Multinet Bayesian Network Models for Large-scale Transcriptome

Integration in Computational Medicine

by

Tiffany J. Lin

S.B., C.S. M.I.T., 2011

Submitted to the Department of Electrical Engineering

and Computer Science

in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

at the Massachusetts Institute of Technology

May 2012

Author: _____

Department of Electrical Engineering and Computer Science
May 21, 2012

Certified by: _____

Gil Alterovitz, Research Affiliate, Thesis Supervisor
May 21, 2012

Accepted by: _____

Prof. Dennis M. Freeman, Chairman, Masters of Engineering Thesis Committee

Multinet Bayesian Network Models for Large-scale Transcriptome Integration in Computational Medicine

by

Tiffany J. Lin

Submitted to the

Department of Electrical Engineering and Computer Science

May 21, 2012

In Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

Motivation: This work utilizes the closed loop Bayesian network framework for predictive medicine via integrative analysis of publicly available gene expression findings pertaining to various diseases and analyzes the results to determine which model, single net or multinet, is a more accurate predictor for determining disease status. Results: In general, it is suggested to use the multinet Bayesian network framework for predictive medicine instead of the single net Bayesian network, because for large numbers of samples and features, it is highly likely that it is the stronger predictor, and for smaller numbers of samples and features, if the multinet returns good results, it is likely to be a better predictor than the single net Bayesian network.

Table of Contents

1 Introduction 6

1.1 Background 6

1.1.1 GEO Database 6

1.1.2 Bayesian Multinet versus Single net 7

1.1.3 Previous Work 8

1.2 Goals of Research 9

1.2.1 General Approach 9

1.2.2 Criteria for Success 10

2 Procedure 10

2.1 Materials 10

2.2 Procedure 10

3 Results 13

4 Discussion 19

5 Conclusions 20

6 Recommendations 21

7 Acknowledgements 23

8 Appendix A: List of Diseases/Disorders Used 24

9 Appendix B: Data Gathered 25

10 Appendix C: Location of Data and Code 27

11 Bibliography 28

List of Figures

Figure 1: Details of using Weka's Attribute Selected Classifier. 12

Figure 2: Details of using Weka's training set. 12

Figure 3: The difference between the multinet AUROC and the single net AUROC versus the ratio of DataSets used versus DataSets available. 15

Figure 4: Difference between AUROC versus total number of DataSets available. 15

Figure 5: The difference between the multinet AUROC and the single net AUROC versus the number of DataSets picked. There is a little more correlation here. 16

Figure 6: The difference between the multinet AUROC and the single net AUROC versus the number of DataSets after the pipeline. 17

Figure 7: The difference between the multinet AUROC and the single net AUROC versus the multinet AUROC. 17

Figure 8: Difference between the multinet AUROC and the single net AUROC versus the number of samples in the experiments. 18

Figure 9: Difference between the multinet AUROC and the single net AUROC versus the number of features (genes) in the experiments. 18

List of Tables

Table 1: Summary of statistical data run on the difference between the multinet Bayesian network AUROC and the single net Bayesian network AUROC. 14

Table 2: T-Test Results for experiments with more than 105 samples. 19

Table 3: T-Test Results for experiments with more than 1100 features. 19

Table 4: A summary of data gathered for each disease that, after the pipeline, fit the profile needed for this project. 26

1 Introduction

There has been research on the usage of multinet Bayesian networks versus single net Bayesian networks, especially in integrative gene expression analysis and predictive medicine. In particular, Parikh, et. al.'s paper on automated Bayesian frameworks for interactive gene expression analysis [1] claims that the accuracy of the model is higher with the multinet approach, which is where integration is performed in the assignment step, rather than the single net approach, which is the data collection step. However, the analysis to this part of the paper is sparse, with only a few trials.

Therefore, we wish to delve deeper into the problem, and try to determine if the multinet is always the better approach to predictive medicine, and, if not, it is possible to determine whether one model, multinet or single net, would be better than the other, or, if neither is better than the other as a whole, if there is a way to determine when one model would be better than the other.

1.1 Background

1.1.1 GEO Database

The Gene Expressions Omnibus, from now on referred to as GEO, is a publicly accessible repository of genomic data [2]. The GEO offers a flexible platform for the submission and retrieval of heterogeneous data from high throughput gene expression and genomic hybridization experiments, and categorizes all user submitted experiments into samples, series, and platforms, some of which are then manually transformed into DataSet records.

In this study, all the gene expression data used was found in GEO database's DataSet records. The identification numbers of experiments relating to certain diseases can be found via the GEO database web site.

1.1.2 Bayesian Multinet versus Single net

Bayesian networks, which are represented by directed, acyclic graphs, are commonly used in statistical analysis, and more importantly for this research, predictive medicine. The Bayesian approach starts with an established probability distribution, called a prior distribution, and uses previously gathered sample data, in this case the genomic information stored in the GEO DataSets, to update the prior distribution to a posterior distribution. Each node on the graph represents one of the attributes that might affect the subject of research and a joint probability table.

