Research Statement

Amit Dhurandhar

I have realized, through performing academic and now industrial research, that appropriately modeling and incorporating domain knowledge or extra information available in practical settings can be critical for the research to be useful and thus impactful. General theories [1] with ”minimal” assumptions are good for providing insight and weak guidelines, but the results themselves are rarely useful in practice. I believe that the challenge for us as a research community is to make (maybe stronger or just different) assumptions that are acceptable in practice but which help us design more relevant frameworks and techniques that are truly applicable. A good example of this is the theory of Compressed Sensing, which by making more strict but usually seen in practice assumptions, we obtain much tighter and more useful results, compared with Shannon's more general theory. I have approached my own research in the same manner, where although the methods and frameworks I have developed are principled, most of my work has been implemented in real systems in and outside of IBM. My Ph.D. work too, was motivated by the commonly encountered in practice classification model selection problem, for which I was able to propose a formal data dependent framework. This required rigorous probabilistic analysis to not only make the problem tractable but also to characterize specific algorithms in this setting. My results although less general than [1], were much more tightly linked to the behavior of the particular classifiers and hence, the classification algorithm relative to the available data. Being able to do principled yet practically useful work has been the underlying theme of my research.

At IBM, I have also been able to abstract out interesting and more generally applicable research problems from specific projects and applications in which I have been involved over the past few years. More details are in the teaching statement, but based on this experience I have ideas for new courses, which will focus on the importance of defining relevant and novel research problems from practical applications. I believe this is a skill that needs to be developed to be successful in industrial research and even to some extent in conducting innovative and useful academic research, but has not received nearly as much attention as skills that need to be learned for solving known research problems.

In the future too, I hope to build principled frameworks and techniques motivated by real world applications. Doing this I wish to not only advance my and other related fields but also to build tools that make a difference in practice. In what follows, I will describe my major research directions and the potential impact they are likely to have in advancing the current state of the art.

1. Efficient and Accurate Prediction in Complex Instrumented Domains

Modern-day instrumented manufacturing is a multi-billion dollar industry, which encompasses diverse industries ranging from semiconductor to pharmaceutical to consumer and snack products. Starting from the initial crude state, the final product is produced by the application of hundreds of steps and tools, over a few weeks to even a few months. However, in each of these industries, there are millions (to even billions) of dollars of losses every year because of out-of-spec products. It would be extremely useful if one could effectively utilize the accrued intermediate measurements or features to accurately predict the quality of the final product, a process referred to as *virtual metrology*. Such predictions would be useful in taking early corrective actions or discarding faulty/damaged products, leading to huge savings in time, money and raw materials. In fact, I believe that *better interfacing with technology through wearable computing that may use high precision sensors to diagnose physical condition* is the next big wave, for which precision manufacturing that can be accomplished using advanced analytics is critical.

As it turns out, efficient and accurate prediction in these domains is ridden with difficulty. From a data science standpoint, we have a) non-stationary populations that drift with time, b) significant amounts of missing data and c) hundreds to thousands of features being added incrementally at each step, which makes efficient and accurate updating of models learned on the data until the previous step extremely challenging. In the last few years I have tried to satisfactorily address these issues and the algorithms I have developed have been implemented in the semiconductor fab in IBM which collects close to 350 GB of data every day. To address a) I have come up with meta-techniques applicable to any base regression method [2] that dynamically identifies stable regimes in the data and then can use (aggregate) drift information to provably improve the quality of the base predictions [3,4]. This work showcases how to use even multiple levels of not necessarily nested aggregate or coarse information along with (estimated) target mean and variance information to improve predictions. To handle b) I have developed a combination of domain specific and machine learning procedures [5] that work well in practice. For c), I have come up with a novel meta-technique [6] that is significantly more efficient than learning from scratch or low rank update procedures and is optimal for a range of methods, such as generalized least squares, ridge regression and generalized linear models, which are non-linear. The problems addressed here and the methods developed are also applicable to a wide variety of applications in computer vision, finance, retail, computational biology and many other such domains.

