Self-Adaptive Neural Networks Based on a Poisson Approach for Knowledge Discovery

Haiying Wang, Huiru Zheng, Francisco Azuaje
School of Computing and Mathematics, University of Ulster at Jordanstown
Newtownabbey, Co. Antrim, Northern Ireland, UK, BT37 0QB
{hy.wang, h.zheng, fj.azuaje}@ulster.ac.uk

Abstract

The ability to learn from data and to improve its performance through incremental learning makes self-adaptive neural networks (SANNs) a powerful tool to support knowledge discovery. However, the development of SANNs has traditionally focused on data domains that are assumed to be modeled by a Gaussian distribution. The analysis of data governed by other statistical models, such as the Poisson distribution, has received less attention from the data mining community. Based on special considerations of the statistical nature of data following a Poisson distribution, this paper introduces a SANN, Poisson-based Self-Organizing Tree Algorithm (PSOTA), which implements novel similarity matching criteria and neuron weight adaptation schemes. It was tested on synthetic and real world data (serial analysis of gene expression data). PSOTA-based data analysis supported the automated identification of more meaningful clusters. By visualizing the dendrograms generated by PSOTA, complex inter- and intra-cluster relationships encoded in the data were also highlighted and readily understood. This study indicate that, in comparison to the traditional Self-Organizing Tree Algorithm (SOTA), PSOTA offers significant improvements in pattern discovery and visualization in data modeled by the Poisson distribution, such as serial analysis of gene expression data.

1 Introduction

Knowledge discovery has been defined as a nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [Fayyad *et al.*, 1995]. Data mining is a particular step in this process, which involves the application of specific algorithms for extracting patterns from data [Fayyad *et al.*, 1996; Fayyad *et al.*, 1997]. There are a wide variety of techniques suitable for various data mining tasks. From the knowledge discovery perspective, unsupervised learning-based clustering analysis has become a fundamental approach, which has resulted in a large number of clustering techniques. Examples of powerful and meaningful techniques include the development of self-adaptive neural networks (SANNs)-based clustering models. This paper focuses on this clustering principle because SANNs have demonstrated several unique and interesting features in data mining and knowledge discovery.

1.1 SANNs: Overview of Principles, Applications and Limitations

SANNs represent a family of unsupervised learning models, which follow the basic principle of the *self-organizing feature map* (SOM) [Kohonen, 1995] with a focus on adaptive architecture. A key advantage of these models is that they allow the shape, as well as the size, of the network to be determined during the learning process rather than by a predetermined grid of neurons. For example, the *Growing Self-Organizing Map* (GSOM) [Alahakoon *et al.*, 2000] is initialized with a map of 2 x 2 neurons and new neurons are incrementally grown from a boundary neuron where the network exhibits a large cumulative representation error. After learning, GSOM can develop into different shapes depending on the clusters present in the data. In the *Growing Cell Structures* (GCS) [Fritzke, 1994], the initial topology consists of a two-dimensional output space where the neurons are arranged in triangles. A new neuron is inserted by the splitting of the longest edge emanating from the neuron with maximum accumulated error. GCS performs an adaptation of the overall structure in those regions that represent large portions of the input data. Based on both the SOM and the GCS principles, Dopazo and Carazo [1997] proposed the *Self-Organizing Tree Algorithm* (SOTA). One of the main contributions of SOTA is that the output space is arranged following a binary tree topology, in which the number of output neurons is adapted to the intrinsic characteristics of the input data [Dopazo and Carazo, 1997; Herrero *et al.*, 2001].

Due to its dynamic, self-evolving nature, the resulting maps of SANN can reveal relevant patterns from the underlying data in a more meaningful fashion. For example, due to the ability to separate neurons into disconnected areas, the GCS can produce explicit representations of cluster boundaries. Thus, patterns hidden in the data become more apparent [Fritzke, 1994]. The GSOM, on the other hand, can indicate the patterns in the data by its shape and attract attention to such areas by branching out. Such a flexible struc-

ture may provide a meaningful visualization of clusters in the data [Alahakoon *et al.*, 2000].

