Growing Hierarchical Self-Organizing Map (GHSOM) for Mining Gene Expression Data

Dipti D. Patil, Ph.D.
Associate Professor
Computer Engineering Dept.,
MIT COE, Pune

Prachi Gupta
Post Graduate Student
Computer Engineering Dept.,
MIT COE, Pune

ABSTRACT
This paper introduces a comprehensive review of a Growing Hierarchical Self-Organizing Map (GHSOM) reported in the specified writing. Investigating gene expression data is a very difficult problem due to the large amount of genes inspected. Computational methods have proved reliable to make sense of large amounts of data like the data obtained from microarray analysis. In this paper, we present inadequacies of standard algorithms K-Mean and self-organizing Map (SOM) and how GHSOM overcome these.

General Terms
GHSOM and SOM algorithms

Keywords
Self-organizing Map (SOM), Growing Hierarchical Self-Organizing Map (GHSOM)

1. INTRODUCTION
Cancer is a major cause of all the natural mortalities and morbidities throughout the world. Nearly thirteen percent of fatalities caused are due to cancer. Biomarkers typically refer to specific genes or their products that can be used to measure the progress of disease or the effects of certain treatment. The condition of the cell whether typical or cancerous can be determined according to genes that are expressed. Humans have approximately 20,000 to 25,000 genes, each one comprises of a sequence of bases [7]. Before genes can carry out their function, they are first transcribed into messenger RNA (mRNA), in a process called transcription. Molecule is in hence used as a template for the synthesis of a protein molecule during translation. Complete process including transcription of RNA and translation into protein, is called to as gene expression. Microarray analysis is helpful in detecting whether genes are active, hyperactive or inactive in different tissues. Since an immense number of genes are measured against a few samples, classification task becomes a big challenge, and other microarray data analyses. Main objective of this study is to identify the shortcomings of standard data mining algorithms for classification and why GHSOM is a better alternative for microarray analysis.

2. ALGORITHMS
Clustering involves partitioning a collection of objects into non overlapping groups, or clusters of objects where objects in a cluster are more similar to one another than to objects in other clusters.

2.1 K-means
K-means clustering (K-Means) [4] is a simple and fast method used commonly due to its straightforward implementation and small number of iterations. This algorithm splits the data set into k disjoint subsets. An estimation of the clusters (k) count is made by the user and calculated as an input where the computer randomly assigns each gene to one of the k-clusters. The distance between center of each cluster and each gene is promptly calculated resulting in an optimal grouping of data to clusters.

The most broadly utilized convergence criteria (1) for the K-Means algorithm is minimizing the SSE (Sum Squared Error):

$$SSE = \sum_{j=1}^{k} \sum_{x_i \epsilon c_j} \|x_i - \mu_j\|^2$$

Where

$$\mu_j = \frac{1}{n_j} \sum_{x_i \epsilon c_j} x_i$$

defines the mean of cluster c_j and n_j denotes the number of instances in c_j.

Multiple iteration for different values of k can be cumbersome especially when the magnitude of data is very large. Also, with the huge inherent noise in the gene expression data, most of it is "forced" into the clusters; that is questionable for the data integrity and correctness.

2.2 Self-organizing map (SOM)
A Self-Organizing Map (SOM) was proposed by Kohonen in 1995 [9], which is an unsupervised learning method. It projects high dimensional data onto lesser dimension mostly 2 or 3. The cluster count in the pattern is selected based around the expected number of resulted groups, which assumes some prior information on the input data. The distance between two objects in the space gives the level of similarities of these objects. The focus of SOM is to discover best matching between input data vectors and a two dimensional space of objects. The model comprises of number of units (neural processing components). An n dimensional weight vector mi is appointed to every unit. It is important to note that the weight vectors have the same dimensionality as the input patterns.

One of the limitations of the SOM lies with its static architecture that has to be defined initially. For microarray analysis applications, the user is not aware of the microarray data structure. Additionally it is not ready to represents data hierarchal way which helps navigate larger data quicker.

2.3 GHSOM
It is an unsupervised and adaptive architecture for clustering. According to data distribution, it grows both in a hierarchical way, permitting a hierarchical decomposition and navigation in sub-parts of the data, and in a horizontal way. GHSOM produces multiple layers with a hierarchical structure, where each layer includes independent SOMs.

Each layer is initialized with four units. Each map contains a number of units that represent clusters. That is, the concept of GHSOM is to grow in horizontally and vertically; until the resulting structure is appropriate to the corresponding input data. GHSOM starts at layer zero by a two by two map. At the start, each unit is assigned with a random weight vector. The size of the vector of weights is equal to the number of samples considered by the microarray.

Training at every layer starts with randomly selecting a data vector v from a unit's data vectors and calculating the Euclidean distances between v and the weight vectors w_i of all units in the concerned map. Based on these distances, the best matching (winner) unit (BMU) is determined and v_j (gene j) gets reallocated to this BMU. That is, the gene associated with v_j changes cluster.

This procedure is applied to all vectors (genes). Also, the weights of all the updated units (clusters) are recomputed based on the new genes' reallocation, a learning rate, and a neighbourhood function. This whole process is then repeated until a map (clusters) stabilizes, for all maps. After this, the mean quantization errors of units and maps are computed to determine whether to expand a map vertically or/and horizontally.

