ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 51-56

Improving Robustness of Fuzzy Gene Modeling

Robert Reynolds, Habtom Ressom, Mohamad Musavi, Cristian Domnisoru

Department of Electrical and Computer Engineering, University of Maine
201 Barrows Hall, Orono, ME 04469

Abstract. This paper proposes modifications to current fuzzy models of gene interaction. Current algorithms apply all combinations of genes to a fuzzy model (i.e. activator/repressor/target), evaluating how well each combination fits the model. The models are susceptible to noisy signals in the gene expression data. Since the margin of error in current microarray technology can be high, the results generated may not properly reflect valid relationships. This paper investigates different methods of creating fuzzy models. We explore methods of conjunction and rule aggregation that produce valid results while being resilient to minor changes to model input.

1 Introduction

Any attempt to model or analyze DNA microarray data is likely to be affected by noise in the data. There are many potential causes for noise, most originating from the stochastic nature of gene interactions and microarray technology. It has been demonstrated that the minimum detectable difference between Cy3 and Cy5 concentrations is 1.8 [1], implying an error of up to 29%. Attempting to create a model from data that is corrupted by such a high noise ratio is extremely difficult.

There are two methods that can be used to improve the modeling of gene expression. The first is to improve microarray technology to lower the noise ratio. Variance analysis [2] may also provide more accurate readings and knowledge of noise ratios. However, some issues, such as the stochastic nature of the process, may not be eliminated by new technology, and some degree of error will have to be dealt with. The second approach is to use models that are more resilient to minor variations in the expression data. In this paper, we will take the second approach to improve the robustness of Woolf's fuzzy modeling algorithm [3].

Woolf developed a fuzzy model of a known activator/repressor model of gene interaction. Using a normalized subset of *saccharomyces cervisae* data from Cho *et al* [4], Woolf's algorithm applies every possible combination of activators and repressors for each gene to a fuzzy model. The model output is compared to the expression level of the gene. Gene combinations are given scores based upon both the mean-squared error between the model and the target gene and as the variance between the application of the fuzzy rules over the time period. Those combinations of genes that have a low error and cover most of the fuzzy rule base (i.e. have low error and variance scores) are the most likely to exhibit an activator/repressor relationship. Since the model deals with qualitative terms (such as "High" or "low") rather than actual expression levels, it is able to deal with imprecision or low levels of noise with minimal changes to output. However, as will be shown, Woolf's model is susceptible to noise and can result in the creation of inaccurate model outputs.

2 Algorithm

For a given set of membership functions and fuzzy rules, fuzzy models can differ in many ways, including fuzzy conjunction (AND) operations, rule aggregation, and deffuzification. Woolf's algorithm uses addition for fuzzy AND, averaging for rule aggregation, and a modified centroid method for defuzzification. As will be shown, this method produces an unusual output space with sharp gradients in several regions. We decided to use a few different methods and analyze their output and error responses. We used Mamdani's model [5]. Kosko's Standard Additive Model [6], and a hybrid model that attempts to take the best attributes of the Mamdani and SAM models.

Mamdani's model is a classic model that uses the drastic product (i.e., minimum) operation for conjunction and a drastic sum (i.e., maximum) operator for rule aggregation. The model does not provide a set method for defuzzification; it is up to the model designer to decide the method, which can include mean of median (MOM), center of area (COA or centroid), or any other method. A minimum operator on fuzzy inputs makes intuitive sense for gene interaction; the truth value of a particular rule is going to be bound by the minimally-expressed gene. For example: if an activator's expression level is mostly MED and a little HIGH, while a repressor's expression level is mostly LOW and a little MED, the rule "If activator is HIGH and repressor is LOW, then target is HIGH" should be limited completely by the fact that the activator is not particularly HIGH.

Kosko's Standard Additive Model (SAM) uses a product operation for conjunction, a sum operation for aggregation, and the centroid method for defuzzification. Centroid defuzzification is performed by scaling membership functions instead of clipping them at the level of rule application.