SANNs are well adapted to various application domains. For instance, they represent a promising way to improve biomedical pattern discovery and visualization. The *Growing cell structure visualization toolbox* [Walker *et al.*, 1999], for example, is an implementation of GCS networks in the *MatLab 5* computing environment. This tool has been commonly used for the visualization of high-dimensional biomedical data. SOTA has been shown to be capable of performing pattern discovery across various biomedical domains. Dopazo and Carazo [1997] used SOTA to cluster aligned sequences. It has also been applied to the supervised [Wang *et al.*, 1998a] and unsupervised [Wang *et al.*, 1998b] classification of protein sequences. More recently, Herrero and colleagues [Herrero *et al.*, 2001] extended its application to the analysis of gene expression data derived from DNA array experiments.

However, most of current SANNs are based on some heuristic criteria that take the accumulated quantization error into account to guide the growth of neural networks. For example, during the learning process GSOM [Alahakoon *et al.*, 2000] applies Euclidean distance to determine the winning neuron for each input data and a cumulative error is calculated for each winning neuron using the Euclidean distance-based metric. In the growing phase, the network keeps track of the highest error value and determines when and where to grow a new neuron. Such a criterion, however, is not suitable for problems in which the data are better approximated by a Poisson distribution (i.e. a mixture of separate Poisson-distributed data sources), such as phenomena in which events are observed a number of times over specific intervals. Emerging problem domains in bioinformatics such as the study of *Serial Analysis of Gene Expression* (SAGE) data [Velculescu *et al.*, 1997] may also be approximated by a Poisson distribution. Euclidean distance-based clustering analysis has demonstrated poor performance in these domains [Cai *et al.*, 2004]. Without taking into account the statistical nature of the data during the learning process, the full potential of SANNs may not be realized.

1.2 Objectives of This Study

This paper aims to present a new SANN model, which takes into account the specific statistical nature of data approximated by a Poisson distribution, to improve data mining and knowledge discovery. The main objective of this study is, based on the incorporation of a Poisson statistics-based distance function, to develop a SANN model tailored to the data approximated by a Poisson distribution. This required the implementation of new strategies for weight adaptation and network growth.

The remainder of this paper is organized as follows. Section II describes important statistical properties of the Poisson distribution, followed by a detailed description of the new SANN learning algorithm: *Poisson-based Self-Organizing Tree Algorithm* (PSOTA). Two datasets, including synthetic and real world data, are described in Section III. Results and a comparative analysis are presented in Section IV. This paper concludes with the discussion of results and future research.

2 Algorithms and Implementation Protocols

2.1 Statistical Nature of a Poisson Distribution

The Poisson distribution describes a wide range of natural phenomena. This distribution may be used to model the number of events occurring within a given time interval when such events are known to occur with an average rate. The formula for the Poisson probability mass function can be represented as:

$$p(m) = \exp(-\lambda) \times \lambda^m / m! \qquad (1)$$

where $p(m)$ is the probability of observing m occurrences, and λ is the shape parameter that estimates the average number of events in a given time interval.

The Poisson distribution has several unique features. Most distinctively, the mean of any Poisson distribution is equal to its variance. In other words, the larger the value of the mean, the less significant the deviation between a count value observed and its expected value.

2.2 Description of PSOTA

PSOTA is based on the same principle of the SOTA [Dopazo and Carazo, 1997]. Its structure is started by generating an initial network composed of two terminal neurons connected by an internal neuron, as shown in Figure 1(a). The output topology is incrementally constructed by generating two new terminal neurons from the leaf neuron having higher *resources* (measured as the mean distance between the weight of each neuron and all the data samples assigned to this neuron) after each cycle (Figure 1(b) and (c)). For a given training dataset, *T*, consisting of *N* samples, a *learning cycle* consists of a series of learning epochs, within which the network is sequentially presented with each training sample. However, by taking into account the statistical nature of data closely following a Poisson distribution, PSOTA adopts novel matching criteria (1) to determine the winning neuron for each input sample and (2) to update the weight vectors of the winning neuron and its neighborhood.