Short - Medium Term Goals (1-5 yrs): I wish to study in depth better ways of incorporating exact or approximate aggregate information in regression and classification models by adding appropriate regularizations. The work I have done will then serve as a baseline. Moreover, such aggregate information is available in many domains and so the methods would be applicable across multiple industries.

I would like to create new algorithms for updating a wider class of non-linear methods efficiently and accurately in the setting where many new features are added. As mentioned before, this has applications in vision, where one gets new frames in video sequence or computational biology, where we might want to learn over a small feature space and then update our model with the remaining features. I would also like to explore the possibility of simultaneously updating models with the addition of features and instances.

Long Term Goals (>5 yrs): As a long term goal, I would like to characterize what it means to be **operationally significant**. In complex instrumented domains and also in many other real applications we get statistically significant signals but they are in quite a few cases not important from the application standpoint, or in other words, not operationally significant. I would like to characterize this through learning from past actions and through appropriate abstractions based on the additional available information.

Funding Potential: Other than NSF, there is a potential here to obtain funding not only from chip manufacturing companies, such as IBM, Samsung, Intel but also from pharma and consumer product companies as building operationally significant scalable models is something that in my experience deeply interests them.

2. Data Dependent Framework for Studying Learning Algorithms

In my thesis, I proposed a novel data dependent moment based framework to study classification algorithms and model selection measures such as cross-validation, hold-out set estimation accurately and efficiently in the non-asymptotic regime [7,8,9,10]. Classification model selection is an extremely important problem in machine learning as there are many models to choose from. In particular, by focusing on the probabilistic space of classifiers induced by the classification algorithm and datasets of size N drawn independently and identically (i.i.d.) from an (estimated) joint distribution, I was able to obtain efficient characterizations for computing the moments of the generalization error. The derived formulas in these characterizations although exact (not approximations) had exponentially fewer terms than the straightforward method. This leap turned something that is surely intractable into something that is tractable for multiple commonly used algorithms [7,8,9,10]. The characterization is a robust way of studying and choosing classification algorithms as we compute these moments over all possible datasets of a certain size from the joint distribution. This approach is *closely related to the philosophy that Michael Jordan recently recommended* [11] for performing classification/regression model selection using bagged predictors rather than just building a single model. Moreover, the results being (estimated) distribution and classification algorithm dependent is a weakness and a strength. This is so because, they are less general than Vapnik-Chervonenkis (VC) theory [1], but are much more indicative of the behavior classification algorithm in the particular setting. Deploying the methodology, I was also able to study well known model selection techniques such as cross-validation and hold-out-set estimation. In fact, I was able to provide an interesting intuitive explanation for the favorable behavior of cross-validation at an intermediate number of folds (viz. 10) using my framework [8].

More recently, other such moment based methodologies have gained prominence for parameter estimation over traditional maximum likelihood and bayesian approaches that are comparatively inefficient and have a tendency to get stuck in poor local minima [12,13].

Short - Medium Term Goals (1-5 yrs): I would like to analyze other popular classification algorithms and model selection measures in this framework. I would like to devise more scalable solutions by accurately approximating the terms involved in the exact formulations. On making sufficient progress, I plan to build a software tool that runs our analysis as backend and serves as an exploratory tool that guides practitioners and academicians in choosing the appropriate algorithm.

Long Term Goals (>5 yrs): As a more long term goal, it would be interesting to see just how far such kind of analysis can be pushed to study not only classification problems but also to regression problems and other learning frameworks where the data is not necessarily i.i.d. such as in statistical relational learning or structured output prediction over finite sample sizes. This would require significantly more sophisticated analysis, which is not attenable by straightforward extensions of the current work.

Funding Potential: NSF would be the main funding source as the problem is central to data science. Moreover, given that the moments can be computed parallely in a distributed fashion, there is also a Big Data angle that is appealing.