Our hybrid model combines attributes of the Mamdani's model and SAM. It uses a product operation for conjunction, a maximum operator for aggregation and a centroid defuzzification that involves scaling as in SAM.

To analyze the output space of each method, we calculated the output of each fuzzy system when presented with all combinations of activator and repressor levels in increments of 0.01. The gradient was calculated using the output surface. The mean and standard deviation of the gradient matrices for each method was calculated.

To analyze the effect of noise on the fuzzy models, the output of the algorithm in [3] with the data from [4] was obtained for each method. The output of the algorithms includes all gene triplets that fit the model well. A Monte Carlo simulation was run on the gene triplets that fit each model well by distorting each time point by a random noise percentage and analyzing the result on the model output. Each triplet was rerun 20000 times with maximum distortion set at different levels from 5-30%. The mean and standard deviation of the model MSE for the 20000 experiments was stored. Each triplet's average MSE was plotted versus its original MSE. Similar plots were made for the standard deviation of MSE for each time point. Linear regressions were found for the average error graphs.

Sensitivity to noise can be related to the regression line; a model is generally less sensitive to noise if the slope of the regression line is close to 1 and the y-intercept is close to 0 (i.e. no average change in MSE due to noise). A slope of 1 would imply that on average, all model outputs would be distorted by the same factor, regardless of the original MSE. Minimizing the y-intercept value is also of interest, but is not necessary for proper operation; if we know that all error scores are offset by the same value due to error, we can simply raise our error cutoffs to get the same results.

Model validation was performed in a similar manner to Woolf and Wang's method. Known activator/repressor complexes were checked in the results; known complexes should appear in the results. We searched for 45 known transcription factors in the yeast genome and compared the number of times they appeared in the results to how many of them are present in the dataset; transcription factors directly affect gene expression, so they should be present in a high percentage of the results. We searched for the most frequently-appearing gene pairs in the triplets to see if they exhibit known biological relationships.

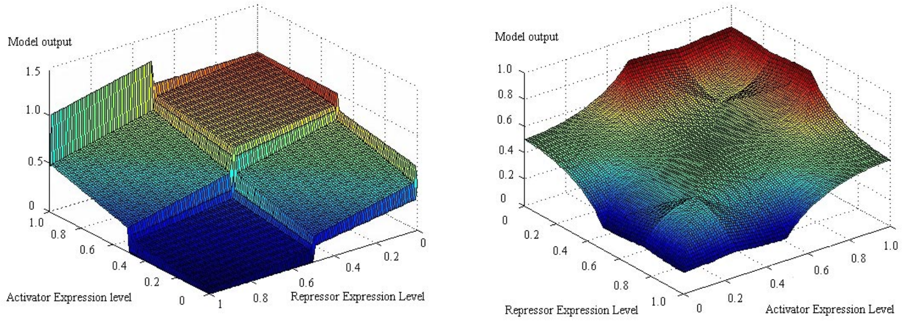

Figure 1: Output space of Woolf (left) and Mamdani models (right)

3 Simulations

The output spaces for each of the Woolf and Mamdani models are depicted in Figure 1. The other two models (SAM and Hybrid) are quite similar to the Mamdani model and are not presented. The analysis of the gradient of the output space of each model is presented in Table 1.

Model Type	Woolf	Mamdani	SAM	Hybrid
Mean Gradient	10.31	5.48	6.57	6.02
Std Dev of Gradient	13.68	4.64	0.68	2.88

Table 1: Gradient analysis of the fuzzy models

The highly irregular response of Woolf’s model is reflected in a high average gradient as well as the high standard deviation; most of the change in output is localized in small areas of the input space. The Mamdani model has a much lower average gradient and standard deviation. The Standard Additive Model has a higher average gradient, but an extremely low change in standard deviation shows that the model has a more consistent gradient. The Hybrid model appears to be a compromise between the Mamdani model and the SAM.

