REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
20-05-2015	Final Report	29-Sep-2010 - 28-Sep-2013

4. TITLE AND SUBTITLE
Disease Modeling via Large-Scale Network Analysis

5a. CONTRACT NUMBER
W911NF-10-1-0529

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Inderjit S. Dhillon, Edward Marcotte

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Texas at Austin
101 East 27th Street
Suite 5.300
Austin, TX 78712 -1532

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
58343-MA.9

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
A central goal of genetics is to learn how the genotype of an organism determines its phenotype. We address the implicit problem of predicting the association of genes with phenotypes or traits. Our primary goal is to develop pragmatic data analytic methods for linking specific genes to traits and diseases, especially polygenic traits, which are the most challenging. We are also interested in developing theoretical guarantees for the methods. In the past, we have developed predictive methods general enough to apply to potentially any genetic trait, varying from plant traits relevant to desirable agricultural properties to important human diseases. Our methods Katz on

15. SUBJECT TERMS
Boinformatics, Genes, Diseases, Social Network Analysis, Noisy Labels, Inductive Matrix Completion

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Inderjit Dhillon
UU	UU	UU	UU		19b. TELEPHONE NUMBER 512-471-9725

Report Title

Disease Modeling via Large-Scale Network Analysis

ABSTRACT

A central goal of genetics is to learn how the genotype of an organism determines its phenotype. We address the implicit problem of predicting the association of genes with phenotypes or traits. Our primary goal is to develop pragmatic data analytic methods for linking specific genes to traits and diseases, especially polygenic traits, which are the most challenging. We are also interested in developing theoretical guarantees for the methods. In the past, we have developed predictive methods general enough to apply to potentially any genetic trait, varying from plant traits relevant to desirable agricultural properties to important human diseases. Our methods, Katz on heterogeneous network and CATAPULT[1], for predicting gene-disease associations were published during the last project period in the PLOS One journal. The biological problem has also led us to pursue a significant problem in machine learning. One of the fundamental questions in machine learning relating to the classification problem is if we can efficiently learn classifiers that can provably achieve low misclassification rates in the presence of certain type of random label noise in the training data. We have answered the question in affirmative in our recent paper[2], and in particular, developed important theoretical results and robust algorithms for dealing with random label noise in classification. Our CATAPULT system employs "biased" support vector machines (SVM), to cope with the lack of negative examples. Biased SVMs have been empirically shown to be successful for similar tasks, but there has not been any theoretical study previously. In the last project period, we were able to show that they exhibit a certain noise tolerance property, a novel result in machine learning. We have also developed a novel matrix-completion method called Inductive Matrix Completion to the problem of predicting gene-disease associations [3] that combines multiple types of features for diseases and genes for discovering new gene–disease associations.

While the primary objective is the biological problem, the mathematical models and techniques developed for the problem are expected to shed light on some open problems related to PU(Positive-Unlabeled) learning, such as (a) multiple sources in PU learning, (b) multi-task PU learning, and (c) theoretical guarantees for PU learning algorithms.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received		Paper
10/01/2013	6.00	U. Martin Singh-Blom, Nagarajan Natarajan, Ambuj Tewari, John O. Woods, Inderjit S. Dhillon, Edward M. Marcotte. Prediction and Validation of Gene-Disease Associations Using Methods Inspired by Social Network Analyses, PLoS ONE, (05 2013): 0. doi: 10.1371/journal.pone.0058977
10/11/2012	3.00	I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, E. M. Marcotte. Prioritizing candidate disease genes by network-based boosting of genome-wide association data, Genome Research, (05 2011): 0. doi: 10.1101/gr.118992.110
TOTAL:	**2**	

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received	Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 1.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received		Paper
05/20/2015	8.00	Nagarajan Natarajan, Inderjit S. Dhillon. Inductive matrix completion for predicting gene–disease associations, . , . : ,

TOTAL: **1**

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received		Paper
10/02/2013	7.00	Nagarajan Natarajan, Inderjit Dhillon, Pradeep Ravikumar, Ambuj Tewari. Learning with Noisy Labels, Advances in Neural Information Processing Systems (NIPS). 05-DEC-13, . : ,

