Clustering using Adaptive Self-Organizing Maps (ASOM) and Applications

Yong Wang[1], Chengyong Yang[1], Kalai Mathee[2] and Giri Narasimhan[1]

[1] Bioinformatics Research Group (BioRG)
School of Computer Science, Florida International University,
Miami FL 33199, USA
{cyang01, giri}@cs.fiu.edu

[2] Department of Biological Sciences, Florida International University,
Miami FL 33199, USA
matheek@fiu.edu

Abstract. This paper presents an innovative, adaptive variant of Kohonen's self-organizing maps called ASOM, which is an unsupervised clustering method that adaptively decides on the best architecture for the self-organizing map. Like the traditional SOMs, this clustering technique also provides useful information about the relationship between the resulting clusters. Applications of the resulting software to clustering biological data are discussed in detail.

1. Introduction

In today's data-driven world, it has become increasingly important to analyze large amounts of data in order to extract information from it. Such data analysis is now an integral part of genomic and proteomic studies. Data analysis methods can be either exploratory or confirmatory, based on the availability of appropriate models for the data source. Cluster analysis, a dominant technique of exploratory data analysis [1], aims to group a collection of objects into subsets or "clusters", such that those within each cluster are more closely related to one another than objects assigned to different clusters [2].

Data clustering schemes can be described in terms of the following three steps: (1) feature selection or extraction, (2) similarity (or dissimilarity) computation, and (3) clustering. Clustering methods can be broadly classified as hierarchical or partitioning. Strategies for hierarchical clustering divide into two basic paradigms: agglomerative (bottom-up) and divisive (top-down) [2]. Typical partition clustering algorithms include: K-means clustering, and Self-Organizing Maps (SOM). K-means is an iterative descent clustering algorithm [3], while the SOM can be viewed as a constrained version of K-means clustering or a neural network [4, 5]. SOMs have the advantage that it is possible to easily display the output as a two dimensional grid of samples.

The main drawback of the existing algorithms is that they require either specifying the number of clusters in advance (K-means clustering and SOM), or leave the decision of the number of clusters to the user (hierarchical clustering). To address this shortcomings of fixing *a priori* the number of clusters, Pelleg and Moore proposed a new algorithm called X-means [6], a modification of the traditional K-means algorithm. Instead of specifying the number of clusters in advance, X-means only requires a range of the number of clusters, and then searches the space of cluster locations and the number of clusters using the Bayesian Information Criterion (BIC) measure.

In this paper, we propose a new clustering algorithm: Adaptive Self-Organizing Maps (ASOM), which corresponds to SOMs in the same way that X-means does to K-means. The architecture of the SOM is adaptively modified using a variant of the BIC measure. ASOM only requires a range of architectures to be specified in advance. Our results showed that ASOM is an efficient and reliable tool for cluster analysis of biological data. It turns out to be much more stable than X-means, and as with SOMs, also reveals potential neighborhood relationships among the resulting clusters.

It is well known that it is difficult to ascertain the validity of inferences drawn from the output of these clustering applications. We discuss simple tools that efficiently display information about significant clusters.

1.1. CONCEPTS AND DEFINITIONS

Samples (or patterns) for clustering are represented as a vector of d measurements or features, i.e., as a point in *d*-dimensional space: $x = (x_1, . . . x_d)$. A cluster is simply a set of samples. The centroid of a cluster is denoted by μ, and is obtained by doing the coordinate-wise averaging of the points in the cluster. Let *N* denote the number of clusters. A distance measure or a similarity measure is a metric on the feature space used to quantify the similarity of the samples. Any set of *n* samples can be viewed as an $n \times d$ matrix. Thus the input to a cluster analysis is an ordered pair (*D*, *s*), where *D* is the matrix of samples and *s* is the similarity measure. The output from the clustering algorithm is a partition $\Lambda = \{\mathrm{G}_1, \ldots, \mathrm{G}_\mathrm{N}\}$, where the clusters G_k, $k = 1, \ldots, N$, are subsets of *D*, such that $G_1 \bigcup G_2 \cdots \bigcup G_k = D,$ and $G_i \bigcap G_j = \Phi, i \neq j$.

1.2. Similarity Measures

A measure of similarity between samples is fundamental to the definition of a cluster, and the quality of clustering depends on its choice. The most popular metric for continuous feature spaces is the *Minkowski metrics*, defined as follows:

$$d_p(X_i, X_j) = \left(\sum_{k=1}^{d} \left| x_{i,k} - x_{j,k} \right|^p \right)^{1/p} = \left\| X_i - X_j \right\|_p$$

When $p = 2$, this is the Euclidean distance metric. Other popular similarity measures include: *Pearson Correlation Coefficient*, *Rho*, *Dice*, *Jaccard*, *Simpson*, and others.