The Bayesian network is very popular because graphical models can help break down large complex systems into simpler parts. Because there is a joint probability table attached to each node, the Bayesian network is particularly useful for predictive medicine. This is because it is often important in predictive medicine to know at what degree of certainty one can make judgment calls. For example, having a 70% chance of the symptoms and genomic data stating that the patient has disease A is very different from having a 98% chance that the patient has disease B. Since the Bayesian network allows one to know exactly at what percent chance the disease is what one believes it is, one can make decisions on further actions easily.

In previous research, it was common to use the Bayesian multinet to analyze disease. This is a set of distinct, yet related, Bayesian networks, where the dataset is first

partitioned by class and a single Bayesian network is constructed for each partition to maximize each individual posterior probability. The aim of this was to model the underlying pattern of dependency between different features. The single net, on the other hand, does not model the dependency, and the classifier is forced to be static across all classes.

1.1.3 Previous Work

Neena Parikh's research created an automated framework that allows for the genome wide expression data, using the GEO database, in regards to diseases and disorders, while also creating a predictive model for disease-related phenotypes that will illustrate the relationship between genetic factors and pathways. The automated framework, written in R, a free software environment for statistical computing and graphics [3], that she made also created files that could be used to create both Bayesian single nets and Bayesian multinets using a Weka [4], a data mining software.

The program took an input of a series of DataSets from the GEO database, and examined each one to find relevant DataSets. For example, it searched for key words such as "control", "infected", "normal type", and "wild type". The pipeline would then filter out DataSets that were not relevant, and return an arff, attribute relation file format, file, an ASCII text file that describes a list of instances that share a set of attributes, which then can be read by Weka to create models.

To test the pipeline, Parikh studied Huntington's disease, obesity, leukemia, and lymphoma, and from the results from Weka, she ran an external cross-validation on the data, which can be used to evaluate any kind of predictive model one can construct. To

judge how accurate the model was, Parikh examined the area under the receiver operating curve (AUROC), which is a graph that illustrates the performance of the model as its discrimination threshold is varied. This value can be from 0 to 1, with 1 being an indication of the best model. Using that parameter, she declared that the Bayesian multinet was a better model for predictive medicine than the Bayesian single net model, as the multinet had a higher AUROC value, which was judged to be either fair (0.7-0.8) or good (0.8-0.9), while the single net seemed to produce lower AUROC values, which were judged to be poor (less than 0.7).

1.2 Goals of Research

As Parikh was only using four diseases to make her claim, the claim, although seeming true, is weak and unsupported. Therefore, the goals of this project are to either confirm or deny Parikh's claim that the multinet is a more accurate model by using the same methods she used. If the multinet method does not always create a better model, then we wanted to determine if, in general, one model would be better than the other, or if there was a way to determine when one should use one model over the other.

1.2.1 General Approach

The general approach we took to the problem was to manually sort through the DataSets available for various diseases, use the pipeline that Parikh created, and determine the AUROC values. While Parikh compared the AUROC values to a strict slide scale, we decided to, instead, compare the AUROC values that resulted from the multinet and the single net Bayesian networks directly by looking at the difference between the two values. We wished to do this because the goal of this research is to see

which method is better on average for different experiments, which requires a direct comparison between the two methods.

We then wanted to look at how the difference in the values could be affected by the factors we controlled, namely, how many DataSets went into the pipeline. We also wished to examine the same difference in AUROC values as it related to the number of DataSets that survived the pipeline.

1.2.2 Criteria for Success

Success would be determined by how well we fulfilled the three stated goals of this research: if we determine that either the Bayesian single net or the Bayesian multinet network should always be used, if we show that either model is the better model as decided via the AUROC found, and if we show that there is a criteria for when using one model will be better than using the other.

2 Procedure

2.1 Materials

The materials used in this project include the automated pipeline created by Parikh et. al, the R project, Weka 3 Data Mining, and the GEO database. I also had a list of diseases as a reference.

2.2 Procedure

To gather the data, we looked at each individual disease and searched for relevant DataSets on the GEO database. We then screened the results so we had the identification numbers of a series of DataSets that we know are relevant. This step is necessary

because of two things: 1) sometimes the system returns odd results, such as an experiment on Down syndrome mixed in with the results of lung diseases and 2) we want to confuse the system as little as possible, so we filter out experiments that are run on two or more diseases and focus only on the ones dealing only with the disease in question, for example Alzheimer's syndrome is commonly analyzed alongside schizophrenia.

We then input the identification numbers of the experiments (called GDSIDS for GEO database identifications) into the pipeline and allow it to create the arff file for both Bayesian single net and Bayesian multinet models.

At this step it is necessary to further narrow down the number of diseases used. Because the pipeline does its own filtering to find relevant experiments, we may end up with only one DataSet left after the pipeline, or even none. These diseases must be discarded because with both, the model for both the multinet and the single net Bayesian network is identical, and therefore, the external cross validation is also identical, leaving us with no relevant information, and instead data that can heavily skew our decision on whether to use multinets or single net Bayesian networks greatly.