TOTAL: **1**

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received		Paper
10/11/2012	4.00	Edward M. Marcotte, Inderjit S. Dhillon, U. Martin Singh-Blom, Nagarajan Natarajan, Ambuj Tewari, John O. Woods. Prediction and validation of gene-disease associations using methods inspired by social network analyses, PLoS ONE (09 2012)
10/11/2012	5.00	Nagarajan Natarajan, Ambuj Tewari, Inderjit S. Dhillon. Convex Loss Minimization with Noisy Labels, International conference on machine learning (10 2012)
TOTAL:		**2**

Number of Manuscripts:

Books

Received	Book
TOTAL:	

Received	Book Chapter
TOTAL:	

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Nagarajan Natarajan	0.25	
Donghyuk Shin	0.19	
FTE Equivalent:	**0.44**	
Total Number:	**2**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
Piyush Rai	0.25
FTE Equivalent:	**0.25**
Total Number:	**1**

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Inderjit Dhillon	0.11	No
FTE Equivalent:	**0.11**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):..... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: 0.00

Names of Personnel receiving masters degrees

NAME

Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

Abstract

A central goal of genetics is to learn how the genotype of an organism determines its phenotype. We address the implicit problem of predicting the association of genes with phenotypes or traits. Our primary goal is to develop pragmatic data analytic methods for linking specific genes to traits and diseases, especially polygenic traits, which are the most challenging. We are also interested in developing theoretical guarantees for the methods. In the past, we have developed predictive methods general enough to apply to potentially any genetic trait, varying from plant traits relevant to desirable agricultural properties to important human diseases. Our methods, Katz on heterogeneous network and CATAPULT[1], for predicting gene-disease associations were published during the last project period in the PLOS One journal. The biological problem has also led us to pursue a significant problem in machine learning. One of the fundamental questions in machine learning relating to the classification problem is if we can efficiently learn classifiers that can provably achieve low misclassification rates in the presence of certain type of random label noise in the training data. We have answered the question in affirmative in our recent paper[2], and in particular, developed important theoretical results and robust algorithms for dealing with random label noise in classification. Our CATAPULT system employs "biased" support vector machines (SVM), to cope with the lack of negative examples. Biased SVMs have been empirically shown to be successful for similar tasks, but there has not been any theoretical study previously. In the last project period, we were able to show that they exhibit a certain noise tolerance property, a novel result in machine learning. We have also developed a novel matrix-completion method called Inductive Matrix Completion to the problem of predicting gene-disease associations [3] that combines multiple types of features for diseases and genes for discovering new gene–disease associations.

While the primary objective is the biological problem, the mathematical models and techniques developed for the problem are expected to shed light on some open problems related to PU(Positive-Unlabeled) learning, such as (a) multiple sources in PU learning, (b) multi-task PU learning, and (c) theoretical guarantees for PU learning algorithms.

Approaches

We pose the problem of inferring associations between genes and phenotypes, as (a) identifying missing links or formation of new links(as in a social network), (b) classifying links as positive or negative in a Positive-Unlabeled learning framework. We have developed unsupervised and supervised learning models that combine information from multiple species. As an analogy with social networks, we model the interactions among genes as a social network, and the associations between genes and phenotypes as a bipartite affiliation networks. Alternate sources of information, like the gene-phenotype networks from other species are similarly modeled as bipartite affiliation networks. A heterogeneous network consisting of the gene interactions network, gene-phenotype networks (of multiple species) and phenotype-phenotype similarity network is formed. The task is then to predict links (or discover missing links) in the gene-disease network of humans. We have developed unsupervised and supervised learning models for linking phenotypes to genes that effectively combine the information from multiple networks. From the perspective of classification, we consider the problem of learning a function that classifies a gene-phenotype pair as positive or negative. The traditional supervised learning problem requires both positive and negative examples to learn meaningful functions. However, we don't have negative examples for the problem, which makes it unique in this respect and calls for a special-type of machine learning model, which is referred to as PU (positive-unlabeled) learning. In [1], we propose an unsupervised method inspired from social network analysis, the Katz method and a supervised learning method, CATAPULT based on PU learning for predicting gene-disease associations. We perform experiments on the human diseases from Online Mendelian Inheritance in Man(OMIM) database as well as benchmark drug-target interactions data (enzymes, ion channels, G-protein-coupled receptors, and nuclear receptors). Our methods incorporate information from eight other model organisms, namely, plant (Arabidopsis thaliana), worm (Caenorhabditis elegans), fruit fly (Drosophila melanogaster), mouse (Mus musculus), yeast (Saccharomyces cerevisiae), Escherichia coli, zebrafish (Danio rerio), and chicken (Gallus gallus). We use two kinds of human gene interactions: (a) Human Net, a large-scale functional gene network which incorporates 21 different data sets, which are results of over 50 million individual experimental observations, and (b) Human Protein Reference Database, commonly used in the published literature for studying gene-disease associations.