The mean quantization error of a unit/object i is calculated as follows:

$$mqe_i = \left(\frac{1}{N_U}\right)\sum \|w_i - v_j\| \quad for\ v_j \epsilon\ v_i\ N_U = |v_i|, v_i \neq \emptyset$$

where N_U represents the number of projected vectors v_j to unit i; weight vector is w_i of the unit i, and Vi is set of input vectors. The mean quantization error (MQE_{mapi}) of map_i is defined as follows:

$$MQE_{mapi} = \frac{1}{Ms_{mapi}}\sum mqe_i,\ \ for\ i \epsilon\ Ms \quad Ms_{map} = |Ms|$$

where Ms is the subset of the maps' units/objects; the breath expansion is continued until MQE_{map} reaches a certain fraction T1 of the mqe_u of object y, where y is the corresponding parent unit/object in the upper layer. T1 controls the horizontal expansion of maps and has values within [0..1].

After a possible expansion, the above described methodology is repeated at the new layer, until a global stability convergence in allocating genes to clusters. The outcome will be a hierarchy of maps that refines as we go towards lower layer. The weight vectors in the final maps corresponds to the allocated genes in the clusters.

3. COMPARISON

GHSOM has remarkable benefits over existing SOM algorithm especially for larger data. Here is the summary:

Table 1. Comparison among SOM and GHSOM [3]

Criteria	Result
Quantization Errors	Gives significantly less quantization errors on 14 out of the 16 experiments.
Speed	Runs faster than SOM by more than 30%
Improvements	Improvement ranges between 29% and 81%

4. CONCLUSION

In this paper we have given review on the different algorithms for clustering gene expression microarray data. GHSOM is adopted to overcome the limitation of SOM for huge microarray data. It provides improvements over SOM both in terms of speed and quantization errors.

5. ACKNOWLEDGMENTS

Our gratitude to the specialists who have helped towards development of the work.

6. REFERENCES

[1] Shaurya Jauhari, S.A.M. Rizvi, "Mining Gene Expression Data Focusing Cancer Therapeutics: A Digest" appeared in 2014 in IEEE/ACM transactions on computational biology and bioinformatics.

[2] Jonnel L. Dela Rosa, Alvin Edwin A. Magpantay, Alex C. Gonzaga, Geoffrey A. Solano "Cluster Center Genes as Candidate Biomarkers for the Classification of Leukemia " appeared in 2014 in (IEEE) Information, Intelligence, Systems and Applications, IISA, The 5th International Conference.

[3] Mansour, Nashat ; Department of Computer Science and Mathematics, Lebanese American University, Beirut, Lebanon ; Zantout, Rouba ; El-Sibai, Mirvat" Mining breast cancer genetic data" on 17/6/2014.

[4] Radha, R.; Dept. of Comput. Sci. , S. D. N. B. Vaishnave Coll. of Women, Chennai, India ; Rajendiran, P. "Using K-Means Clustering Technique to Study of Breast Cancer" on 17/6/2014.

[5] Hicks C, Asfour R, Pannuti A, Miele L. "An integrative genomics approach to biomarker discovery in breast cancer", Cancer Inform 2011;10:185-204.

[6] http://www.ifs.tuwien.ac.at/~andi/ghsom/

[7] Stein, L. D., "Human genome: End of the beginning", Nature, 431, 915-916, 2004.

[8] S.Tavazoie, D.Hughes, M.J.Campbell, R.J.Cho, G.M.Church, "Systematic determination of genetic Network architecture", Nature Genet, pp.281-285, 1995.

[9] T. Kohonen, Self-Organizing Maps. Springer, Berlin, 1995.

[10] A. Gruzdz, A. Ihnatowicz, and D. Slezak, "Interactive gene clustering A case study of breast Cancer microarray data," Information Systems Frontiers archive, vol. 8, no. 1, pp. 21-27, 2006.

[11] D. Covell, A. Wallqvist, A. Rabow, and N. Thanki, "Molecular
Classification of Cancer: Unsupervised Self-Organizing Map Analysis of Gene Expression Microarray," Data National Cancer InstituteFrederick, Molecular Cancer Therapeutics, vol. 2, pp. 317–332, March 2003.

[12] M. Markeya, J. Loa, G. Tourassib, and C. Floyd, "Self-organizing map for cluster analysis of a breast cancer database," Artificial Intelligence in Medicine, Elsevier, vol 27, pp. 113–127, 2002.

[13] S. Yobon, S. Wichaidit, and W. Wettayaprasit, "Microarray Gene Selection Using Self-Organizing Map" in Proc. Of the 7th WSEAS International Conference on Simulation, Modelling and Optimization, 2007, pp. 239-244.

[14] M. Markeya, J. Loa, G. Tourassib, and C. Floyd, "Self-organizing map for cluster analysis of a breast cancer database," Artificial Intelligence in Medicine, Elsevier, vol 27, pp. 113–127, 2002.

[15] M. Dittenbach, D. Merkl, and A. Rauber, "The Growing Hierarchical Self Organizing Map," in Proc. of the International Joint Conference on Neural Networks, (IJCNN'2000), Como, Italy, IEEE Computer Society Press, Los Alamitos, CA, July 24-27, 2000, pp. VI-15 - VI-19.