A Critical Review: Learning Structure in Gene Expression Data Using Deep Architectures, with an Application to Gene Clustering

Paper by: Madhavi Ganapathiraju, Aman Gupta, Haohan Wang
Review by: Priyanka Nigam, Jared Dunnmon, Darvin Yi, Naveen Arivazhagan, Carson Lam

Introduction

The explosion of rapid genome sequencing technologies has led to the availability of sequence-level data at a level never before seen. In this paper, the authors propose the use of stacked Denoising Autoencoders (DAs) originally developed by Vincent et al. (2008) to support such tasks as inference of regulatory pathways and functional gene clustering by leveraging DAs to learn highly explanatory low-dimensional representations of gene expression data. The authors utilize the output of stacked DAs from raw genomic data as input into the unsupervised learning task of gene clustering. They hypothesize that multi-gene interdependencies contained in gene expression profiles lend themselves well to estimating relationships in empirical genomic input distributions, similarly to how DA-style features perform well in imaging contexts due to integration of spatial locality in key features. While the results presented in this paper show promise, there exist a multitude of inconsistencies and unexplained behaviors that suggest additional work will be required to robustly apply DAs to gene clustering problems.

Methods Summary and Critique

This paper's major contribution to the literature is the idea of applying stacked DAs to gene clustering. Although the main components of DAs are similar to those of the standard autoencoder, DAs are intended to directly reconstruct the input from an intermediate low-dimensional representation. DAs generally encode a noise-corrupted input x to a lower-dimensional representation at layer y, and then decode y back to an output z, of the same dimensionality as x. Thus, z is effectively a reconstruction of x from low-dimensional features contained in y. A "stacked" DA refers to aspects of both the architecture and training procedure. Rather than training all layers together via global optimization, the middle layer, layer I, is trained first, without layers i+n. The output from layer i is then used to train layer i+1 until finally the entire deep network is fine-tuned using a global loss function.

The DA presented in this article takes in sequence data and outputs a "denoised" version of this sequence reconstructed from the hidden low-dimensional representation. The input is synthetically corrupted by masking noise (randomly zeroing out some fraction of the input), followed by a single hidden layer, and finally a data reconstruction layer. In the hidden layer, a mapping of the form $y = f(Wx' + b)$ is learned, where the dimensionality is less than that of the input. Finally to regenerate the data, a mapping of the form $z = g(W'y + b')$ is learned, which returns the dimensionality to that of the original input. Note that f and g are nonlinear functions and W, W', b, b' are the learned parameters. The loss function used is the cross-entropy loss between the original uncorrupted data and the regenerated output. The regenerated outputs then are used as inputs to both k-means and spectral clustering algorithms.

Despite the well-developed literature on DAs, there exist several steps in this paper's DA construction methodology, preprocessing routines, training procedures, and clustering mechanisms that remain unclear. For instance, the article fails to mention what nonlinearities are used for f and g in the DA that was constructed. It is also mentions that both k-means and spectral clustering algorithms are used, but there is no specification of which algorithm was used to generate the results shown in the paper or whether the two clustering algorithms showed differing performance. This is an important point, as previous studies have shown that certain algorithms perform better than others on gene clustering problems (Altman and Raychaudhuri, 2001).

There are also several aspects of the architecture that remain unexplored. The main utility of deep learning, for instance, is the extraction of higher level hierarchical representations from raw data. The authors describe how stacked denoising autoencoders can be trained efficiently, but the present results using only a single hidden layer, varying only the number of nodes in the layer. Although deeper networks are not always superior in or appropriate for every application, in 2015 the ability to train networks that are several layers deep became widely accessible via the integration of Graphics Processing Units (GPUs) into deep learning pipelines. Thus, training a slightly deeper network with even a small hyperparameter search space while reporting both performance and training time would enable the reader to understand whether using a deeper stacked DA architecture could provide either better or more robust results. This is particularly true given that previous work has shown improvements (particularly at moderate levels of corruption) by increasing the number of layers from 1 to 3 on the MNIST digit classification problem (Vincent et al., 2010).

Results Summary and Critique

The authors compare the results of gene clustering using the original sequence-level input data, the principal components of this data, and the DA-reconstructed input data on two different yeast gene expression datasets, originally gathered by Spellman et al. (1998) and Cho et al. (1998). The first dataset is a subset of 384 out of 420 RNA expression levels that can be clustered into one cell cycle phase. The ground truth is established by the original authors who classified these 384 genes into one of 5 cell cycle phases: early G1, late G1, S, G2, or M phase. The second dataset is a similar to the first, but contains 237 genes and 4 clusters. The authors argue that genes in the same cluster are likely to either be coexpressed or be part of a cohesive regulatory pathway. However, because these datasets incorporate temporal relationships between the amount of transcription of the different genes, it may be difficult to infer causal or regulatory relationships beyond these from these particular data series.

The evaluation metric used in this study is the Adjusted Rand Index, which is a standard clustering metric (Hubert and Arabie, 1985) that assumes a generalized hypergeometric distribution as the underlying model of randomness. However, it would be very helpful in understanding the mechanism by which clustering results are improved using DA-regenerated data (as opposed to the raw data) if the authors performed an analysis of which pairs of data were misclassified by each of the DA-based, PCA-based, and raw data based methods. At present, it is unclear if a particular type of error dominates inaccuracy in the clustering results. Additionally, it seems dubious to use an arbitrary (albeit biomedically relevant) clustering metric to prove the validity of a dimensionality-reduction method; it would be advisable to demonstrate robust performance using a variety of metrics in future work.

