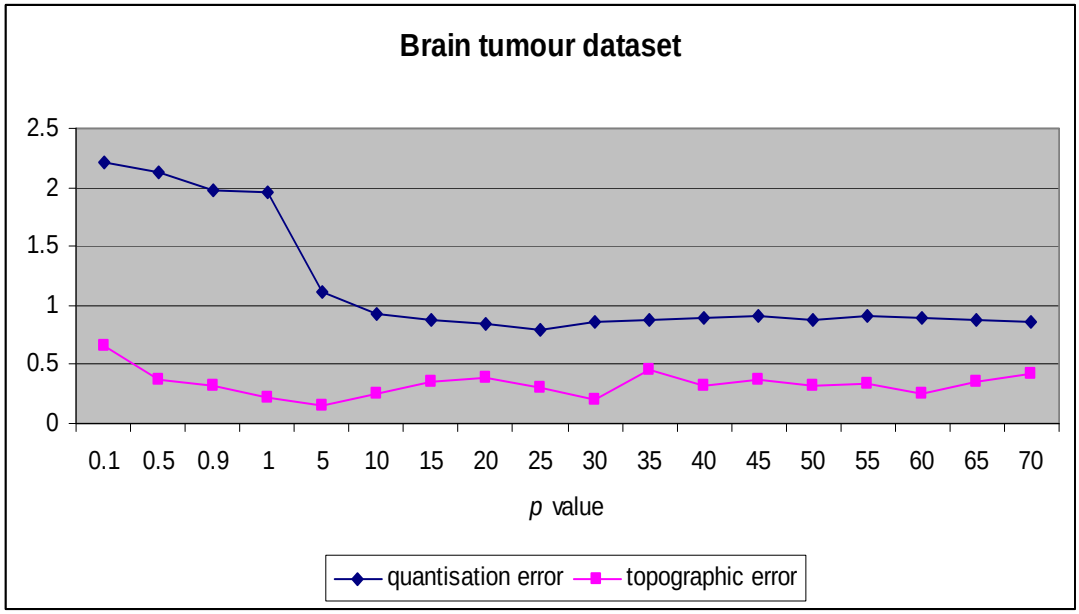

Figure 1: Comparative graph of varying *p* values for brain tumour dataset.

