

Goals for Scholarly Activity

Amit Dhurandhar

In my three and a half years as a P.h.D. student one of the things I have realized is that, appropriately incorporating domain knowledge is essential in deriving useful results. General theories with "minimal" assumptions are good in providing weak guidelines but the results themselves are rarely useful in practice, since they are pessimistic. An example of this in machine learning is Vapnik-Chervonenkis (VC) theory, which is a fundamental piece of work that provides a principled approach to study learning algorithms but the generalization bounds provided by this theory are usually weak. Consequently, I believe that the challenge for us as a research community, is to make (may be stronger or just different) assumptions that are acceptable in practice but which help us in deriving tight results that are truly applicable. A striking example of a recent idea that follows this trend is the theory of Compressed Sensing. By assuming a special structure on the input signals we can greatly reduce the sample complexity of reconstructing the original signal when compared with Shanon's theory. Hence, though Shanon's theory is applicable to a wider range of signals, by assuming some structure – which is seen in many real life applications – we obtain much tighter results. I have incorporated this ideology in my own research and I hope to continue doing so in the future, in an attempt to eventually build tools that make a difference in practice.

In what follows, I will describe some specific problems and the potential impact my research is likely to have in advancing the current state-of-the-art.

1. Classification Model Selection

In my current research I have proposed a novel moment based method to study classification algorithms and model selection measures accurately and efficiently over finite sample sizes. Deploying the methodology I was able to thus far provide interesting explanations for the behavior of cross-validation and study popular algorithms such as Random Decision Trees, Naive Bayes and K-Nearest Neighbor accurately and efficiently. Here are some promising directions for the future.

Short - Medium Term Goals (1-3 yrs): In the near future 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 that I have developed in my P.h.D. 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 (>3 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 problems over finite sample sizes. The current analysis applies when the output is discrete which is alright for classification but for regression we would have to extend the theory to be applicable to continuous outputs.

Intellectual Merit and Broader Impact: I believe that this work has the potential of impacting industrial and academic research alike, in the near future.

- *Impact on industry and other fields:* We know that in today's day and age adaptive classification models find applicability in a wide spectrum of applications ranging over various domains. Financial Firms deploy these models for security purposes such as fraud detection, intrusion detection. Giant chains of Supermarkets use these models to figure out which group of items are generally bought together by the customer. These models are used extensively in Bioinformatics for problems such as gene classification based on functionality, DNA/protein sequence matching, etc. Today's state-of-the-art search engines also use classification models. This is just a snapshot of the entire range of applications they are used for.

Noticing the wide applicability of classification models and the sheer extent of their number, it is but a desired goal that we choose the correct model for our specific application. What is currently missing is a principled approach that accomplishes this in a scalable and accurate fashion. I through my research hope to develop a tool that will help realize this goal.

- *Impact on the machine learning and data mining research:* I believe that the research will assist in providing new insight into the behavior of classification models and model selection measures. The framework may be used as an exploratory tool for observing and understanding models and selection measures under specific circumstances that interest the user. It is possible that other related problems may also be framed in an analogous fashion leading to interesting observations and consequent interpretations.

2. Usable Theory for Relational Domains

A major portion of the theory developed in machine learning is based on the assumption that the data is independently and identically distributed (i.i.d.). This assumption however is rarely justifiable in practice. In fact, most of the real life data is in the form of graphs, relational databases etc. where individual datapoints are related.

Short - Medium Term Goals (1-3 yrs): In my current research I have derived distribution free bounds for relational classification algorithms which have this unique feature that they depend on the degree of auto-correlation between individual datapoints. This property makes the bounds much tighter than previously derived bounds for similar applications. In the near future I would like to tighten these bounds especially at high auto-correlations making them more useful in practice. It would also be desirable that the derived bounds reduce to known inequalities in the i.i.d. case which will help in better understanding the behavior of the bound potentially leading to further tightening of the bound.

Long Term Goals (>3 yrs): As a more long term goal I would like to build theory that would be useful in practice for relational or the more general class of non-iid data that is omnipresent in today's world. The theory would focus on better bounds, extending the moment based methodology to relational data thus presenting an avenue to study relational algorithms accurately and efficiently over finite sample sizes and developing efficient approximation schemes to conduct inference in these models.

Intellectual Merit and Broader Impact: I believe, this theory can potentially have the following kind of impact.

- *Impact on industry:* Most of the real world data is relational in nature (i.e. graphs, relational databases etc.). Building usable theory for such data will inexorably lead to better decision making leading to savings in time and money. The act of obtaining practically useful results in real applications will become more of a science than an art through ad-hoc experimentation.
- *Impact on the machine learning and data mining research:* The derived bounds that depend on the degree of auto-correlation besides being useful in practice, also are proof of concept as to how such bounds can be derived in the future. Moreover, such bounds can provide insight relating several important notions. For example, the current auto-correlation dependent bound I have derived relates actual sample size to an important notion called effective sample size. Effective sample size is essentially the size of an i.i.d. sample which has the same amount of information as the relational dataset at our disposal.

3. Classification and Sampling in Relational Domains

Unlike the classification algorithms in i.i.d. domains, the classification algorithms in relational domains (collective classification algorithms in particular) are more complicated and computationally intensive to train and infer. In general, even other tasks such as sampling are much more involved in relational domains. With this in mind I have the following set of goals which I plan to accomplish.

Short - Medium Term Goals (1-3 yrs): Given that collective classification algorithms are usually more complicated than independent classification algorithms it is important to know when this added complexity is worthwhile to handle. In my current research I have provided necessary conditions for using collective classification algorithms. In the future I plan to find what conditions might be sufficient to justify their use. I also plan to suggest simple yet effective collective classification methods that may be used as baselines to evaluate more sophisticated collective classification algorithms.

Long Term Goals (>3 yrs): As mentioned before sampling in relational domains is a major challenge. In the future, I would like to propose sampling methods with proven guarantees that are also efficient. I would also like to delve into other interesting problems in SRL such as entity resolution, link prediction which have wide range of applications (e.g. in social networks).

Intellectual Merit and Broader Impact:

- *Impact on industry:* Providing necessary and sufficient conditions for using collective classification can guide practitioners choose the appropriate algorithm. This will lead to better performance and higher efficiency in industrial applications. Similarly, efficient relational sampling methods can too have a huge impact on industrial applications in terms of performance and speed.
- *Impact on the machine learning and data mining research:* The necessary and sufficient conditions for using collective classification algorithms may be used to develop new specialized algorithms that are better suited for the particular setting. Thus, a new specialized relational classification algorithm would be better than the current ones when the auto-correlation is medium to high but may be worse when the auto-correlation is low, in which case we may use an independent classification algorithm.