Noise-Resistant Bicluster Recognition

Huan Sun[1], Gengxin Miao[1], Yu S. Huang[2], and Xifeng Yan[1]
[1]University of California, Santa Barbara, [2]Center for Neurobehavioral Genetics, UCLA
huansun@cs.ucsb.edu

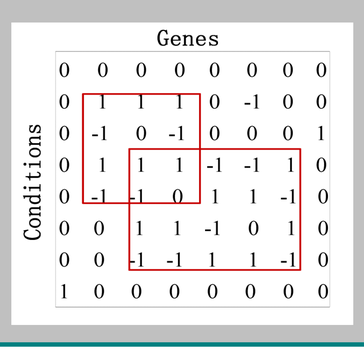
Sponsored in part by the U.S. National Science Foundation under grant IIS-0954125 and by the Institute for Collaborative Biotechnologies through Grant, No. DFR3A-8-447850-23002 from the U.S. Army Research Office

Introduction

- Biclustering is a dominant unsupervised technique for gene expression data analysis, which refers to simultaneously group genes and conditions in the data. Various kinds of algorithms have been proposed over the past decade. It becomes hard to further improve the performance without resorting to a new methodology.
- We propose a two-layer neural network model, named AutoDecoder (AD), to unmask biclusters hidden in gene expression data. AutoDecoder is a generalization of unsupervised neural networks that can automatically learn features from unlabeled data. To the best of our knowledge, this is the first biclustering algorithm that leverages neural network techniques to recover biclusters. **Our software is available at** http://www.cs.ucsb.edu/~huansun/ad

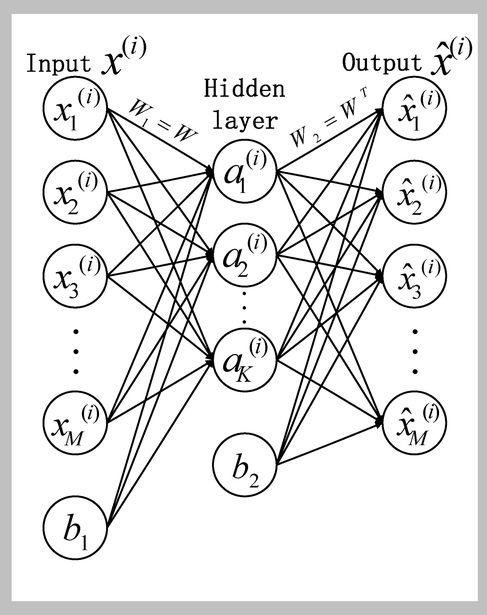
Challenges

- Overlapping in both genes and conditions;
- Severe noise existing in the data;
- Not necessarily full coverage: One gene (condition) might not belong to any bicluster;
- Various kinds of bicluster patterns: Genes (conditions) can be positively and negatively correlated.



Preliminaries: Sparse Autoencoder

· Two-layer neural network



Input: gene expression data

$$X^{M\times N} = [x^{(1)},...,x^{(i)},...,x^{(N)}]$$

Output: recovered data $\hat{X}$, weights W

activation function $a^{(i)} = sigmoid(W * x^{(i)} + b_1)$

activation rate $\hat{\rho}_k = \sum_{i=1}^{N} a_k^{(i)} / N$

Optimization formulation:

$$\underset{W,b_1,b_2}{argmin} \quad H = \frac{1}{2N} * \sum_{n=1}^{N}\sum_{m=1}^{M}(\hat{x}_m^{(n)} - x_m^{(n)})^2 \quad (i)$$
$$+ \beta_2 * KL(\rho\|\hat{\rho}) \quad (ii)$$
$$+ \frac{\lambda}{2} * \|W\|_F^2 \quad (iii)$$

Sparse Autoencoder Uniformly reconstructs the zero part and non-zero part in the data. However, these two parts might be corrupted to different degree.