If the disease survives the filtering step above, we then run the Weka program on it, using the Weka explorer to run the external cross validation. To do that, we used the classify tab in the Weka explorer, and picked the Attribute Selected Classifier under classifiers of the type meta, then for the details of this classifier, we used the Naïve Bayes classifier, with an evaluator of wrapper subset evaluation, and a search of linear forward selection, which is an extension of the best first search (shown in Figure 1).

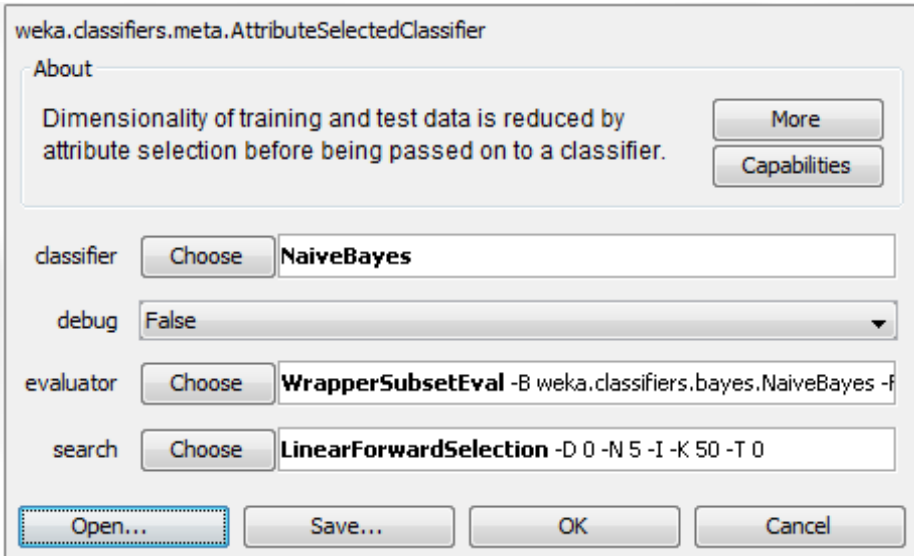

Figure 1: Details of using Weka's Attribute Selected Classifier.

We then ran the classifier on cross-validation with three folds, with the class (Control or Infected in the case of single net Bayesian network and Control and a list of experiments for the multinet Bayesian network) as the subject of interest in the Bayesian network (as shown in Figure 2). Running the external cross-validation with three folds means that we create the model with two thirds of the data, and testing with the last third of the data available.

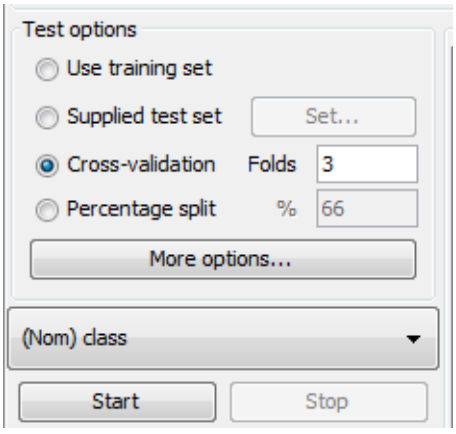

Figure 2: Details of using Weka's training set.

We then recorded the AUROC data for each of the models for comparison later, the number of DataSets that survived the pipeline's filter, the number of attributes (genes) associated with the disease, and the number of samples we have to build and test the model with for future use.

To find correlation, we used scatter plots to find the relationship between the difference in AUROC values and the other recorded data: total number of experiments, number of experiments before the pipeline, number of experiments after the pipeline, number of features, and number of samples available. Once we had some likely relationships from the plots, we ran t-tests to determine if these relationships we see are statistically significant.

More details about where to find the code and the GDSIDs can be found in Appendix C below.

3 Results

The detailed results are listed in Appendix B. A summary of statistics of the data gathered is below. Because Parikh claimed that the multinet Bayesian network was stronger than the single net, we chose to use the multinet AUROC minus the single net AUROC so a positive value would make Parikh's claim valid.

Multinet AUROC – single net AUROC	
Mean	-0.01029
Standard Error	0.020305
Median	-0.0075
Mode	-0.022
Standard Deviation	0.12517
Sample Variance	0.015668

Kurtosis	2.172319
Skewness	0.32475
Range	0.69
Minimum	-0.302
Maximum	0.388
Sum	-0.391
Count	38
Largest(1)	0.388
Smallest(1)	-0.302
Confidence Level (95.0%)	0.041142

Table 1: Summary of statistical data run on the difference between the multinet Bayesian network AUROC and the single net Bayesian network AUROC.

Below is a sample of a few of the diseases, with the models judged with the same measure that Parikh et. al. used.

Inflammation		
Single net AUROC	.914	Excellent
Multinet AUROC	.932	Excellent
Ischema		
Single net AUROC	.695	Poor
Multinet AUROC	.88	Good
Leiomyoma		
Single net AUROC	.88	Good
Multinet AUROC	.951	Excellent
Aortic Aneurysm		
Single net AUROC	.949	Excellent
Multinet AUROC	.878	Good

Table 2: Examples of diseases and the AUROC values associated with the two models, along with the ranking.