3. Other Research

Above were two of the major research directions, however, here are some other research problems that I have worked on and I am interested in furthering the agenda.

Statistical Relational Learning: A lot of today's real world data is relational in nature (e.g., social networks, relational databases, citation graphs, etc). Statistical relational learning tries to learn from this non-iid data and perform probabilistic inference. Although effective methods have been developed in literature [20], there is still a lot that has to be done to

scale these methods in today's Big Data age and improve their performance especially when the data graph is sparsely labeled. Moreover, although a lot of theory has been developed and insight gained in the i.i.d. setting, there is a lot that can be done in the relational setting with realistic assumptions. I have done some initial work on both of these accounts. I have developed a method that works well on sparsely labeled relational graphs recently [14], by providing a simple abstraction that enables the effective use of laplacian based graph transduction methods, which are arguably the most commomly used in the graph based semi-supervised learning literature [15]. I believe that this *linking of the two sub-fields* will help in creating much better methods in both sub-fields and it is something I wish to further investigate. With regards to developing better theory, in my current research, I have derived distribution free bounds for relational classification algorithms [16,17] based on what I believe are reasonably realistic assumptions. In fact, the bound has the unique feature that it depends on the degree of dependence between instances. This property makes the bounds much tighter than previously derived bounds for similar applications.

In the future, I would like to not only derive tighter bounds but also take into account the temporal aspect as the graphs are evolving. Moreover, I would like to derive bounds that become tighter rather than looser with increasing dependence as they are an intrinsic part of how many real graphs are generated and not an artifact of having a biased sample, which is the standard statistical perspective. Network science is an extremely hot topic and hence, besides NSF, there is a potential for getting faculty awards from companies such as Facebook, LinkedIn, Google amongst others.

Active Instance Completion: Recently I have been working on a real problem that is commonly seen in the targeted advertising industry, healthcare and education. The problem entails trying to acquire missing information at a reasonable cost so as to gain accurate insight. For example in targetted advertising, we may not know the age and salary of some people, while for others we may know their age and salary but not their education. The goal is to figure out which attributes are important in determining people's buying behavior. However, because of incomplete information, the classification models have high error. Given this, we want to choose the (few) people to acquire information from – as asking everyone is impractical – so as to maximize the improvement in classification accuracy. This problem is complementary to traditional active learning and is considerably harder to design effective techniques for, since there could be multiple different sets of (input) attributes missing for each of the instances as opposed to just one (class label). I have done some work [18,19] related to this problem, where we propose ways to estimate the score for each instance based on joint density estimation efficiently for popular classification methods such as SVMs and logistic regression. I have seen that these methods perform reasonably well in practice. However, this is a relatively new problem and there is much scope to design and analyze better methods and address other variants with specific cost structures. As a long term goal, a variant of this problem is critical in building **dialog systems**, which many companies such as Apple, IBM, Google are interested in. Such systems would be useful for tech support and in call centers besides having many other applications.

Practical Frameworks for Clustering: I have recently abstracted out from real business problems a new framework for making clustering actionable, which is not captured by the current frameworks [21]. In particular, I have come up with a novel constraint, which leads to clusters that provide actionable insight. The constraint is based on the observation that organizations can take actions only at certain predefined aggregate levels consistent with their internal organizational hierarchy. There is a lot of potential to develop effective techniques under this framework that will be useful in multiple domains. There is also scope to do more theoretical work and develop clustering algorithms with guarantees for certain classes of functions (viz. submodular) [22]. Developing other or more relevant frameworks from this is also a distinct possibility.

I have shared these ideas with well established researchers who have worked in clustering such as Kiri Wagstaff, Joydeep Ghosh, Thorsten Joachim, Ian Davidson and Cynthia Rudin, some of who provided useful feedback, while all felt that the idea was novel and interesting.