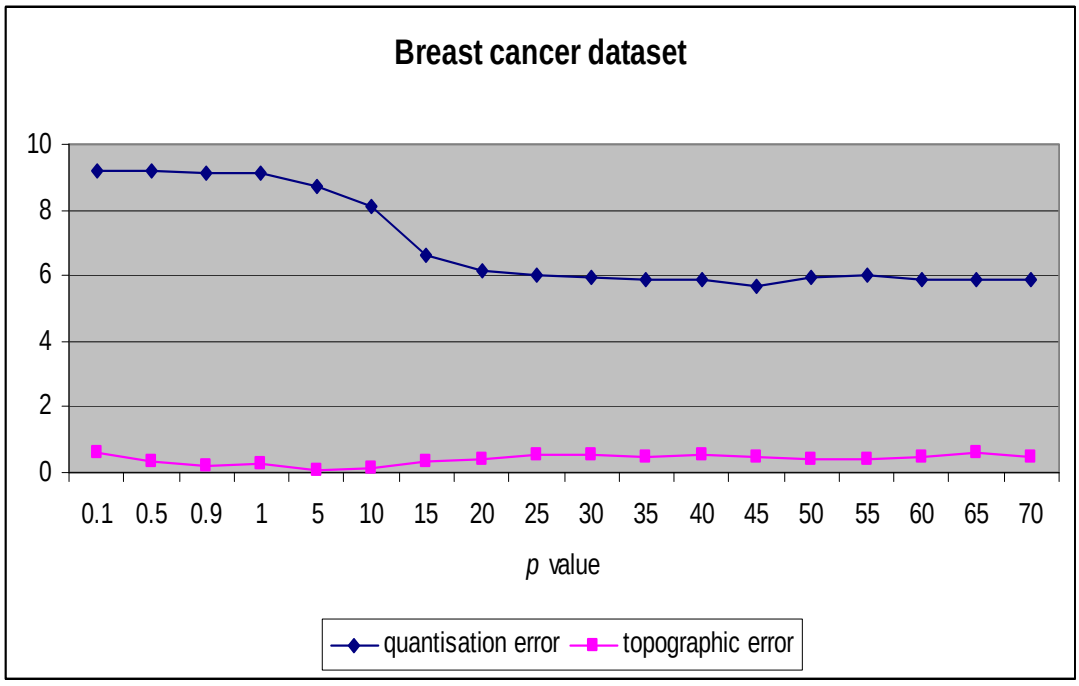

Figure 2: Comparative graph of varying *p* values for breast cancer dataset.

Table 2: Comparison of SOM algorithms for brain tumour dataset. MFSOM uses *p* value of 25

error	normal batch SOM	FSOM	MFSOM
quantisation	1.320	2.409	0.797
topographic	0.043	0.000	0.304

Table 3: Comparison of SOM algorithms for breast cancer dataset. MFSOM uses *p* value of 45.

error	normal batch SOM	FSOM	MFSOM
quantisation	8.029	9.381	5.686
topographic	0.010	0.061	0.495

5 Simulation Results

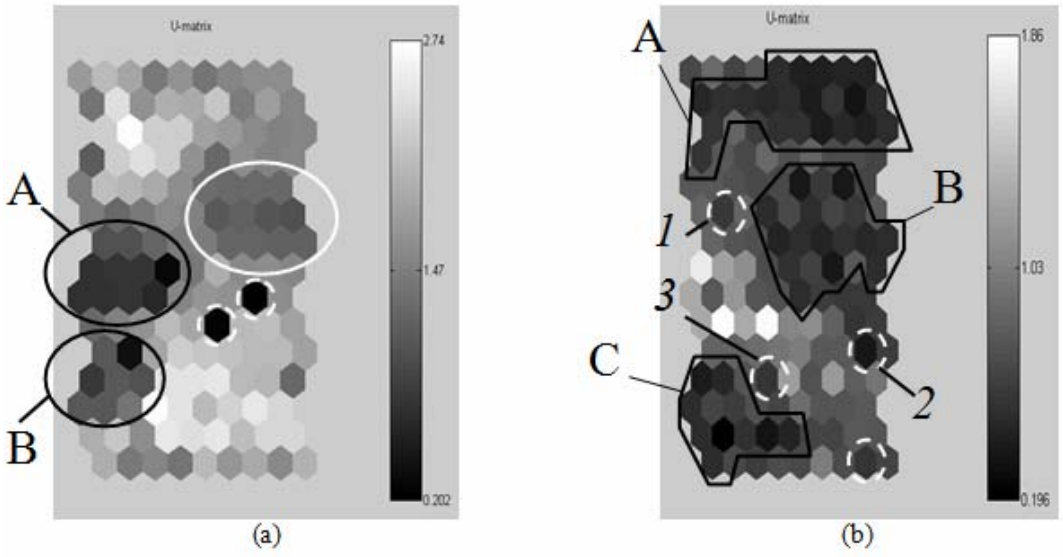

Figure 3: U-matrix for brain tumour dataset using MFSOM algorithm (a) and normal SOM (b). Colour bars indicating distance between data points.

The U-matrices in Figure 3 show the U-matrix projections for the brain tumour dataset. The U-matrix in Figure 3a was made using the MFSOM algorithm. The black circled areas, labelled A and B, show clusters while white circled areas

show possible clusters. The 2 white dashed circles indicate low distance between the surrounding neurons. These can be considered to be trivial clusters as it involves very small and negligible areas. However, it must be noted that this interpretation is still subjective.