One of the first things we decided to analyze was if the ratio of the number of DataSets we chose to put into the pipeline to the number of DataSets available would have any effect on the difference between the AUROC data.

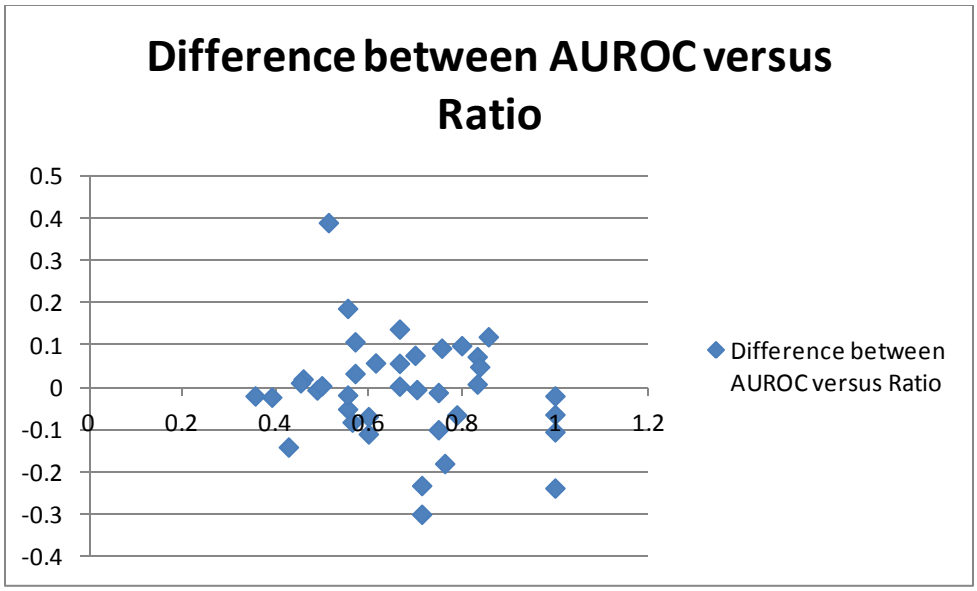

Figure 3: The difference between the multinet AUROC and the single net AUROC versus the ratio of DataSets used versus DataSets available.

We also decided to see if there was correlation between the differences and the number of total DataSets available in the GEO database, which is related to how popular the disease was for genomic research. With more interest in the disease, there would be more data, and perhaps the model will be more accurate.

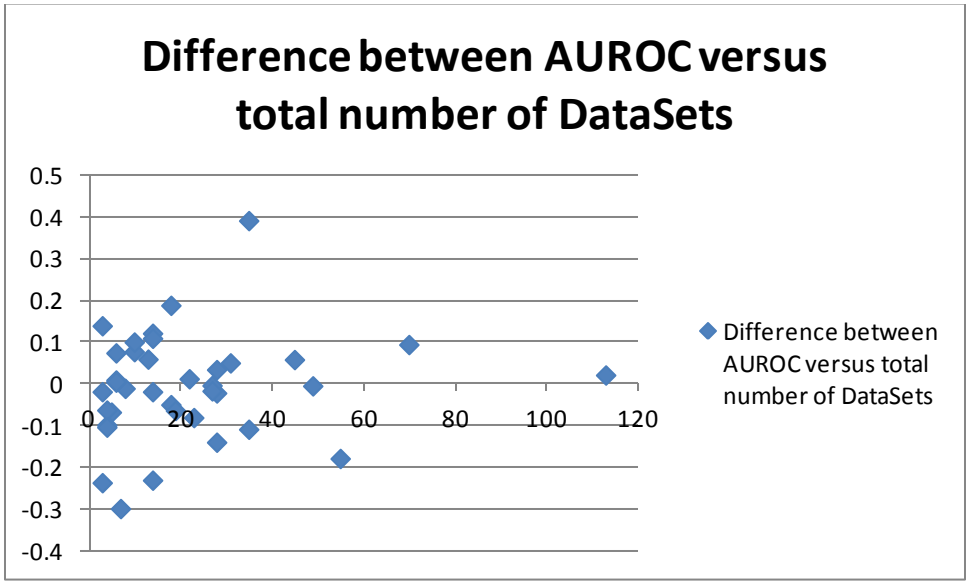

Figure 4: Difference between AUROC versus total number of DataSets available.