The authors further claim that clustering performance using the DA reconstructed data improves as the number of nodes in the hidden layer increases, and surmise that this may occur due to the network's ability to retain more information when the number of nodes in the low dimensional encoding is increased. However, it is distressing that the improvement is not monotonic with increasing hidden node number, implying that the representation is not getting better with increasing number of model parameters. This is echoed in the PCA-based clustering results, wherein adding more PCA components does not give monotonically better results.

There are other inconsistencies that remain unaddressed, such as the fact that the effect of introducing various levels of masking noise seems unclear. Though both datasets describe cell cycle RNA expression patterns from yeast, there seems to be no clear trend in performance as masking noise is increased. In the first dataset, for instance, adding noise leads to better performance, but sensitivity to the amount of noise is small. Conversely, in the second dataset, no noise seems to outperform or perform just as well as data with noise. Thus, though Vincent et al. (2010) indicate that adding appropriately specified training noise (such as salt-and-pepper noise to MNIST) can improve performance of DAs, the present paper does not seem to have robustly demonstrated this type of improvement in the context of genomic data.

On this point, it would be ideal for the authors to explore different noise models to better understand these effects, especially as the ability to learn missing or corrupted data is highlighted as one of the benefits DAs provides over ordinary autoencoders. This is particularly true given that the noise model chosen, masking noise, does not use any biological knowledge and would appear to model missing data rather than corrupted data. Fundamentally, it is an open question as to whether using masking noise is a useful approach in biological/DNA-based systems. It would seem, for instance, that corruption of certain genes in a random fashion ignores the fact that disruptions may well be spatially correlated on a sequence, etc., and thus that the datasets trained on here may not represent faithful physical noise distributions.

In addition to the above performance and specification drawbacks, it can be argued that the results presented in this paper show no substantial improvement over baselines. Yeung et al. (2001), for instance, compared quality of clusters on the same dataset used here using the both the original data normalized to mean 0 and variance 1 and this data projected onto subsets of its principal component axes. Yeung et al. (2001) cluster using both k-means with Euclidean distance and CAST, a clustering algorithm developed specifically for co-regulated yeasts genes by Ben-Dor and Yakhini (1999). In the Yeung et al. (2001) study, using PCA features does not improve cluster quality. Less encouragingly for the present work, the Adjusted Rand Index of the clustering generated by Yeung et al. (2001) performs very similarly (and in some cases is slightly superior) to the results obtained in this paper using data regenerated by a stacked DA. The change in clustering quality with different clustering algorithms reported by the present paper is also comparable to that reported by Ben-Dor and Yakhini (1999). Thus, while the underlying method for DA-based representation of genomic features presented in this paper may well prove useful in gene clustering problems, it remains to be convincingly shown that DAs can robustly outperform traditional dimensionality-reduction methods in a gene clustering context.

Conclusion

Overall, the need for better clustering within genetic expression is clear, as it enables elucidation of key pathways and potentially implies coexpression, and the introduction of deep autoencoders into this field could enable substantial progress. The preliminary results presented here suggest that well-trained DAs could be a powerful tool for representing genomic data in a manner that enables more effective clustering. However, the results of the present paper are also plagued by a number of critical limitations, including the appropriateness of masking noise in a genomic context, insufficient hyperparameter tuning and training time in model construction, non-monotonicity in performance improvement with model dimension, lack of clarity around the effect of training noise on DA performance, and questionable ability to outperform previous baselines. Further, it is particularly distressing that the authors only report the "best results," and give no sense as to how DAs perform over a broader spectrum of conditions. In the end, then, while the idea of applying DAs to gene clustering may well have substantial merit, the above questions should be addressed before DAs are presented to the scientific community as a robust method to extract features that improve gene clustering.

References

1. Vincent, Pascal, et al. "Extracting and composing robust features with denoising autoencoders." *Proceedings of the 25th international conference on Machine learning.* ACM, 2008.
2. Vincent, Pascal, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." *The Journal of Machine Learning Research* 11 (2010): 3371- 3408.
3. Altman R. and Raychaudhuri S. "Whole-genome expression analysis: challenges beyond clustering." *Current Opinion in Structural Biology* 11 (2001): 340-347.
4. Spellman, Paul T., et al. "Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." *Molecular biology of the cell* 9.12 (1998): 3273-3297.
5. Cho, Raymond J., et al. "A genome-wide transcriptional analysis of the mitotic cell cycle." *Molecular cell* 2.1 (1998): 65-73.
6. Ben-Dor, Amir, Ron Shamir, and Zohar Yakhini. "Clustering gene expression patterns." *Journal of computational biology* 6.3-4 (1999): 281-297.
7. Hubert, L. and Arabie, P. (1985) Comparing partitions. *Journal of Classification*, 193– 218.
8. Yeung, Ka Yee, and Walter L. Ruzzo. "Principal component analysis for clustering gene expression data." *Bioinformatics* 17.9 (2001): 763-774.
9. Bengio, Y. et al. "Greedy Layer-Wise Training of Deep Networks." A*dvances in Neural Information Processing Systems 19 (NIPS'06)*, (B. Schölkopf, J. Platt, and T. Hoffman, eds.), pp. 153- 160, MIT Press, 2007.