Error simulations for 30% noise for the Woolf and Mamdani models are shown in Figures 2-3. The results for the Hybrid and Standard Additive Models are not included as their noise response. While they perform better than Woolf response, they are not as robust as the Mamdani model. A more complete set of error simulation graphs can be found in [7].

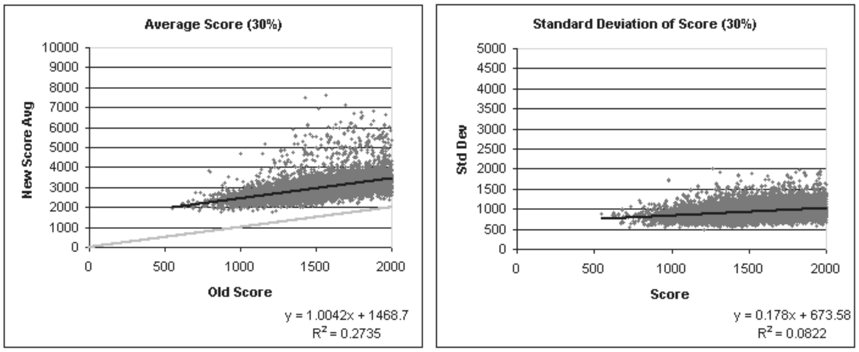

Figure 3: Monte Carlo error simulations for the Mamdani model

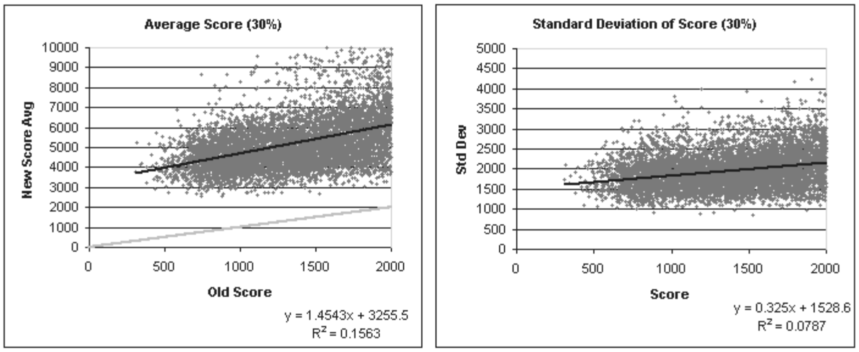

Figure 2: Monte Carlo error simulations for the Woolf model

From the graphs, it appears that Mamdani model produces regression lines with a slope closest to 1 for all potential noise distortions. This implies that, *on average*, the primary effect of noise on the model is only to add a constant error offset to the noise-free error score. The original error score for the inputs (i.e. the noise-free error score) has little or no effect upon the noise-distorted data’s error score. If the standard deviation of noise-distorted error score is also low, as is the case with the Mamdani model, we can say that the majority of gene input combinations are distorted by approximately the constant error offset. If the dataset’s noise interval can be estimated [2], one could raise the desired error cutoff by the value of the constant error offset to obtain the majority of the genes that are likely to fit the model under noise-free conditions. The other three models (Woolf, Sam, and Hybrid) have regression line slopes significantly greater than 1, implying that high-error gene combinations will be distorted by a higher amount in the presence of noise.

The standard deviation of scores around the results from Woolf’s algorithm is much higher than the other three models.

Transcription factor enrichment results are indicated in Table 2. The percentages are derived from the results of each model with an error score cutoff of 2000 (MSE of 2%) and a variance cutoff of 20000. The “Ratio of Enrichment” is the ratio of percentages of results with transcription factors in them to the percentage of tran-scription factors in the input set (3.97% of the input genes).