There is some correlation, so we decided to examine

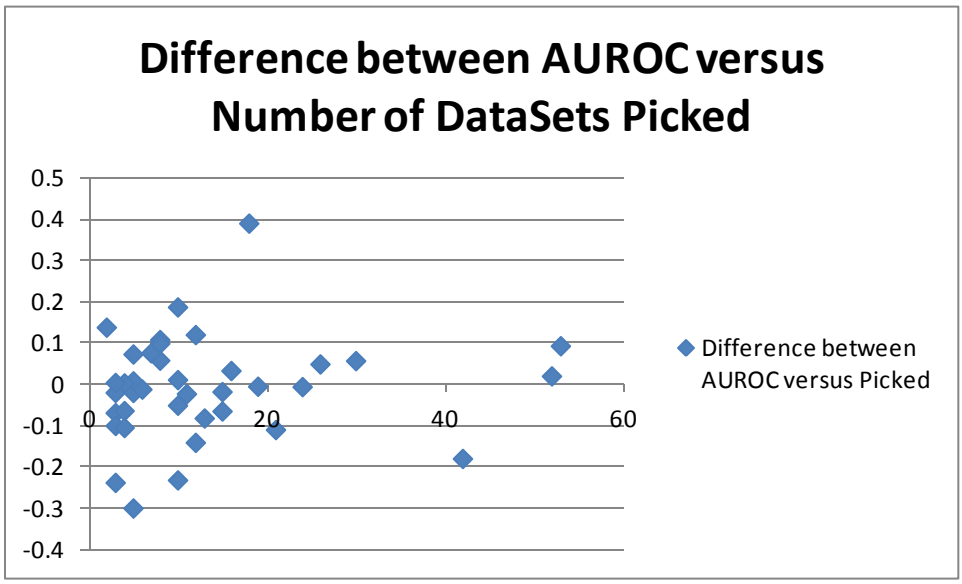

Figure 5: The difference between the multinet AUROC and the single net AUROC versus the number of DataSets picked. There is a little more correlation here.

Seeing a slightly stronger correlation, although not one that we could conclusively say showed that either model was the more correct one, we decided to move forward in this direction, and analyzed the correlation between the number of DataSets that managed to pass the filter in the pipeline. This was the next logical step since the pipeline does filter out the number of DataSets that we use to create the models.

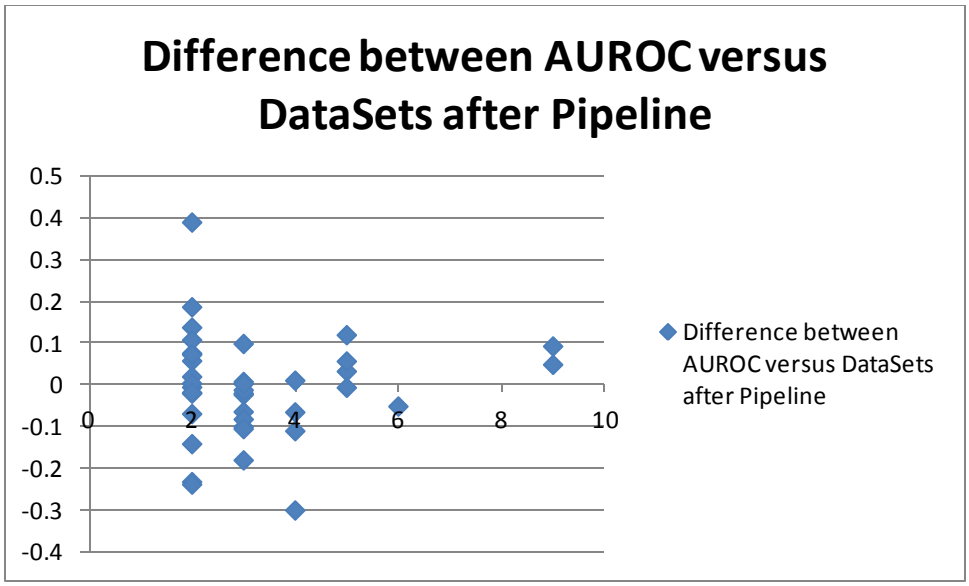

Figure 6: The difference between the multinet AUROC and the single net AUROC versus the number of DataSets after the pipeline.

We then examined the difference between the multinet AUROC and single net AUROC versus the multinet AUROC, because we wanted to see if there was any relation between the strength of the model and how much better the model was.

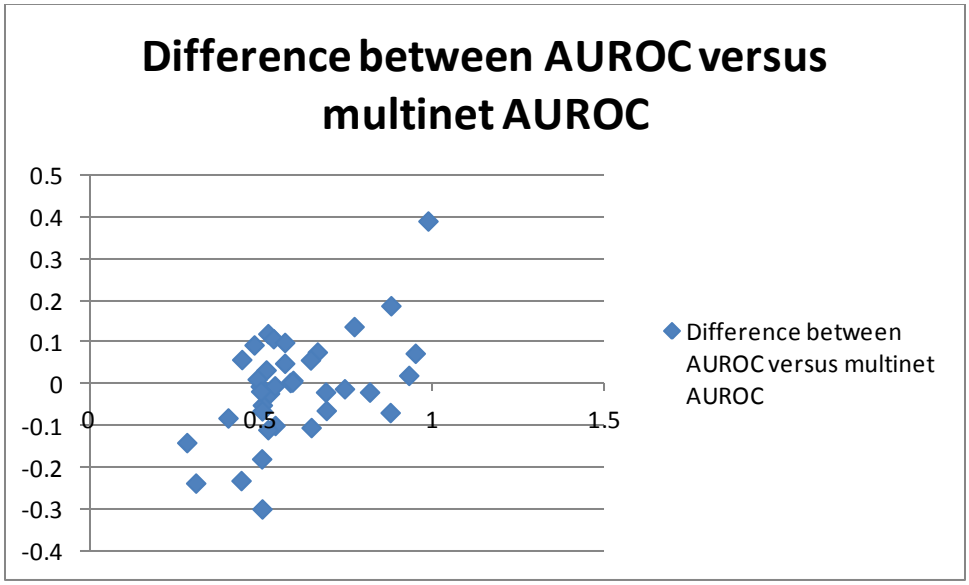

Figure 7: The difference between the multinet AUROC and the single net AUROC versus the multinet AUROC.