1.3. Evaluating Clusters

Clustering is considered difficult because data can reveal clusters with different shapes and sizes in a *d*-dimensional feature space. To compound the problem further, the number of clusters in the data often depends on the resolution (fine vs. coarse) with which we view the data. The examples in Figures 1(a) and (b) show that it is possible to get two different sets of clusters depending on whether one wants 3 clusters or 4. Both sets of clusters appear to be equally reasonable, making it necessary to evaluate clustering results objectively and quantitatively [7]. If, for the given data set, *a priori* grouping information is available, then *entropy* can be used to evaluate the clustering results. However, its chief limitation is that it can only be used to compare clusters with the same architecture (i.e., same number of clusters).

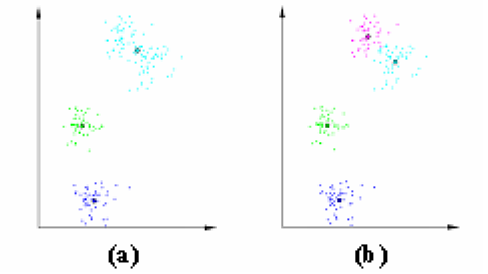


Fig. 1. Input can be partitioned into 3 or 4 clusters

In practice, we may not have *a priori* grouping information. In such cases, the *Bayesian Information Criterion* (BIC) measure can be used to compare the clustering results [8]. The BIC measure is based on the maximization of a log-likelihood score [2]. For *X*-means and the proposed algorithm, ASOM, the BIC measure is also used as the criterion for model selection. For a given model M, the BIC measure is given by $BIC(M) = l(D) - P / 2 \bullet (\log R)$, where l is the log-likelihood of the data according to model *M* taken at the maximum likelihood point, and *P* is the effective number of parameters. Further details may be found in Pelleg and Moore [6]. For finite samples, BIC often chooses simple models to avoid placing heavy penalty on complexity. As the size of samples increases, the probability that the BIC measure favors the correct model also increases [2].

2. Adaptive Self-Organizing Maps (ASOM)

One of the shortcomings of SOM algorithm is its fixed network architecture [9]. For a large dataset, it is very difficult to guess the right architecture. This has motivated the development of a number of adaptive variants.

Growing Grid introduces the notion of a *resource* [10]. This is associated with each unit and used to gather statistical information, which is then used in each adaptation step, to decide where in the map a new row or column is to be inserted. Performance of the

Growing Grid is superior to conventional SOM [11]. Since its parameters are constant, the user does not need to choose a "cooling schedule", as in the conventional SOM. However, when the map grows larger, the algorithm is likely to split more than necessary.

The *Growing Hierarchical Self-Organizing Map* (GH-SOM) [9] uses a hierarchical structure with multiple layers, where each layer consists of a number of independent SOMs. Every cell in an SOM from one layer may be added to the next layer of the hierarchy. The adaptation steps are similar to that of Growing Grid, with the difference that it uses a decreasing learning rate and a decreasing neighborhood radius. The mean quantization error (MQE) of the map is used to decide if a new level of the hierarchy is to be created. Due to its hierarchical structure, GH-SOM shows a lot of structural detail. But the parameters remain hard to choose.

2.1. Algorithm Adaptive Self-Organizing Map (ASOM)

Given lower and upper bounds on the number of columns and rows (col_{min}, row_{min}, col_{max}, and row_{max}, respectively), ASOM starts with the smallest architecture. The steps in each iteration are described below.

1. Run SOM algorithm on current architecture: *colcurr* × *rowcurr*.
2. Split each "parent" column (and row), and run SOM of architecture 2 × *rowcurr* on only the data belonging to the parent. Select the architecture that improves the BIC score by the largest amount.
3. If maximum of rows or columns are exceeded, then stop and report the best scoring model during the search.
4. Else update *colcurr* and *rowcurr*. Go to step 1.

In the splitting step, the algorithm decides how to update the architecture and the cluster centers. The algorithm considers every column separately and checks the value of splitting it into two columns. The rows are also considered in a similar fashion. The architecture with the highest BIC score is then chosen for the next iteration. The algorithm continues until the upper bound on the number of rows or columns is reached. The map with the highest BIC score is finally reported.