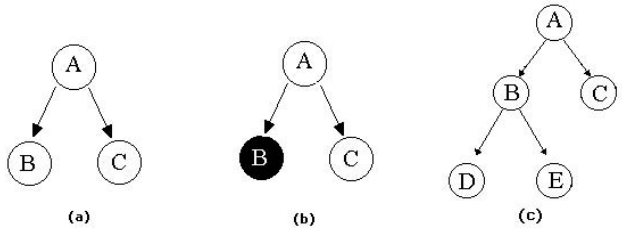


Figure 1: New neurons generation process for PSOTA. (a) The PSOTA initial topology; (b) The accumulation of resources (the heterogeneity of each neuron) during learning process, the neuron marked with a filled circle (neuron B) has the highest cumulative resource after a learning cycle; (c) Neuron B gives rise to terminal neurons D and E (leaf neuron). Thus, D and E are sister neurons, whose ancestor neuron is B.

Matching Criterion for Finding a Winning Neuron, w_c, for a Given Input Vector, x_i

Traditional SANNs, e.g. SOTA, normally apply Euclidean or Pearson Correlation-based distance to determine the winning neuron for each input data. These distance measures have achieved a great success for data approximately following a normal distribution. For data associated with a Poisson distribution, however, these measures have shown poor performance [Cai et al., 2004]. On the basis of the consideration of the statistical nature of a Poisson distribution, two new criteria based on Chi-square statistics and a joint likelihood function are introduced here.

Let $\boldsymbol{x_i}$ be the input vector representing the i^{th} input sample, $\boldsymbol{w_j}$ be the associated weight vector of the j^{th} neuron, and the index k indicate k^{th} value of n-dimensional vector, the winning neuron represented by the subscript c can be determined by the following minimum Chi-square statistics–based distance matching criterion.

$$d_\chi(i,j) = \sum_{k=1}^{n} \left((x_{i,k} - \hat{x}_{i,k})^2 \Big/ \hat{x}_{i,k} \right) \tag{2}$$

$$d_\chi(i,c) = \min d_\chi(i,j), \quad \forall\, j \tag{3}$$

Given that in the Poisson distribution, the probability of a number of events occurring within a given time interval is considered to be independent of events that occurred in previous time intervals, the winning neuron can be also determined by using the maximum joint likelihood function-based matching criterion:

$$p(i) = \prod_{k=1}^{n} (\exp(-\hat{x}_{i,k}) \times \hat{x}_{i,k}^{\,x_{i,k}} \Big/ x_{i,k}!) \tag{4}$$

$$p(i,c) = \max(p(i,j)), \quad \forall j \tag{5}$$

where $\hat{x}_{i,k}$ is the expected value of $x_{i,k}$. After completing a learning process, each weight vector in the SOTA coincides with the centroid of the respective cluster of the input data. Moreover, we are interested in grouping samples with similar relative values rather than the absolute values. Thus, the expected k^{th} value of i^{th} input given the weight vector of j^{th} neuron, $\hat{x}_{i,k,j}$, is calculated as follows:

$$\hat{x}_{i,k,j} = (w_{j,k} \Big/ \sum_{k=1}^{n} (w_{j,k})) \times \sum_{k=1}^{n} x_{i,k} \tag{6}$$

This equation is used, together with Equations (2) and (3) or (4) and (5), to find a winning neuron. The matching criteria expressed in Equations (2) to (6) suggests that when the expected values are large, the deviation between actual and expected count values become less significant. This is consistent with an important property of the Poisson model, i.e. the variance of the dependent variable equals its mean, which is totally ignored by using Euclidean (or other traditional) distance-based error calculation approaches.

Weight Adaptation for a Winning Neuron and Its Topological Neighborhood

Like other SANNs, once the winning neuron has been identified for each input sample, it is necessary to define a method to update the weight vectors of the winning neuron and its neighborhood in order to better match the input vectors and fulfill the overall clustering goals. In PSOTA, the main goal is to assign an input data to a neuron with the most similar relative vector. Thus, instead of performing weight adaptation simply based on absolute values, like in other SANNs (e.g. traditional SOTA), we propose the following weight adaptation strategy, which updates all relative weight values within the neighborhood, $N_c(t)$, of a winning neuron, c, according to the given i^{th} input.