We extend the Katz measure, which has been shown to be successful for link prediction, to the heterogeneous network setting comprising the human gene network, gene-phenotype bipartite networks of humans and model organisms, and the phenotype-phenotype networks. Katz measure is a walk-based measure that summarizes similarities between two nodes in a network based on the number of paths of a fixed length between the nodes. Our supervised learning method, CATAPULT, learns the weights for different types of paths, to improve on the performance of fixed choice of parameters in Katz. We construct a feature mapping that maps a gene-phenotype pair into a feature space of manageable dimensions that represent walks of different types in the heterogeneous network. Absence of negative examples calls for Positive-Unlabeled learning techniques. We use a biased Support Vector Machine that is based on the idea that the false negatives must be penalized heavily as they are known to be positives, while the false negatives much less as they are arbitrarily chosen from the large set of unlabeled examples.

In the last project period, we theoretically studied the problem of learning with label noise, motivated from the positive-unlabeled learning problem arising in settings such as the gene-disease association prediction. In particular, we consider the problem of learning a classifier when the training data has class-conditional random label noise, i.e. positive and negative examples can have labels flipped with respective noise rates. Existing algorithms that work by minimizing convex loss functions can fail at high noise rates. We propose two approaches to suitably modify any given surrogate loss function. First, we provide a simple

unbiased estimator of any loss, and obtain performance bounds for empirical risk minimization in the presence of iid data with noisy labels. We are able to show that if the loss function satisfies a simple symmetry condition, then the method leads to an efficient algorithm for empirical minimization. Second, by leveraging a reduction of risk minimization under noisy labels to classification with weighted 0-1 loss, we suggest the use of a simple weighted surrogate loss, for which we are able to obtain strong empirical risk bounds. This approach has a remarkable consequence --- methods used in practice such as biased SVM and weighted logistic regression are provably noise-tolerant. Our work on classification in the presence of random label noise, in this project period, appeared at the NIPS conference[2].

More recently, we developed a novel matrix-completion method called Inductive Matrix Completion to the problem of predicting gene-disease associations [3]; it combines multiple types of evidence (features) for diseases and genes to learn latent factors that explain the observed gene–disease associations. We construct features from different biological sources such as microarray expression data and disease-related textual data. A crucial advantage of the method is that it is inductive; it can be applied to diseases not seen at training time, unlike traditional matrix-completion approaches and network-based inference methods that are transductive.

Significance
Besides discovering new gene-phenotype associations, any approach improving our ability to link genes to traits will have immediate impact: (1) on understanding molecular mechanisms of the traits, (2) on identifying genetic targets relevant to manipulating the traits, and (3) on targeting and identifying genetic variants associated with the trait, thus enabling genetic diagnostics. Thus, progress on these fronts has the potential for major impacts in our understanding of the genetic basis of physical traits. Our work also has significant impacts on machine learning and related research, in particular, (a) applying models from diverse mathematical areas such as graph theory, network analysis, recommender systems and probabilistic graphical models, (b) developing mathematical techniques that are not only applicable to linking genes with polygenic traits, but also applicable in other areas such as forensics, agriculture, or social network analysis, (c) contributing software and visualization tools to the biology community, which can then use them in analyzing more diverse data sets, including other polygenic traits, (d) exploring multi-task learning strategies for the PU setting, as yet an open problem in machine learning. In the work on learning with random noise, we have answered some long-standing questions in machine learning. The noise model in [2], class-conditional random noise, has already been studied in the literature. While some results are known under certain assumptions on the data distribution, our results generalize known results and we provide efficient algorithms for noise tolerance. It is quite surprising that even under the simple noise model, no guarantees were known for minimizing surrogate losses with noisy training data, especially given that almost two decades have elapsed since the first work on classification with random noise. Our results and algorithms provide promising strategies for more complex noise models.