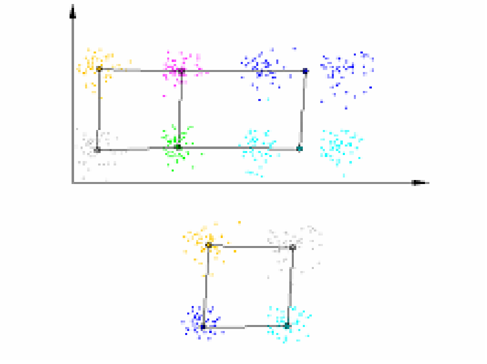
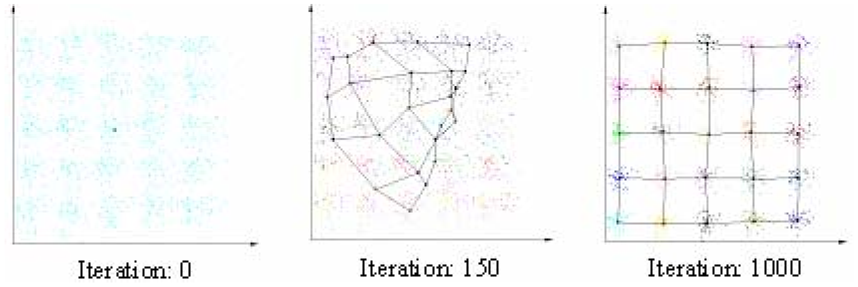


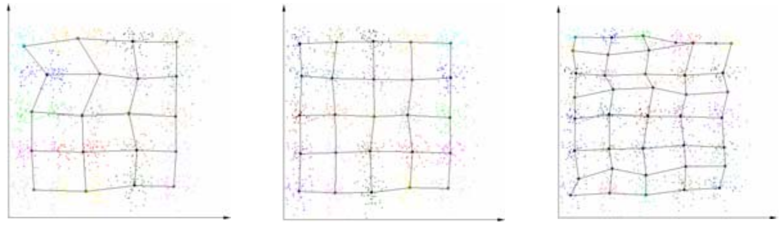
Fig. 2. An example of 2×3 grid being updated to a 2×4 grid. The upper part shows a set of points partitioned into 6 clusters organized into a 2×3 grid. The last column is further partitioned as shown in the lower figure to obtained a 2×4 grid

Figure 3 shows the location of the cluster centers as the ASOM algorithm progresses for a specific chosen architecture. The initial cluster centers are selected at random, and the structure of the network becomes visible after several iterations. As the algorithm progresses, the cluster centers spread out to make meaningful clusters. Figure 4 shows the optimal clustering obtained for each chosen grid architecture. In this example, the best architecture was a 5×5 grid, which had the largest BIC value among the architectures investigated for that data set.



Iteration: 0 Iteration: 150 Iteration: 1000

Fig. 3. The positions of the cluster centers (or prototypes) during the execution of the ASOM algorithm for a specific architecture (5×5 grid architecture). More details in [15]



5 × 4 BIC: 513.26 5 × 5 BIC: 795.98 5 × 8 BIC: 475.05

Fig. 4. Experiments with data consisting of points in two-dimensional space are shown above. The maximum number of rows and columns were set to 8. The reference vectors are shown as large black dots, and the neighboring cluster centers are connected by an edge. The numbers at lower right-hand corner indicate the architecture and corresponding BIC score. More details in [15]

3. Experiments And Results

Four clustering algorithms, *K*-means, *X*-means, SOM, and ASOM, were implemented in Java. In order to visualize the whole process, a visualization package was also implemented using Java Swing.

3.1. Rat Central Nervous System (CNS) Dataset

The rat CNS data set, from Wen *et al.*, contains the expression levels of 112 genes during rat central nervous system development over nine time points [12]. As suggested by Wen *et al.*, the raw data set was normalized by the maximum expression level of each gene, and then augmented with slopes (differences between consecutive time points) to take into account offset but parallel patterns. The dataset is a matrix with 112 genes and 17 conditions after the above preprocessing.

The dataset was provided as input to the ASOM software. The best BIC measure for this set of data was obtained with the 2×3 grid architecture. Each grid has one non-empty cluster, and thus the data set was clustered into six clusters, Wave 1, Wave 2, Wave 3, Wave 4, Constant, and Other [12]. Three other clustering methods in our software were also tested with the following parameter settings: 6 for the number of clusters in K-means and 2×3 for grid architecture in SOM. In order to further compare clustering results, mean normalized expression levels of all gene clusters were plotted. The mean expression plot for one gene cluster (wave 3) is shown in Figure 5 (Plots of all six clusters provided in supplemental website [15]). Average profiles from the four approaches look very similar across all the six clusters. The clusters capture majority significant genes when compared with results in Wen *et al.*.