$$w_{j,k,i}(t+1) = \begin{cases} w_{j,k}(t) + \alpha(t) \times (x_{i,k}(t) \times \dfrac{\sum_{k=1}^{n} w_{j,k}(t)}{\sum_{i=1}^{n} x_{i,k}(t)} - w_{j,k}(t)), \\ \qquad\qquad j \in N_c(t) \\ w_{j,k}(t), \qquad\qquad otherwise \end{cases} \tag{7}$$

where $w_{j,k}(t)$ and $w_{j,k,i}(t+1)$ are the k^{th} weight values of neuron j before and after the adaptation at iteration t. $N_c(t)$ and $\alpha(t)$ represent the neighborhood of the winning neuron c and learning rate at iteration t respectively. The reader is referred to [Dopazo and Carazo, 1997; Herrero *et al.*, 2001] for a more detailed description of the selection of $N_c(t)$ and $\alpha(t)$ for SOTA-based algorithms. The learning algorithm of PSOTA is summarized in Table 1.

1: Initialization
2: **Repeat** cycle
3: **Repeat** epoch
4: For each input sample,
5: Find the winning neuron for each input using (2) to (6)
6: Update the winner and its neighbors using (7)
7: Calculate the resource for each neuron.
8: **Until** a cycle finishes: relative increase of the error between two consecutive epochs falls below a given threshold.
9: Grow new neurons from the one having higher resource
10: **Until** The highest resource reaches a given threshold.

Table 1: A summary of PSOTA learning algorithm

2.3 Implementation Protocols

Both PSOTA and SOTA models were implemented within the software development framework provided by the open-source platform, *TIGR MeV* [Saeed *et al.*, 2003]. Unless indicated otherwise, the learning parameters for PSOTA and SOTA are: the maximum number of learning cycles = 5, the maximum number of learning epochs = 1000, and the learning rates for the winning, ancestor and sister neurons are set to 0.01, 0.005, and 0.001 respectively [Herrero *et al.*, 2001].

3 The Datasets Under Study

Two datasets, including synthetic and real world data, were used to assess the PSOTA algorithm.

3.1 Synthetic Data

The dataset was obtained from a study published by Cai *et al.* [2004]. It included 80 synthetic samples, each represented by five simulated values at five time points: T1, T2, T3, T4, and T5. All the simulated values are generated independently using Poisson distributions. Based on the models they are generated from, the 80 samples are divided into four groups PA, PB, PC, and PD with 12, 16, 24 and 28 samples respectively. Samples within the same group have similar profiles determined by the relative count numbers across different time points, as illustrated in Figure 2, which shows the profiles of Groups PA and PB.

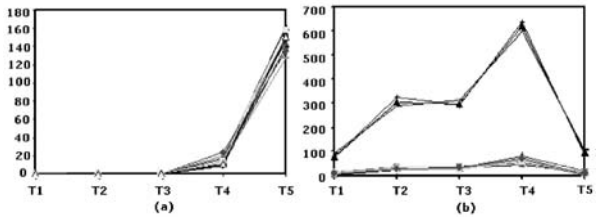


Figure 2: An example of profiles for synthetic data. (a) Group A (12 samples). (b) Group B (16 samples). Five time points are shown on the x-axis, while the y-axis represents the absolute simulated count numbers. Different greys stand for different samples.

3.2 Mouse Retinal Gene Expression Data

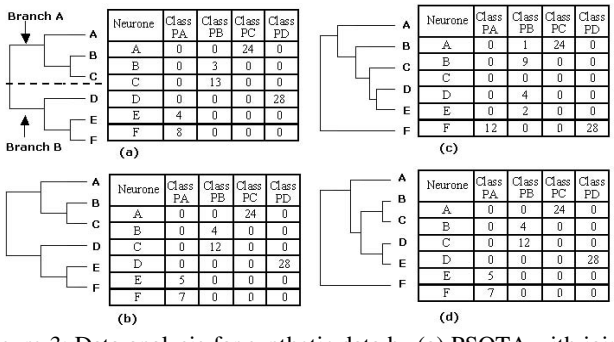
To further evaluate the algorithm, a real world dataset generated by SAGE in mature and developing mouse retina was analysed [Blackshaw *et al.*, 2004]. SAGE is a global gene-expression profiling technique designed to provide quantitative measures of gene expression in a particular cell or tissue obtained from different developmental stages or pathological processes [Velculescu *et al.*, 1997]. The result of a SAGE experiment, known as a *SAGE library*, is a list of *tags* and the number of times each tag is observed within a biological sample. It has been suggested that the count values of SAGE tags observed in a specific library can be approximated by a Poisson distribution [Cai *et al.*, 2004].