As you can see, there seems to be some correlation. We also decided to examine the relationship between the number of samples available and the difference between the AUROC values, as well as the relationship between the number of features and the difference between the AUROC values.

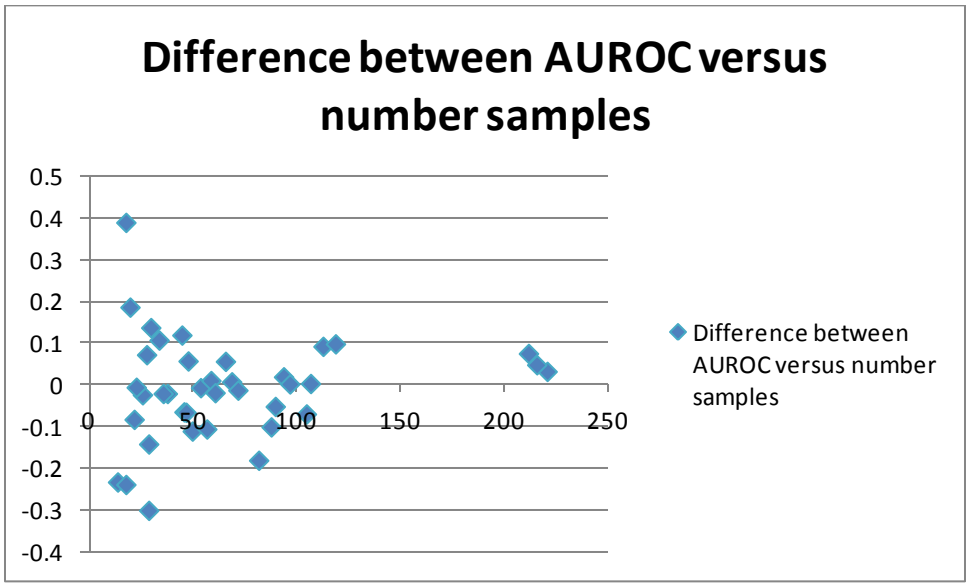

Figure 8: Difference between the multinet AUROC and the single net AUROC versus the number of samples in the experiments.

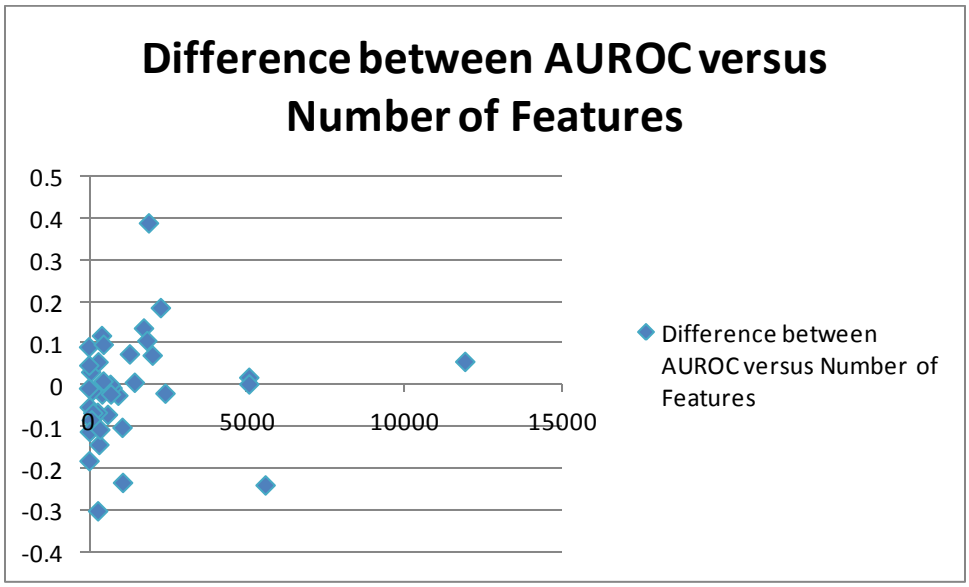

Figure 9: Difference between the multinet AUROC and the single net AUROC versus the number of features (genes) in the experiments.

From these graphs, we decided to examine the t-test results for number of samples and number of features.

T-Test Result for # samples > 105	Average of Single Net AUROC of # samples > 105	Average of Multinet AUROC of # samples > 105
0.006489673	0.509666667	0.566666667

Table 3: T-Test Results for experiments with more than 105 samples.

T-Test Result for # features > 1100	Average of Single Net AUROC of # features > 1100	Average of Multinet AUROC of # features > 1100
0.075590775	0.616083333	0.68125

Table 4: T-Test Results for experiments with more than 1100 features.