Framework

• AutoDecoder (AD) optimization formulation

$$\underset{W,b_1,b_2}{argmin} \quad H = \frac{1}{2N}\sum_{i=1}^{N}\sum_{m=1}^{M}[I_{m,i} + \alpha(1 - I_{m,i})](\hat{x}_m^{(i)} - x_m^{(i)})^2 \quad (i)$$
$$+ \beta\, KL(\rho\|\hat{\rho}) \quad (ii)$$
$$+ \frac{\gamma}{2}\sum_{k=1}^{K}\sum_{i\neq j} S_{i,j}^{(condition)}(|W_{k,i}| - |W_{k,j}|)^2 \quad (iii)$$
$$+ \lambda\,\|W\|_1 \quad (iv)$$

- Term (*i*): non-uniform reconstruction;
  I: an indicator matrix;
  To be robust against noise;

- Term (*iii*): $S_{i,j}^{(condition)}$ is the absolute value of cosine similarity between the ***i***th condition and ***j***th condition;
  To be robust against bicluster overlaps;

- Term (*iv*): weight decay regularizer.

• Model solution
- Minimize the objective function through L-BFGS algorithm;
- Derivatives of parameters obtained by backpropagation.

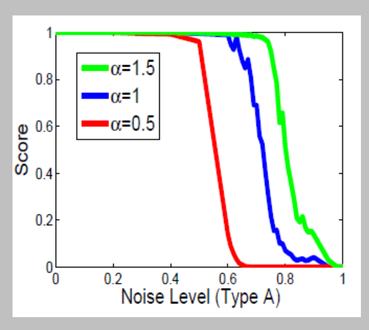
• Bicluster recognition
For each hidden neuron *k*,

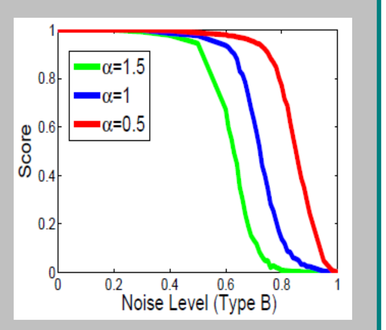
- Gene selection
Pick any gene ***i*** if $a_k^{(i)} > \delta$ $(\delta \in (0,1))$;

- Condition selection
Pick any condition ***m*** if $|W_{k,m}| > \xi$ $(\xi \in (0,1))$.

Characteristics of AutoDecoder

1. Robustness against noise (Term (*i*))
Type A noise: noise outside bicluster patterns;
Type B noise: noise inside bicluster patterns.



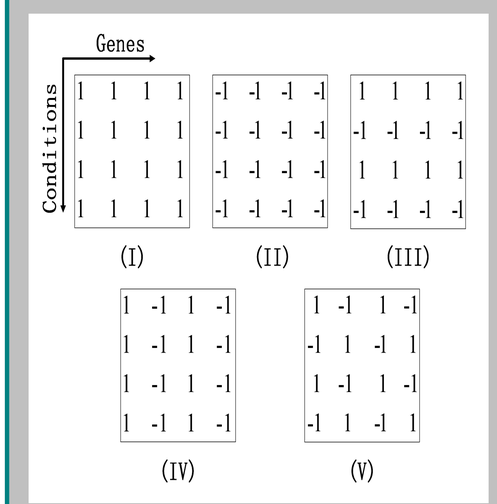


2. Robustness against overlaps (Term (*iii*))

Weights W controls the condition membership. our intuition to incorporate Term (*iii*) is that if two conditions are similar, they should have similar membership. Furthermore, if one condition is similar to conditions in multiple biclusters, it should belong to these biclusters simultaneously, thus resulting in biclusters overlapped in conditions.

Due to the neural network structure, if two genes are similar ($x^{(i)} \sim x^{(j)}$), they will naturally have similar activation values ($a^{(i)} \sim a^{(j)}$).

3. Bicluster patterns



•(I-III) Sigmoid activation

• (IV-V) Tanh activation

Experiments

• Evaluation measures

$$Relevance = \frac{1}{m}\sum_{i=1}^{m} \max_{j\in\{1,2,..,m^*\}} \frac{M_i \cap M_j^*}{M_i \cup M_j^*}$$

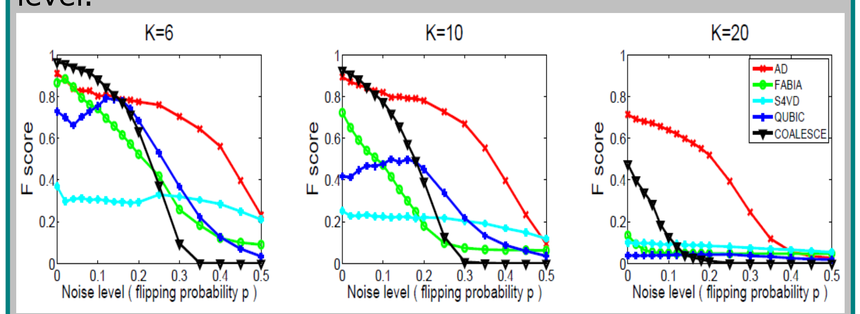
$$Recovery = \frac{1}{m^*}\sum_{i=1}^{m^*} \max_{j\in\{1,2,..,m\}} \frac{M_i^* \cap M_j}{M_i^* \cup M_j}$$