Such distributions tend to be independent across different tags and libraries. A detailed description of the SAGE technique and relevant applications can be found in Velculescu *et al.* [1997]. The dataset under study includes 10 murine SAGE libraries from developing retina taken at 2-day intervals from embryonic day 12.5 to postnatal day 10.5 and adult retina: E12.5, E14.5, E16.5, E18.5, P0.5, P2.5, P4.5, P6.5, P10.5, and Adult. The reader is referred to Blackshaw *et al.* [2004] for a full description of the generation and biological meaning of these libraries. A subset of 92 tags with known biological functions and distinctive expression patterns were analyzed. On the basis of their biological functions and temporal expression patterns during retinal development, these 92 tags may be divided into six distinctive clusters: (1) *P10Cluster* (14 tags), which show high but transient expression at P10.5; (2) *PrenrichedCluster*, which includes 21 tags that were found to be highly enriched in photoreceptor (PR)-enriched genes; (3) *PerinatalCluster* (11 tags), whose expression peak appears around P0.5; (4) *CystallinCluster*, which includes 12 cystallin proteins; (5) *EmbryonicCluster* (17 tags), which show strong expression levels during embryonic days, and (6) *NeuroD4Cluster*, which includes 13 tags having similar expression patterns as gene *NeuroD4*. These "natural clusters" have been defined as key functional classes in previous studies [Blackshaw *et al.*, 2004; Blackshaw *et al.*, 2001]

4 Results

4.1 Analysis of Synthetic Data

We first implemented a comparative analysis using the synthetic data with SOTA (Figure 3). By incorporating Poisson statistics-based distance into the learning process, PSOTA correctly constructed a dendrogram that reflect significant inherent relationships between the data samples. For example, PSOTA with joint likelihood function-based distance produced a hierarchical topology with 6 terminal neurons, each neuron uniquely representing one natural class (see the class distribution over terminal neurons given in the right panel in Figure 3(a)). Moreover, by visualizing the whole hierarchical clustering process, a more comprehensive picture that highlights the similarity between all the data samples can be obtained. For instance, as can be seen from Figure 3(a), PSOTA first grouped 80 samples into 2 clusters (Branches A and B). All samples from Classes PA and PD are clustered together (Branch A), while all of samples from Classes PB and PC are grouped into Branch B. This is consistent with the characteristics exhibited by this synthetic data. Similar results were obtained when using Chi-square statistic-based distance as shown in Figure 3(b). Clustering analysis with traditional SOTA (based on Euclidean distance and Pearson correlation, Figure 3(c) and (d), however, fails to detect the underlying data structure. For example, SOTA with Euclidean distance groups Classes PA and PD into the same cluster.



Neurone	Class PA	Class PB	Class PC	Class PD
A	0	0	24	0
B	0	3	0	0
C	0	13	0	0
D	0	0	0	28
E	4	0	0	0
F	8	0	0	0

Neurone	Class PA	Class PB	Class PC	Class PD
A	0	1	24	0
B	0	9	0	0
C	0	0	0	0
D	0	4	0	0
E	0	2	0	0
F	12	0	0	28

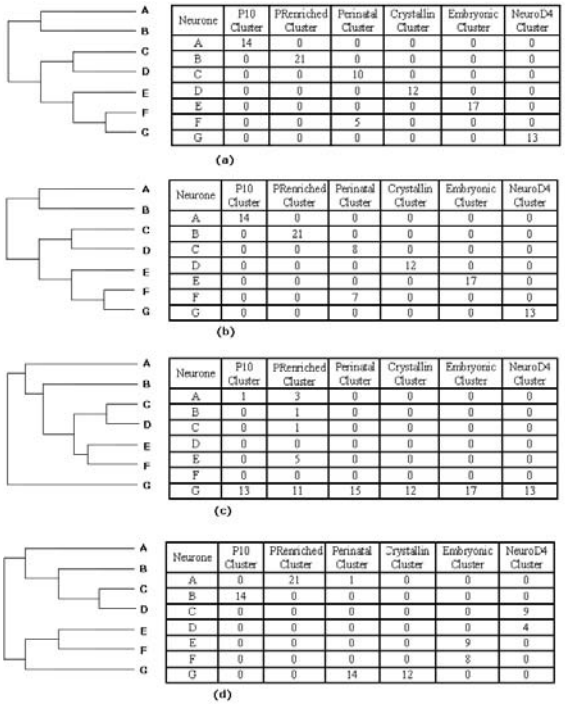
Neurone	Class PA	Class PB	Class PC	Class PD
A	0	0	24	0
B	0	4	0	0
C	0	12	0	0
D	0	0	0	28
E	5	0	0	0
F	7	0	0	0