The U-matrix in Figure3b was processed by normal batch SOM. There are 3 areas that can be identified as clusters, and these have been marked in the figure. Cluster A and B are separated unambiguously, but cluster B is not clearly defined. Its can be extended further downwards, but is entirely dependent on an individual. There are also 4 trivial clusters, represented by the dashed white circles. The 3 labelled trivial clusters, marked as 1, 2 and 3, are ambiguous as they could easily be incorporated into another cluster, with marker 2 to cluster B, marker 3 to cluster C and marker 1 to either cluster A or B. Here, we consider smaller distances between map units in order for a cluster to be taken into account, and thus removing any uncertainty about the membership of the above clusters.

The result of processing the brain tumour dataset using normal batch SOM and MFSOM has been illustrated in Figure 3. It can be seen that while there are some similarities, there are also obvious differences. However, differences are to be expected when running a different algorithm. Firstly, an entirely different U-matrix is to be expected, as the flow of the algorithm is significantly different between the normal batch SOM and the MFSOM. In the batch SOM U-matrix, there are 3 visible clusters, while the MFSOM had 2 visible clusters and 1 possible cluster. In the MFSOM, the status of the 3rd cluster as possible rather than definite could be drawn from the conclusion that while there is some correlation between those map units, it is not sufficient. Nonetheless, it remains subjective to the individual about how close a distance is necessary to count for a cluster. Therefore, while in the normal batch SOM, a 3rd cluster is conclusive, it is not in the MFSOM. Furthermore, the clusters in the MFSOM are more unambiguously separated and well defined, while it is the inverse for the normal batch SOM.

Secondly, based solely on the visual aspect, the component planes of the MFSOM

do not look as “nice” as the component planes for the batch SOM. It is inevitable that with a different algorithm, the component planes, much like the U-matrix, would not look the same as the normal batch SOM. A different distribution of map units in the U-matrix will result in the genes being represented differently in the component planes. The difference in the outlook of the U-matrix and component planes is expected as the MFSOM has a higher topographic error. Nonetheless, the error remains reasonable, and the mapping is satisfactory.

This brain tumour dataset was used as a preliminary test for the MFSOM due to the small number of genes (145 genes) used in the dataset. It was used to test the MFSOM algorithm on a small scale microarray dataset before being used for a large dataset. Here, we have seen the feasibility of the MFSOM on a small scale, and will proceed to implement it on a larger scale that is characteristic of microarray datasets.

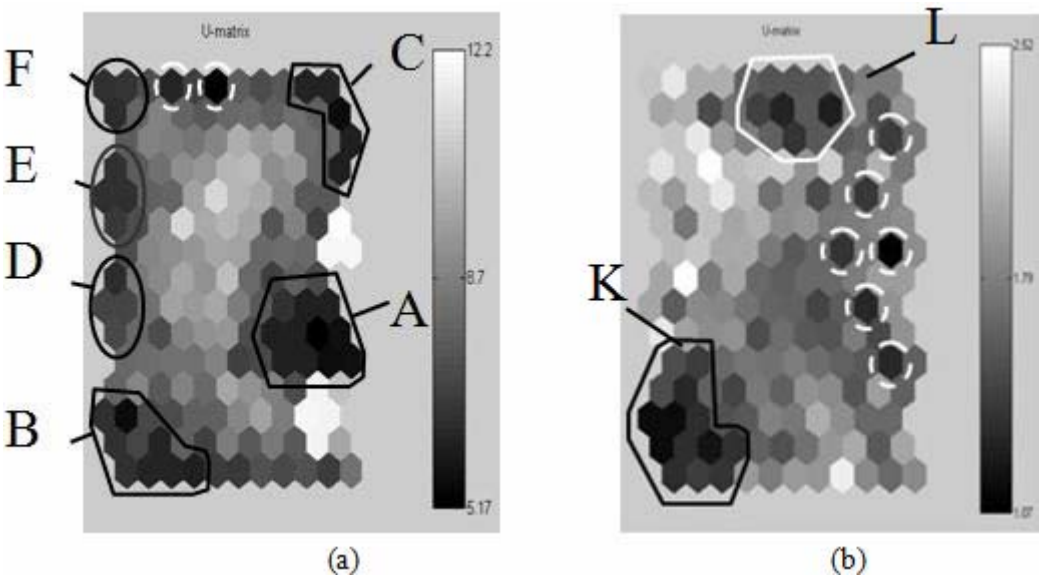

Figure 4: U-matrix for breast cancer dataset using MFSOM algorithm (a) and normal SOM (b). Colour bars indicating distance between data points.