Accomplishments
We have developed two methods for identifying potential gene-disease associations. The methods have experimentally been verified to produce interesting results: In particular, biologists have verified that some of the top predictions (potential gene-phenotype connections) from our methods, namely, Katz on the heterogenous network and CATAPULT[1], indeed have been referenced in the literature. We have published the aforementioned work in PLOS One, a popular open-access journal in the field of computational biology. In [1], we have shown some of our predictions for certain human diseases catalogued in Online Mendelian Inheritance in Man(OMIM) database (www.omim.org). Some of the genes that are predicted to be associated with certain diseases, have been mentioned in the literature in the context of the corresponding diseases, though their associations have not been established biologically. We have developed a web interface for biologists, wherein a set of known gene associations (for a phenotype) can be submitted, and the CATAPULT system will recommend potential genes. The web interface can be accessed at http://marcottelab.org/index.php/Catapult.

Our experimental results indicate that combining multiple sources of information from different species, significantly improves the quality of the prediction or classification. Extending the Katz method to a heterogeneous network setting, and in particular, for the gene-disease association prediction problem, and using walk-based features to learn a supervised classifier in a positive-unlabeled learning setting are novel. In [1], we show that our methods are qualitatively and quantitatively better than a number of recently proposed methods. We also show that the performances observed are not an artifact of the data set, but is indeed the efficacy of the methods using similar extensive experiments on benchmark drug-target interactions data. Another important contribution we make in [1] is about the evaluation of the methods. We observe that a leave-one-out validation may not be “right” method if one were to evaluate how biologically novel a given method’s predictions are. In particular, we show that evaluating the methods on genes with no previous associations is a better strategy, as confirmed by our qualitative and quantitative comparisons of proposed and state-of-the-art methods on OMIM and drug data sets.

The positive-unlabeled learning aspect of the biological problem has prompted us to ask important and fundamental questions in machine learning. In [2], we consider risk minimization in the presence of class-conditional random label noise --- the data consists of iid samples from an underlying ``clean'' distribution, and the learning algorithm sees samples drawn from a noisy version, where the noise rates depend on the class label. General results in this setting have not been obtained before; we are the first to provide guarantees for risk minimization under random label noise in the general setting of convex surrogates, without any assumptions on the true distribution. To this end, we develop two methods for suitably modifying any given

surrogate loss function, and show that minimizing the sample average of the modified proxy loss function leads to provable risk bounds where the risk is calculated using the original loss on the clean distribution. Our general results include certain existing results for random classification as special cases. A significant outcome of this line of research is that we resolve an elusive theoretical gap in the understanding of practical methods like biased SVM (which is also a part of CATAPULT framework) and weighted logistic regression. We show that such methods are provably noise-tolerant. Besides theoretical significance, our algorithms have an added advantage of being efficient and easy to implement. Our results will appear at NIPS, a leading conference in machine learning. In the paper, we also show experiments on synthetic and benchmark data sets demonstrating the success of the proposed methods. In particular, our methods achieve over 88% accuracy even when 40% of the labels are corrupted, and are competitive at high noise rates in many UCI benchmark data sets. The proposed algorithms in the paper already give a new family of methods that can be applied to the positive-unlabeled learning problem, but the implications of the methods here should be more carefully analyzed. There are some potential open problems we want to study, involving more complex noise models, that look more promising given our results.