Neurone	Class PA	Class PB	Class PC	Class PD
A	0	0	24	0
B	0	4	0	0
C	0	12	0	0
D	0	0	0	28
E	5	0	0	0
F	7	0	0	0

Figure 3: Data analysis for synthetic data by (a) PSOTA with joint likelihood function-based distance. (b) PSOTA with Chi-square statistic-based distance; (c) Traditional SOTA with Euclidean distance; (d) SOTA with Pearson correlation-based distance. The left panel on each figure shows the dendrogram obtained by each method, while the right panel shows the class distribution over each neuron.

4.2 Analysis of Mouse Retinal SAGE Data

The outcomes of a comparative analysis of mouse retinal gene expression data with PSOTA and SOTA are illustrated in Figure 4. Only the dendrograms generated by PSOTA, with either joint likelihood function or Chi-square statistics-based distances, correctly depict significant relationships encoded in the SAGE data (Figure 4(a) and (b)). This can be further demonstrated by the analysis of class distributions over the terminal neurons shown in the right panel in Figure 4(a) and (b). For example, 14 P10Cluster tags and 17 EmbryonicCluster tags were grouped together and assigned to the neurons A and E respectively (Figure 4(a)). By contrast, the dendrograms produced by using SOTA with traditional distance measures are less meaningful, especially with Euclidean distance (see Figure 4(c)). This highlights the clear advantages of PSOTA when dealing with datasets that follow Poisson distribution.



(a)

Neurone	P10 Cluster	PRenriched Cluster	Perinatal Cluster	Crystallin Cluster	Embryonic Cluster	NeuroD4 Cluster
A	14	0	0	0	0	0
B	0	21	0	0	0	0
C	0	0	10	0	0	0
D	0	0	0	12	0	0
E	0	0	0	0	17	0
F	0	0	5	0	0	0
G	0	0	0	0	0	13

(b)

Neurone	P10 Cluster	PRenriched Cluster	Perinatal Cluster	Crystallin Cluster	Embryonic Cluster	NeuroD4 Cluster
A	14	0	0	0	0	0
B	0	21	0	0	0	0
C	0	0	8	0	0	0
D	0	0	0	12	0	0
E	0	0	0	0	17	0
F	0	0	7	0	0	0
G	0	0	0	0	0	13

(c)

Neurone	P10 Cluster	PRenriched Cluster	Perinatal Cluster	Crystallin Cluster	Embryonic Cluster	NeuroD4 Cluster
A	1	3	0	0	0	0
B	0	1	0	0	0	0
C	0	1	0	0	0	0
D	0	0	0	0	0	0
E	0	5	0	0	0	0
F	0	0	0	0	0	0
G	13	11	15	12	17	13

(d)

Neurone	P10 Cluster	PRenriched Cluster	Perinatal Cluster	Crystallin Cluster	Embryonic Cluster	NeuroD4 Cluster
A	0	21	1	0	0	0
B	14	0	0	0	0	0
C	0	0	0	0	0	9
D	0	0	0	0	0	4
E	0	0	0	0	9	0
F	0	0	0	0	8	0
G	0	0	14	12	0	0

Figure 4: Data analysis for Mouse SAGE data by (a) PSOTA with joint likelihood function-based distance. (b) PSOTA with Chi-square statistic-based distance; (c) SOTA with Euclidean distance; (d) SOTA with Pearson correlation-based distance. The left panel on each figure shows the dendrogram obtained by each method, while the right panel shows the class distribution over each neuron.