Figure 4 shows the U-matrices projections for the breast cancer dataset. From the MFSOM visualisation in Figure 4a, there are 6 visible clusters marked A to F. Clusters C, D, E and F are small clusters, and their inclusion is subjective. There

are 2 trivial clusters, as indicated by the dashed white circles. The clusters here are unambiguously spaced, and are considered to be very well defined, even for clusters D, E and F that are small. The normal SOM visualisation in Figure 4b shows a single visible cluster K, a potential cluster L and several trivial clusters. The group of trivial clusters are not close enough to be merged into a single cluster, but this remains a subjective interpretation.

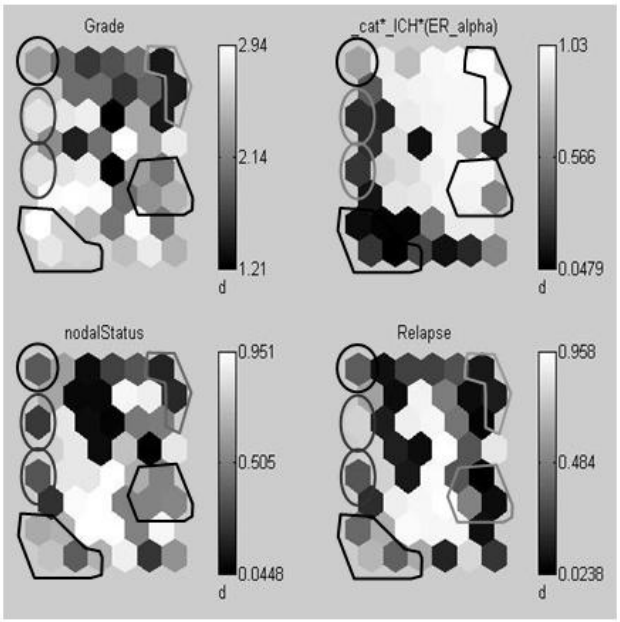

Figure 5: Component planes of patient info of breast cancer dataset using MFSOM (Colour bar indicating approximate values).

From the component planes in Figure 5, it can be seen that Cluster A has grade 2 tumours, positive estrogen receptor, negative nodal status and negative relapse. Cluster B has grade 3 tumours, negative estrogen receptor, uncertain nodal status and uncertain relapse. Cluster C has grade 1 nodal status, positive estrogen receptor, negative nodal status and negative relapse. Cluster D has grade 3 tumours and uncertain estrogen receptor, nodal status and relapse values. Cluster E has similar characteristics to cluster D but with the exception of having positive relapse. Cluster F has similar characteristics to cluster D but with the exception of having uncertain grade and estrogen receptor status. This conclusion is derived

from the fact that the estrogen receptor, nodal status and relapse attributes are Boolean attributes, having positive and negative values only, respectively represented by 1 and 0. As for grade, it has the discrete values of 1, 2 or 3, and thus the conclusions for the grade values have to be approximate to the nearest integer.

5 Conclusion

The difference between the MFSOM algorithm and the normal batch SOM is obvious from the projections of the U-matrix and component planes. While the normal SOM produced “nice” visuals, it only managed to find one cluster, one possible cluster and many trivial clusters. There is ambiguity among the many trivial clusters, as it is possible that they are grouped into one cluster, albeit with high distance between map units. As for the possible cluster, it is tempting to consider it as a cluster as there is only one certain cluster, but the distance between map units is rather insufficient. The MFSOM, on the other hand, has 6 well defined and unambiguous clusters, even though some of them are small. Also from the U-matrix, the 3 small clusters (D, E and F) appear to be somewhat correlated by their small inter-cluster distances. The MFSOM produced the same number of clusters are found by [7], where 3 clusters are found to be sub-clusters of a larger cluster. Cluster A, B and C shows that the estrogen receptor status is correlated to the tumour grade, where positive estrogen receptors indicates the presence of grade 1 and 2 tumours, while negative estrogen receptors indicate grade 3 tumours. This is consistent with the conclusion in [7].