4. Concluding Remarks

Based on my research directions and experience in dealing with experts from different domains (viz. manufacturing, procurement, finance, consumer products, etc.), I would definitely want to collaborate with other departments such as mechanical, electrical, biomedical amongst others as I think it is critical to perform successful research. I would also like to leverage my industry collaborations that I have nurtured over the years.

Both of the above aspects should help in acquiring funding. Regarding my current experience, I have written a couple of grant proposals with my advisor. More recently, as KDD PIC chair at IBM T.J. Watson, I have been *able to secure funding* by writing budget proposals from IBM VPs requesting funding for special events. This is from a fixed pool and is quite competitive given that the other area PICs ($>$ 35 PICs) are also funded from it.

I strongly feel that to dive deeply into any of these research problems an academic environment where one can work with enthusiastic and talented students is essential. Moreover, the excellent faculty at your university can further help me facilitate this. In particular, I would love to collaborate with the labs/people mentioned in my cover letter if given a chance at your prestigious institution.

References

[1] V. Vapnik. *Statistical Learning Theory*. Wiley & Sons, 1998.
[2] A. Dhurandhar. Multistep Time Series Prediction in Complex Instrumented Domains. *IEEE ICDMW*, 2010.
[3] A. Dhurandhar. Using Coarse Information for Real Valued Prediction. *DMKD*, 2013.
[4] A. Dhurandhar. Improving Predictions Using Aggregate Information. *ACM SIGKDD*, 2011.
[5] S. Weiss, A. Dhurandhar and R. Baseman. Improving Quality Control by Early Prediction of Manufacturing Outcomes. *ACM SIGKDD*, 2013.
[6] A. Dhurandhar and M. Petrik. Efficient and Accurate Methods for Updating Generalized Linear Models with Multiple Feature Additions. *JMLR*, 2014.
[7] A. Dhurandhar and A. Dobra. Probabilistic Characterization of Random Decision Trees. *JMLR*, 2008.
[8] A. Dhurandhar and A. Dobra. Semi-Analytical Method for Analyzing Models and Model Selection Measures. *ACM TKDD*, 2009.
[9] A. Dhurandhar and A. Dobra. Probabilistic Characterization of Nearest Neighbor Classifiers. *IJMLC*, 2012.
[10] A. Dhurandhar. Bounds on the Moments of an Ensemble of Random Decision Trees. *KAIS*, 2014.
[11] A. Kleiner and A. Talwalkar and P. Sarkar and M. Jordan. The Big Data Bootstrap. *ICML*, 2012.
[12] A. AnandKumar and D. Foster and D. Hsu and S. Kakade and Y. Liu. A Spectral Algorithm for Latent Dirichlet Allocation. *NIPS*, 2012.
[13] B. Boots and G. Gordon. Two Manifold Problems with Applications to Non-linear System Identification. *ICML*, 2012.
[14] A. Dhurandhar and J. Wang. Single Network Relational Transductive Learning. *JAIR*, 2013.
[15] J. Wang, T. Jebara and S. Chang. Semi-Supervised Learning Using Greedy Max-Cut. *JMLR*, 2013.
[16] A. Dhurandhar and A. Dobra. Distribution-free Bounds for Relational Classification. *KAIS*, 2012.
[17] A. Dhurandhar. Auto-correlation Bounds for Relational Data. *Mining and Learning on Graphs Workshop in ACM SIGKDD*, 2013.
[18] K. Sankarnarayanan and A. Dhurandhar. Intelligently Querying Incomplete Instances for Improving Classification Performance. *ACM CIKM*, 2013.
[19] A. Dhurandhar and K. Sankarnarayanan. Improving Classification Performance through Selective Instance Completion. *submitted*.
[20] L. Getoor and B. Taskar. Introduction to Statistical Relational Learning. *MIT Press*, 2007.
[21] A. Dhurandhar and X. Wang. Actionable Clustering. *submitted*.
[22] A. Dhurandhar and K. Gurumoorthy. Symmetric Submodular Clustering with Actionable Constraint. *DISCML NIPS*, 2014.