A closer examination of the dendrogram constructed by PSOTA (Figure 4 (a) and (b)) reveals that by monitoring the learning process of PSOTA the potential relevance of inter- and intra-cluster relationships hidden in the data can be readily detected and understood. For example, as shown in Figure 4(a), at the early learning stage, samples belonging to P10Cluster and PRenrichedCluster were actually grouped together, suggesting common patterns between these two classes. The heat maps shown in Figure 5(a) and (b) show that both clusters have strong expression levels at the P10.5 time point. Significant relationships can also be obtained when analyzing relationships between other clusters.

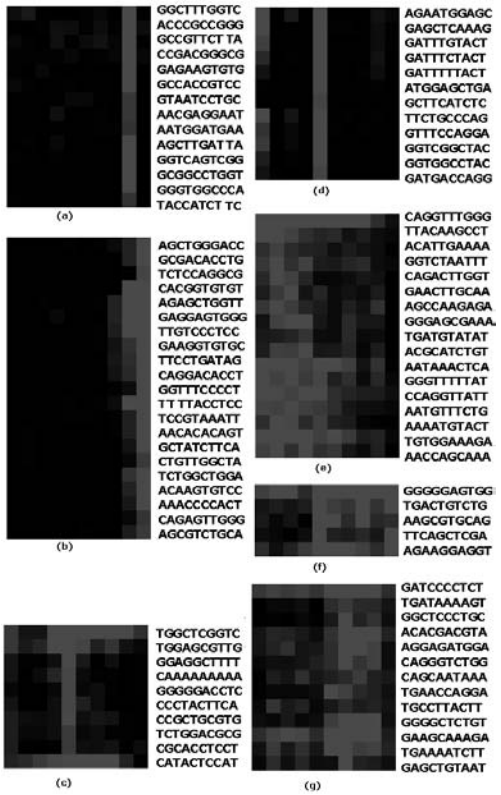


Figure 5: Heat maps (generated by PSOTA) for SAGE tags that fall into (a) Neuron A; (b) Neuron B; (c) Neuron C; (d) Neuron D; (e) Neuron E; (f) Neuron F; and (g) Neuron G, as shown in Figure 4(a). Each row represents expression level of a SAGE tag across SAGE libraries shown as columns in each image. The absolute abundance of each SAGE tag correlates with color intensity, black with the expression level equal to zero. The SAGE tags are displayed on the right side.

5 Discussion and Conclusions

From the pattern discovery perspective, clustering-based techniques have received great attention. However, cluster analysis of data approximated by a Poisson distribution has not been rigorously studied. By incorporating Poisson statistics-based distance functions into the learning process, this paper presented a new SANN model, PSOTA, specially designed to deal with problems modeled by Poisson statistics, such as SAGE data analysis. The results obtained indicate that PSOTA offers several advantages over traditional SANN techniques. Like SOTA [Dopazo and Carazo, 1997], PSOTA not only incorporates some of the advantages demonstrated by hierarchical clustering and SOM, but also it implements unique features such as the generation of clusters at different levels. Moreover, by using new matching criteria to determine the winning neurons and implement weight adaptation, significant improvements in pattern discovery and visualization are accomplished. By visualizing

the dendrogram constructed by PSOTA, complex inter- and intra-cluster relationships encoded in the data may be highlighted and understood.

The fundamental advantages of PSOTA over SOTA are driven by the fact that PSOTA is tailored to the statistical nature of Poisson-distributed data. Equations (6) and (7) include a factor determined by the sum over all the dimensions of i^{th} input and j^{th} weight vectors, which aims to group samples with relative similar profiles (values) into one neuron.

One crucial problem that needs to be further addressed is the optimal determination of learning parameters. Currently, there is no standard way to define, *a priori*, the optimal learning parameters. One possible solution is to combine PSOTA with machine learning-based searching techniques, such as genetic algorithms, to determine optimal parameter values [Jin *et al.*, 2003]. This is part of our future research.