Where, $M^* = \{M_1^*, M_2^*, \cdots, M_{m^*}^*\}$ and $M = \{M_1, M_2, \cdots, M_m\}$ are respectively the true set and the discovered set of biclusters. We define F score as the harmonic mean of Relevance and Recovery.

· Synthetic datasets

Given a matrix size 100x500, bicluster number *K*, for each bicluster, randomly select the number of rows in the bicluster from the range [10,30], the number of columns from the range [50,100].

Each bicluster is originally filled with "1". We flip the 1's inside biclusters to 0's with probability *p* and 0's outside biclusters to 1 or -1 with probability *p*/2. *p* is named noise level.

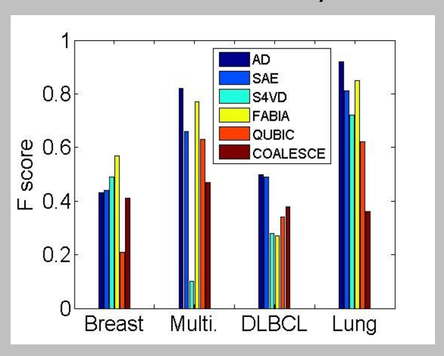


• Real datasets

Breast Cancer, Multiple Tissue, DLBCL, and Lung Cancer.

0. Preprocessing
According to [3], preprocess the original data to {-1,0,1}.

1. Condition clusters evaluation by F score.

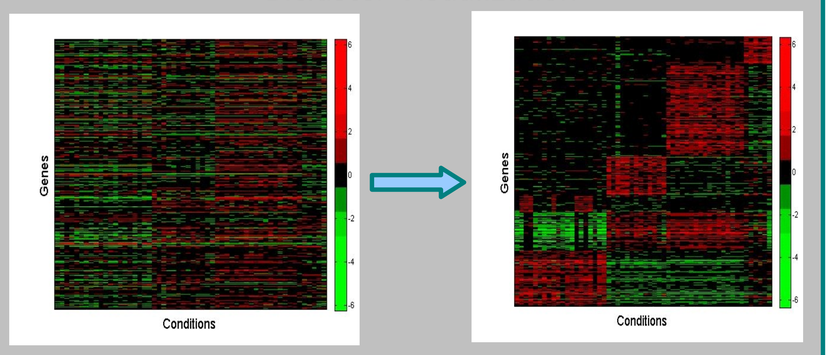


2. Gene clusters evaluation by gene enrichment analysis

Calculate the P-value of each functional class *w.r.t* the gene set in a discovered bicluster. The smaller the P-value, the more significant the bicluster.

AD can generally discover biclusters with P-value less than 10^{-4}, much often less than 10^{-10}.

Bicluster Visualization



Conclusion: AD guarantees the biological significance of the biclusters while improving the performance on condition clusters.

References

[**1**] S. Hochreiter, U. Bodenhofer, M. Heusel, et al. Fabia: factor analysis for bicluster acquisition.*Bioinformatics*, 26(12):1520–1527, 2010.

[**2**] M. Sill, S. Kaiser, A. Benner, and A. Kopp-Schneider. Robust biclustering by sparse singular value decomposition incorporating stability selection. *Bioinformatics*, 27(15):2089–2097, 2011.

[**3**] G. Li, Q. Ma, H. Tang, A. Paterson, and Y. Xu. Qubic: a qualitative biclustering algorithm foranalyses of gene expression data. *Nucleic acids res*, 37(15):e101–e101, 2009.

[**4**] C. Huttenhower, K.T. Mutungu, N. Indik, W. Yang, M. Schroeder, J.J. Forman, O.G. Troyanskaya, and H.A. Coller. Detailing regulatory networks through large scale data integration. *Bioinformatics*, 25(24):3267–3274, 2009.

[**5**] H. Lee, C. Ekanadham, and A. Ng. Sparse deep belief net model for visual area v2. *NIPS*, 20:873–880, 2008.