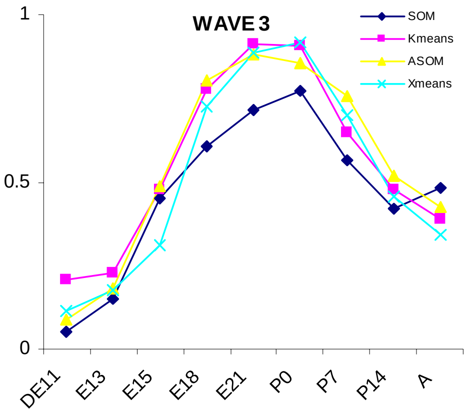


Fig. 5. Mean normalized gene expression levels of gene clusters generated by ASOM, SOM, K-means, X-Means

3.2. Yeast Cell Cycle Dataset

The ASOM algorithm was also tested on the yeast cell cycle data of Spellman *et al* [13]. A total of 799 genes were identified as being regulated by cell cycle, and were used as input. To assess the classification capability of the ASOM clusters, gene ontology information was used to evaluate whether the clusters have significant enrichment of one or more function groups (below ontology level 2); this was done using GoMiner [14]. Table 1 shows details of 5 typical clusters with enriched functional groups. For example, cluster 1, with 42 genes, was enriched by DNA and nucleic-acid binding genes. Enriched function groups included helicase activity, cytokinesis, transferase activity, transcription regulator activity and others, suggesting that the ASOM clusters are biologically meaningful (Details of all 8 clusters provided in supplemental website [15]).

Table 1: Enrichment of ASOM clusters by GO function category.

Cluster	# of Genes	Enriched functional category (total genes)	Clustered genes	-log10 (p-value)
1	42	nucleic acid binding (37) DNA binding (27)	16 15	12 15
2	130	cell proliferation (77) cell cycle (66)	34 33	9 10
3	20	cell proliferation (77) cytokinesis (14)	10 6	5 6
4	53	helicase activity (17) DNA helicase activity (16)	11 8	9 5
5	26	amino acid metabolism (26)	10	9

3.3. Experiments with Synthetic Data

Random data sets were generated with coordinates from the Gaussian distribution. Two types of data sets were generated. The first one consisted of random point sets where the cluster centers were organized into a pre-specified number of rows and columns. An example of such a data set was shown in Figure 4. The second one consisted of randomly generated point sets where the cluster centers were manually provided by a user. The results showed that the ASOM algorithm was extremely successful in identifying the clusters that were present in the clustered synthetic data that was generated.

4. Discussion

We have developed a new, adaptive variant of Kohonen's self-organizing maps. We apply it to two different gene expression datasets. The proposed adaptive SOM permits the detection of the best architecture for the self-organizing map effectively. The software package provides effective visualization tools to facilitate the analysis. The software is available from the authors upon request.

Supplemental Website: http://biorg.cs.fiu.edu/ASOM

Acknowledgements: Research of GN was supported by NIH Grant P01 DA15027-01.

References

1. Jain, A.K., M.N. Murty, and P.J. Flynn, Data Clustering : A Review. ACM Compting Surveys, 1999. **31**(3): p. 264-323.
2. Hastie, T., R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. Springer Series in Statistics. 2001, New York: Springer. 583.
3. Nagy, G., State of the art in pattern recognition. Proc. IEEE, 1968. **56**: p. 836-862.
4. Kohonen, T., Self-organized formation of topologically correct feature maps. Biological Cybernetics, 1982: p. 43.
5. Kohonen, T., Self-Organizing Maps. 1995.
6. Pelleg, D. and A. Moore. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. in Proceedings of the Seventeenth International Conference on Machine Learning. 2000.
7. Jain, A.K. and R.C. Dubes, Algorithms for Clustering Data. Prentice-Hall advanced reference series. 1988: Prentice-Hall, Inc., Upper Saddle River, NJ.
8. Schwarz, G., Estimating the dimension of a model. Ann. Statist., 1978. **6**(2): p. 461-464.
9. Dittenbach, M., D. Merkl, and A. Rauber. The Growing Hierarchical Self-Organizing Map. in Proc. Intl. Joint Conf. on Neural Networks (IJCNN'00). 2000.
10. Fritzke, B., A growing neural gas network learns topologies, in Advances in Neural Information Processing Systems 7, T.K. Lean, Editor. 1995, MIT Press: Cambridge MA. p. 625-632.
11. Fritzke, B., Kohonen feature maps and growing cell structures - a performance comparison, in Advances in Neural Information Processing Systems 5, J. Cowan, Editor. 1993, Morgan Kaufman Publishers: San Mateo, CA. p. 123-130.
12. Wen, X., et al., Large-scale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci U S A, 1998. **95**(1): p. 334-9.
13. Spellman, P.T., et al., Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell, 1998. **9**(12): p. 3273-97.
14. Zeeberg, B.R., et al., GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol, 2003. **4**(4): p. R28.
15. Additional results and supplemental information for ASOM. http://biorg.cs.fiu.edu/ASOM