The Poisson distribution has been used to model a wide range of natural phenomena. For example, in bioinformatics, transcription-factor binding sites and SAGE data may be modeled by Poisson statistics. The pattern discovery and visualization techniques described in this paper have the potential to contribute to the improvement of data mining and knowledge discovery in these areas, in which the data represent a number of events occurring within a fixed time interval and when such events are known to occur with an average rate. Nevertheless, if the data do not encode these types of situations other (traditional) methods may be equally recommended.

Acknowledgments

We thank Dr H. Huang at the University of California, Berkeley, for providing synthetic data and for helpful discussions.

References

[Fayyad *et al.*, 1995] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery: an overview. In *Advances in Knowledge Discovery and Data Mining*, Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth Eds., Menlo Park, CA: AAAI Press, 1-34, 1995.

[Fayyad *et al.*, 1996] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11): 27-41, 1996.

[Fayyad and Stolorz, 1997] Usama Fayyad and Paul Stolorz. Data mining and KDD: promise and challenges. *Future Generation Computer Systems*, 13: 99-115, 1997.

[Kohonen, 1995] Teuvo Kohonen. *Self-Organising Maps*. Heidelberg, Germany: Springer-Verlag, 1995.

[Alahakoon *et al.*, 2000] Damminda Alahakoon, Saman K. Halgamuge, and Bala Srinivasan. Dynamic self-organising maps with controlled growth for knowledge discovery. *IEEE Trans. Neural Networks*, 11(3):601-614, 2000.

[Fritzke, 1994] Bernd Fritzke. Growing cell structures -- a self-organising network for unsupervised and supervised learning. *Neural Networks*, 7: 1441-1460, 1994.

[Dopazo and Carazo, 1997] Joaquin Dopazo and Jose M. Carazo. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *Journal of Molecular Evolution*, 44: 226-233, 1997.

[Herrero *et al.*, 2001] Javier Herrero, Alfonso Valencia, and Joaquin Dopazo. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17:126-136, 2001.

[Walker *et al.*, 1999] Andy J. Walker, Simon S. Cross, and Rob F. Harrison. Visualisation of biomedical data sets by use of growing cell structure networks: a novel diagnostic classification technique. *The Lancet*, 354: 1518-1521, 1999.

[Wang *et al.*, 1998a] Huai-Chun Wang, Joaquin Dopazo, Luis G. de la Fraga, Yun-Ping Zhu, and Jose M. Carazo. Self-organising tree-growing network for the classification of protein sequences. *Protein Science*, 7:2613-2622, 1998.

[Wang *et al.*, 1998b] Huai-Chun Wang, Joaquin Dopazo, and Jose. M. Carazo. Self-organising tree growing network for classifying amino acids. *Bioinformatics,* vol. 14, no. 4, pp. 376-377, 1998.

[Velculescu *et al.*, 1997] Victor E. Velculescu, Lin Zhang, Bert Vogelstein, and Kenneth W. Kinzler. Serial analysis of gene expression. *Science*, 276:1268-1272, 1997.

[Cai *et al.*, 2004] Li Cai, Haiyuan Huang, Seth Blackshaw, Jun S. Liu, Connie Cepko, and Wing Wong. Clustering analysis of SAGE data: A Poisson approach. *Genome Biology,* 5: R51, 2004.

[Saeed *et al.*, 2003] Alex Saeed, Vasily Sharov, Joe White, Jerry Li, Wei Liang, Nirmal Bhagabati, *et al.* TM4: a free, opensource system for microarray data management and analysis. *BioTechniques*, 34(2): 374--378, 2003.

[Blackshaw *et al.*, 2004] Seth Blackshaw, Sanjiv. Harpavat, Jeff Trimarchi, Li. Cai, Haiyuan Huang, Winston Kuo, *et al.* Genomic analysis of mouse retinal development. *PLoS Biology*, 2(9), 2004.

[Blackshaw *et al.*, 2001] Seth Blackshaw, Rebecca E. Fraioli, Takahisa Furukawa, and Constance L. Cepko. Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes. *Cell* 107:.579-589, 2001.

[Jin *et al.*, 2003] Hui D. Jin, Kwong S. Leung, Man L. Wong, and Z.-B. Xu. An efficient self-organizing map designed by genetic algorithms for the traveling salesman problem. *IEEE Trans. Systems, Man, and Cybernetics-part B: Cybernetics*, 33(6):877-888, 2003.