ACKNOWLEDGEMENTS. This research is supported by a grant from the School of Mathematical Sciences, Universiti Sains Malaysia.

References

[1] J. Abonyi, S. Migaly and F. Szeifert, Fuzzy Self-Organizing Map based on Regularized Fuzzy c-means Clustering, *Advances in Soft Computing, Engineering Design and Manufacturing*, (2002), 99-108.

[2] J.C. Bezdek, E.C.K. Tsao and N.R. Pal, Fuzzy Kohonen clustering networks, *IEEE International Conference on Fuzzy Systems*, (1992), 1035 – 1043.

[3] G.E. Garrigues, D.R. Cho, H.E. Rubash, S.R. Goldring, J.H, Herndon and A.S. Shanbhag, Gene expression clustering using self-organizing maps: analysis of marcophage response to particulate biomaterials, *Biomaterials*, **26**(16), (2004), 2933-2945.

[4] S. Hautaniemi, O. Yli-Harja, J. Astola, P. Kauraniemi, A. Kallioniemi, M. Wolf, J. Ruiz, S. Mousses and O. Kallionimi, Analysis and Visualization of Gene Expression Microarray Data in Human Cancer Using Self-Organizing Maps, *Machine Learning*, **52**(1), (2003), 45-66.

[5] W. Hu, D. Xie, T. Tan and S. Maybank, Learning activity patterns using fuzzy self-organizing neural network, *IEEE Transactions on Systems, Man and Cybernetics, Part B*, **34**(3), (2004), 1618-1626.

[6] W.T. Keeton, *Biological Science*, 2nd Edition, W.W. Norton & Company Inc., 1972.

[7] C. Sotiriou, S.Y. Neo, L.M. McShane, E.L. Korn, P.M. Long, A. Jazaeri, P. Martiat, S.B. Fox, A.L. Harris and E.T. Liu, Breast cancer classification and prognosis based on gene expression profiles from a population-based study, *Proceedings of the National Academy of Science of the United States of America*, **100**(18), (2003), 10393 –10398.

[8] M. Sultan, D.A. Wigle, C.A. Cumbaa, M. Maziarz, J. Glasgow, M.S. Tsao and I. Jurisica, Binary tree-structured vector quantization approach to clustering and visualizing microarray data, *Bioinformatics*, **18**(Suppl. 1), (2002), 111-119.

[9] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky,

E.S. Lander and T.R. Golub, Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation, *Proceedings of the National Academy of Science of the United States of America*, **96**(6), (1999), 2907-2912.

[10] K. Torkkola, R.M. Gardner, T. Kaysser-Kranich and C. Ma, Self-Organizing Maps Mining Gene Expression Data, *Information Sciences*, **139**(1), (2001), 79-96.

[11] P. Toronen, M. Kolehmainen, G. Wong and E. Castren, Analysis of gene expression data using self-organizing maps, *FEBS Letters*, **451**(2), (1999), 142-146.

[12] J. Vesanto, J. Himberg, E. Alhoniemi and J. Parhankangas, SOM Toolbox for MatLab 5, *Report A57*, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, 2000.

[13] P. Vuorimaa, Use of the Fuzzy Self-Organizing Map in pattern recognition *Proceedings of the Third IEEE Conference on Fuzzy Systems, IEEE World Congress on Computational Intelligence*, **2**, (1994) 798-80.

[14] S. Wu and T.W.S. Chow, PRSOM: A new visualization method by hybridizing multidimensional scaling and self-organizing map, *IEEE Transactions on Neural Networks*, **16**(6), (2005), 1362-1380.

[15] L. Xiao, K. Wang, Y. Teng and J. Zhang, Component plane presentation integrated self-organizing map for microarray data analysis *FEBS Letter*, **538**(1) , (2003), 117-124.