Model Type	Woolf	Mamdani	SAM	Hybrid
TF % in results	8.96%	10.53%	9.43%	10.51%
Ratio of Enrichment	2.26	2.66	2.38	2.65

Table 2: Transcription factor enrichment of the fuzzy models

All of the models appear to report a disproportionate amount of low-error results containing transcription factors. However, the Mamdani and Hybrid models appear to yield a higher percentage of results with transcription factors than Woolf’s model or SAM. This may imply that these models are better at extracting gene relationships.

The algorithm's output using the Mamdani model was analyzed and compared to the outputs of Woolf's model. The gene relationships of the HAP1 regulatory network were found to have similar error and variance scores as indicated in [3]. Most of the variance scores with the Mamdani model in general, so an increased variance score cutoff would eliminate the problem. This shows that the Mamdani model has the ability to find some known relationships. The most common pairs of genes were found and are summarized in Table 3. ‘A’ denotes an activator, ‘B’ denotes a repressor, and ‘C’ denotes the target. Most of the gene relationships were obtained from the Proteome YPD database [8].

A	B	C	No.	Functions
-	PUS2	LEU4	273	PUS2 alters tRNA-Leu, inhibits LEU synthesis
-	GSC2	LEU4	156	Involved in different metabolisms.
-	PUS2	ARO3	145	PUS2 alters tRNA-Tyr, inhibits ARO3 translation
-	GSC2	CAP20	138	Unknown
MEP2	-	AGP1	127	Both activated by low nitrogen levels
-	HAP1	CYT1	115	CYT1 is directly regulated by HAP1
GLK1	-	MSF1	95	Co-induced in mitochondrial mutant
SPO13	-	INO2	94	Both involved in cell division (mitosis/meiosis)
HPR5	-	GLG2	86	Co-induced during G2

Table 3: Most common gene pairs in results of Mamdani model

The common pairs show us that the Mamdani model extracts many coregulated pairs of genes. There is no known causal relation between the two, but they appear to raise and fall with similar profiles of expression. With the use of the *min* operator for fuzzy conjunction, it is more likely that changes in one model input or the other will not change the output. Thus, it is more likely that a particular gene’s expression time series will have effect on the output compared to another. Thus, we are faced with an increasing likelihood that frequently expressed pairs are in fact co-regulated and do not have a causal relationship.

4 Conclusion

We have shown that the use of the standard Mamdani model can improve the performance of Woolf's fuzzy algorithm by being more resilient to noise, which is important in light of high noise ratios in current microarray technology. While we found that the Mamdani model produces valid results with far less noise distortion, we cannot say that one model is inherently better than the other without further investigation. Some fitness measurement for different models needs to be developed to obtain a relative perspective of model validity. Some heuristics for analyzing model validity [9] may provide insight into which models are most valid.

References

[1] Anonymous: GEM microarray reproducibility Study. Incyte Pharmaceuticals, Inc (1999).

[2] M. K. Kerr, M. Martin, G. A. Churchill: Analysis of variance for gene expression microarray data. Journal of Computational Biology, 7, 819-837 (2000).

[3] P. J. Woolf, Y. Wang: A fuzzy logic approach to analyzing gene expression data. Physiological Genomics, 3, 9-15 (2000).

[4] R. J. Cho: A Genome-wide transcriptional analysis of the mitotic cell cycle. Molecular Cell, 2, 65-73 (1998).

[5] E. H. Mamdani, S. Assilian: An experiment in linguistic synthesis with a fuzzy logic controller. Int. Journal of Machine Studies, 7, No. 1, (1975).

[6] B. Kosko: Fuzzy Engineering. Prentice Hall (1997).

[7] R. Renyolds: Gene expression data analysis using fuzzy logic, masters thesis. ECE Department, University of Maine (2001).

[8] M. C. Costanzo: The Yeast Proteome database (YPD) and caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. Nucleic Acids Research, 28(1), 73-76 (2001).

[9] L. F. A. Wessels, E.P. Van Someren, M.J.T. Reinders: A comparison of genetic network models. Pacific Symposium on Biocomputing, 6, 508-519 (2001).