Scientific Barriers

The most important challenge arises due to the sparsity in the gene-disease associations data. In particular, a majority of the genes are associated to at most one disease. Furthermore, most of the OMIM diseases have one or two gene associations established. Absence of negative examples in the data, that characterizes the PU learning setting, makes the prediction problem unique and challenging. It is infeasible to obtain negative examples in this case --- for any given phenotype, it is very hard to verify that a gene is not associated in some way with the phenotype. While a biological experiment can give clear evidence for the existence of a certain gene-phenotype association, a lack of evidence for a connection does not imply that such a connection does not exist. Biologists therefore tend to report positive associations between genes and phenotypes. There are existing supervised methods that deal with each of the aforementioned problems per se, like for e.g. sparsity of the network. Methods for learning in the PU setting have been proposed in the literature, but the machine learning community has not really considered inherently positive-unlabeled learning applications like the problem at hand. Our method CATAPULT currently tries to address (1) by pooling the examples from multiple species phenotypes to compensate the dearth of positive examples for human diseases, and (2) by using the biological prior that a tiny fraction of the large unobserved associations are likely to be positive, to randomly sample a set of “negatives”. However, it is conceivable that there are better approaches to tackle the aforementioned challenges. For example, the multi-task learning techniques, which try to learn classifiers for individual “tasks” (i.e., diseases), under the constraints that similar tasks have similar prediction functions, could help better link genes specific to diseases. Scalability is one of the challenges typically faced in network analysis. Multi-partite networks, and similarity measure based methods are extremely time consuming; construction of walk-based features is therefore a bottleneck if we are to consider walks of longer lengths. We need scalable approximations of the methods. Graph kernels (node and edge kernels) can help avoid computation of features in large graphs, but computing edge kernels itself is challenging. But the characteristics of the heterogeneous network suggest that the kernel can be computed more efficiently than what the state-of-the-art methods imply. Scalable computation of edge kernels is essential --- as a way to scale walk-based methods, and as a problem in machine learning per se.

Collaborations and Leveraged Funding

The work is done as a collaboration between Inderjit Dhillon, who is a computer scientist and an expert in data analysis, and Dr. Edward Marcotte, an expert in bioinformatics. After all analytical validation, the ultimate test of the predictions produced by the new techniques will be biological. Some of the predictions from our methods are highly relevant and as yet not confirmed by any biological studies. These predictions make good candidates for further biological validation. We will be biologically validating predicted genes for a select few traits and Dr. Marcotte's lab is well-equipped to perform the experiments. Considering the recent success of analytical modeling in identifying gene-phenotype links, we expect to discover new connections between genes and phenotypes.

Conclusions

We have observed that walk-based graph theoretic methods and PU learning yield promising results. The methods perform significantly better than previous state-of-the-art. Computational experiments demonstrate that the success of our supervised and unsupervised methods in predicting genes for diseases, and encourage further biological validation of predictions. On the machine learning front, we addressed the problem of risk minimization in the presence of random classification noise for convex losses. We are the first to provide guarantees in the general setting, and our results generalize some existing results for random label noise. We have developed two methods with provable guarantees for learning under label noise. The algorithms are efficient and achieve impressive performance even at high noise rates and are competitive with state-of-the-art methods. Our work helps shed light on the success of practical methods such as biased SVMs from the point of view of robustness to label noise.

Future Plans

Extending the theoretical results on classification with label noise to more realistic noise models such as constant-partition label noise, where different regions of the labeled space of examples could be corrupted by different noise rates, is next on our agenda. Furthermore, extending guarantees for performance metrics, when labels are noise-free and for the PU learning setting, beyond the accuracy measure is a problem of importance and interest. There has been no work on empirical

minimization under more general noise models and performance metrics yet; our current results for the class-conditional random noise model show promise for this direction of research and could potentially be extended.

Bibliography

[1] Singh-Blom U, Natarajan N, Tewari A, Woods J, Dhillon I, et al. (2013) Prediction and Validation of Gene-Disease Associations Using Methods Inspired by Social Network Analyses. PLoS ONE 8(5): e58977.
[2] N. Natarajan, I. S. Dhillon, P. Ravikumar and A. Tewari. Learning with Noisy Labels. In Proceedings of the Neural Information Processing Systems Conference(NIPS), pp. 1196-1204. 2013.
[3] N. Natarajan, I. S. Dhillon. Inductive matrix completion for predicting gene–disease associations. Bioinformatics 30, no. 12 (2014): i60-i68.

Technology Transfer