4 Discussion

From the results, we can see that the Bayesian multinet is not always the stronger predictive model, which is what Parekh claims. From Table 1, we see that the multinet Bayesian network is, on average, the weaker of the two models. Therefore, we know that the multinet is not always the more powerful predictor. However, as Table 2 shows, the multinet model is a better model sometimes. Therefore, we now need to see when the multinet model should be used, and when the single net should be used.

We first examined the difference between AUROC values and the number of DataSets, comparing the values to the total number of DataSets, the number of DataSets we picked, and the number of DataSets that made it through the pipeline. From the graphs, we can see that there is some correlation between the three, but then, running t-tests, we find that the correlation is not statistically significant when arranged by the total number of DataSets, the number of picked DataSets, or the number of DataSets that

survived the pipeline. Therefore, there must be something else that is related to the number of DataSets that can help determine how powerful the model is.

We then examined the number of samples and the number of features. From Figure 8 and Figure 9, we can see that there is some correlation: larger features and samples results in a better multinet result versus a single net result. We then ran t-tests for these values, and found that when the number of samples available for the disease is greater than 105, the probability that the multinet Bayesian network model's AUROC value is larger than that of the single net Bayesian network model is statistically significant. We also found the same for when the number of features is greater than 1100. Therefore, we can conclude that when one has more than 105 for the disease or more than 1100 features, the multinet model is a stronger predictor than the single net predictor.

Another interesting thing to note is that from Figure 7, we can see that there is definitely a correlation between the multinet AUROC and the difference between the AUROC values. It seems that the better the multinet model is, the more likely it is for it to be a much better predictor than the single net model. Therefore, if one runs the multinet model and finds a high AUROC value, it is more likely that it is a better model than the single net Bayesian network model.

5 Conclusions

Because Parekh et. al. used diseases with a large number of DataSets available, their results were skewed. They claimed that one should always use the multinet Bayesian network method for creating models in predictive medicine, but one can see from this project that that is not necessarily the case. The multinet Bayesian network is not always a more powerful predictor than the single net Bayesian network, as at least

half of the experiments run for this project showed that the single net actually is the more accurate model.

The project does find that at a certain point, when the number of samples is greater than 105 or when the number of features is greater than 1100, the multinet is a significantly better predictor ($P < m$) where m is the p-value. For experiments with fewer samples or features than the suggested values, it is unknown which model would be better. However, because there is a positive correlation between the difference of the AUROC values of the two models and the AUROC value of the multinet model, we know that if the multinet model is a successful predictor, it is very likely that it is better than the single net model.

The suggestion is, therefore, this: for all experiments, it is better to use the multinet model than the single net Bayesian network. For experiments with a large number of samples ore features (genes) available, we know that the multinet model is the stronger model, and at low numbers of experiments available, we know that if the AUROC value of the multinet is higher, then the single net model will likely not be a better model. Furthermore, the multinet Bayesian network model would always provide more detail in illustrating the relationships between the genes, diseases, and experiments, but the single net may provide a more powerful predictor.

Thus, we know that although Parekh, et. al.'s suggestion was made on incomplete data, it is still a valid suggestion.

6 Recommendations

For future research, one may wish to continue in this manner and catalogue the differences in the area under the AUROC curve for more disease, or perhaps branch out

to machine learning in something other than predictive medicine to see if one can generalize the claim that both networks are approximately the same in predictive power. It would be interesting to see if our claims hold only for predictive medicine or if they are universal claims. In this study, it would also be interesting to delve into the reasons why this claim is true, even though the multinet Bayesian network, with its ability to model the underlying pattern of dependency between different features, would appear at first glance to be the better model.

Another step one might take would be to determine if there is something inherent in the DataSets used that could be used to predict which model would be better, as this experiment focused mainly in raw data from an input standpoint.

Furthermore, because the database is constantly updated with more experiments, one can always redo the experiment to confirm or disprove the results we found. These would be good steps to take because, from a list of many diseases, only some had the required minimum of two experiments that survived the pipeline at this time. The results of this project would become more accurate as the GEO database is updated through time (and more experiments).

Another step one might take is to find a way to create another method of modeling disease related phenotypes to find the relationship between genetics and diseases that would be more effective than either the multinet or the single net method currently is. One would then need to compare their proposed model in a similar way to the one we have used to validate the model.

7 Acknowledgements

I would like to acknowledge and thank Gil Alterovitz and Amin Zollanvari for the guidance and helped they provided me. They introduced me to the project and helped me get started, as well as provided me additional people to contact for help. I also want to thank Radhika Malik and Swetha Sampath for their help whenever I got stuck, answering the questions I had about the script that I was using. I also would like to thank Peter Szolovits for introducing me to biomedical computation via the MIT class 6.872.

8 Appendix A: List of Diseases/Disorders Used

Acne Vulgaris, Alzheimer disease, Anemia, Aortic Aneurysm, Rheumatoid Arthritis, Asthma, Atherosclerosis, Bipolar Disorder, Squamous Cell Carcinoma, Cardiomyopathy, Colitis, Dermatitis, Diabetes, Endometriosis, Epilepsy, Fibrosis, Glaucoma, Heart Disease, HIV, Huntington Disease, Hypertrophy, Hypoxia, Inflammation, Ischemia, Leiomyoma, Lung Disease, Lung Neoplasm, Lymphoma, Melanoma, Mental Disorder, Neoplasm Metastasis, Neurodegenerative Disease, Parkinson Disease, Polycystic Ovary Syndrome, Pre-Eclampsia, Progeria, Psoriasis, Pulmonary disease.

9 Appendix B: Data Gathered

# of Data Sets picked	# of Data Sets after pipeline	Total # of Data Sets	single net ROC Area	multinet ROC Area	delta ROC Area	ratio of picked/ total	# of samples	# of features
2	2	3	0.637	0.773	0.136	0.666667	30	1741
8	2	13	0.39	0.446	0.056	0.615385	48	11911
7	2	10	0.592	0.666	0.074	0.7	212	1296
3	2	5	0.949	0.878	-0.071	0.6	105	601
10	2	14	0.678	0.444	-0.234	0.714286	14	1078
16	5	28	0.486	0.517	0.031	0.571429	221	69
11	3	28	0.552	0.527	-0.025	0.392857	26	929
3	3	4	0.645	0.543	-0.102	0.75	88	1062
10	6	18	0.559	0.506	-0.053	0.555556	90	22
12	2	28	0.429	0.286	-0.143	0.428571	29	328
5	3	14	0.712	0.69	-0.022	0.357143	38	423
4	3	4	0.758	0.692	-0.066	1	46	282
53	9	70	0.391	0.482	0.091	0.757143	113	3
5	3	6	0.589	0.595	0.006	0.833333	69	1451
6	3	8	0.759	0.745	-0.014	0.75	72	117
24	5	49	0.508	0.5	-0.008	0.489796	54	2
5	4	7	0.807	0.505	-0.302	0.714286	29	289
12	5	14	0.404	0.522	0.118	0.857143	45	412
15	4	19	0.572	0.505	-0.067	0.789474	47	136
8	2	14	0.432	0.538	0.106	0.571429	34	1846
30	5	45	0.591	0.646	0.055	0.666667	66	302

42	3	55	0.686	0.504	-0.182	0.763636	82	17
52	2	113	0.914	0.932	0.018	0.460177	94	5073
10	2	18	0.695	0.88	0.185	0.555556	20	2276
5	2	6	0.88	0.951	0.071	0.833333	28	2017
19	2	27	0.549	0.542	-0.007	0.703704	23	745
26	9	31	0.524	0.571	0.047	0.83871	216	9
18	2	35	0.6	0.988	0.388	0.514286	18	1898
13	3	23	0.49	0.406	-0.084	0.565217	22	87
4	3	6	0.586	0.587	0.001	0.666667	97	684
21	4	35	0.634	0.522	-0.112	0.6	50	28
10	4	22	0.482	0.491	0.009	0.454545	59	462
8	3	10	0.474	0.571	0.097	0.8	119	467
4	3	4	0.755	0.648	-0.107	1	57	358
3	2	3	0.552	0.312	-0.24	1	18	5591
3	2	3	0.84	0.818	-0.022	1	36	699
3	2	6	0.591	0.593	0.002	0.5	107	5075
15	2	27	0.521	0.501	-0.02	0.555556	61	2421

Table 5: A summary of data gathered for each disease that, after the pipeline, fit the profile needed for this project.

10 Appendix C: Location of Data and Code

To access the code base for this project, please go to the following web site: https://github.com/bcl2group/GroupData/tree/master/Tiffany.

Download the file named GEOpipeline.R, and run it in the R program [3]. The disease list I used was the file called All_diseasesToGDSIDs.txt. To find the GDSIDs of the experiments associated with the diseases, I went to the web site http://www.ncbi.nlm.nih.gov/gds, and searched for the disease, and looked at the DataSets available for the disease. The GDSIDs are then inputted into the R file, for the variable called ids as a string with a comma between each id. Then run the script. You will find files called Myelitis_singlenet_25percent.txt and Myelitis_25percent.txt. You will then need to rename the two, changing the .txt extension to .arff. A change from Myelitis to the disease name is also suggested. From here, simply open the .arff file, and the Weka explorer should open.

11 Bibliography

[1] Neena Parikh, Amin Zollanvari and Gil Alterovitz. An Automated Bayesian Framework for Integrative Gene Expression Analysis and Predictive Medicine.

[2] Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A. GEO: archive for functional genomics data sets—10 years on Nucleic Acids Res. 2011 Jan;39(Database issue):D1005-10

[3] R Development Core Team (2005). R: A language and environment for statistical computing, reference index version 2.2.1. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

[4] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[5] Enabling Integrative Genomic Analysis Of High-Impact Human Diseases Through Text Mining. Joel Dudley and Atul J. Butte.

[6] Davis, Sean and Meltzer, Paul S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics (2007) 23 (14):1846-1847.

[7] Smyth, G. K. (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, pages 397-420.