Xin Du

Research Statement

My research is motivated by the challenge of discovering interesting patterns from the data. Specifically, the objective is to find subpopulations which can be summarized by descriptive variables, like 'age $\geq 25 \wedge$ Smoker $=$ yes'. These subgroups indicate that the interactive dependencies between target variables are exceptionally different from those ones in the whole data. To do so, multiple statistical measure and modeling methods are proposed to summarize those patterns. And specific search algorithms are implemented to enable the discovery of interesting subgroups within sufficient time and computing budget. My research contributes to the machine learning and data mining communities through providing algorithms and models for finding such interesting local patterns on diverse datasets and application scenarios. This is quite useful for understanding the performance of machine learning models. Because the exceptional interactive patterns in subgroups may lead to exceptional behavior of the models. Recent concerns in trustworthy machine learning suggest that machine learning systems need to be rigorously tested before being deployed in safety-critical applications. This problem can be theoretically regarded as the generalization of machine learning models, which is the primary goal of learning theories. The general methodology for testing the generalization ability of a model is to use a small set of hold-out data. Even though this methodology is useful, but due to the reason that the limited test data only include part of the modes which may appear in deployment environments, it may also introduce biases such as over-confident prediction. For instance, an adversarial example which was created by adding some nuisance noise would mislead the prediction of machine learning models. This problem indicates that a single aggregated statistical measure is not sufficient to provide trustworthy evaluation for the performance of machine learning models in complex scenarios.

Exceptional Model Mining

Exceptional Model Mining (EMM) is a local pattern mining approach that cares about how differently a model would perform in subpopulations, as compared to the same model but fitted on the whole data. The main interests of EMM is to discover exceptional interactive patterns in the subset of data with descriptive statistics by comparing the patterns related to models' performance on whole data and subgroup data. Model is a tool to summarize the patterns in the data. Here are two study cases from my previous work: Fairness in network representation: we argue that the structural heterogeneity in networks can bias the network representation models across subgroups, which will prevent people from building fair decision making models for downstream tasks like node classification or link prediction. In this work, graph data are selected by the associated descriptions, as the input of the node representation learning model. The exceptionality of these subgroup data are evaluated by comparing the reconstructed relation matrix between nodes with the relation matrix derived by the whole data. We evaluate the statistical significance of the discovered exceptionality score of subgroups by applying a kernel two-sample test. This work shows that without knowing the graph data itself, with a black-box trained node representation learning model, the indicated node relations could be very different regarding to the certain attributes associated with the data. In other words, attribute information itself can be distinguishable for the graph structures, which raises alarm for the trustworthiness of pre-trained models, for which we do not have access to the original training data.

Spatio-temporal behavior on collective social media: behavior in this setting can be exceptional in three distinct ways: in terms of spatial locations, time, and texts. We develop a Bayesian Non-Parametric Model (BNPM) to automatically identify spatio-temporal behavioral patterns on the subgroup level, explicitly modeling the three exceptional behavior types. The proposed proposed graphical model with Gibbs sampling inference method allows us to learn the summary statistical pattern behind the multi-modal data distributions. This allows us to provide an effective evaluation method to measure the exceptionality of a behavioral pattern and to employ it in finding exceptional subgroups with collective social behavior. The main contribution of this work is to show that Bayesian probabilistic model could be used to extract complex patterns from multi-modal data. These patterns are shown to have strong association with the descriptive attributes which could indicate the exceptionality in subgroups.

Causal Inference

Among those interactions between target variables, causal relation is of great interest and benefits for the current research of machine learning and data mining. Due to the natural property of stability, learning causal effects from observational data greatly benefits a variety of domains such as health care, education, and sociology. For instance, one could estimate the impact of a new drug on specific individuals to assist clinical planning and improve the survival rate. Causal relation could prevent us from learning spurious correlations between variables which would lead to false discoveries. In this study, we focus on studying the problem of estimating the Conditional Average Treatment Effect (CATE) from observational data. Heterogeneity is an important property that exists across the data distributions, with regard to descriptive variables. For instance, one clinical treatment could be quite effective on specific groups of people but ineffective on other groups. In order to infer the causal relation for certain demographics rather than average, we propose a neural network framework ABCEI, based on recent advances in representation learning. To ensure the identification of the CATE, ABCEI uses adversarial learning to balance the distributions of covariates in the treatment and the control group in the latent representation space, without any assumptions on the form of the treatment selection/assignment function.

Trustworthy Machine Learning Systems

Descriptive variables could indicate certain properties in subgroups of data. These groups could be used to make predictions for a specific target variable. Hence, those conjunction and combination of descriptive variables could be used as predictive local decision rules for classification. Local decision rules could provide both high predictive performance and explainability. However, the stability of those explanations given by the rules is still underexplored. We propose two regularization terms to improve the robustness of decision rule ensembles. The graph-based term is built by decomposing invariant features using a given causal graph; the variance-based term relies on an additional artificial feature that can restrict the model's decision boundary within groups. This work indicates that in some situation, we cannot trust the learned representations to generalize well on the prediction tasks in another environment. The main concern that is derived from this work is that: given a learned representation without the access of training data, how much information can we know about the trustworthiness of the model with collected testing data? By referring to the trustworthiness, we mean that the generalization ability of the model under distributional shifts, if we can know such mechanism. Following this direction, in another work, we propose to evaluate the trustworthiness of image classification models by random data augmentation. Specifically, we are interested in investigating the family of vision transformer based image models like ViT and DeiT. Considering their success and high-stake application of image models like clinical health care, it is necessary to develop a systemic model test framework to the evaluation of vision transformer based model families. We propose to develop a set of tools which could generate a bunch of test samples for different test tasks to evaluate the model, and formulate a systemic report to provide a comprehensive understanding for the model's performance. Different from adversarial examples, this work explores augmented data by elementary perturbation with patch level inputs. The aim is to explore and discover potential semantic mismatch between the model prediction and the ground truth labels.

Cognitive Computing and Robustness

Using historical data to predict the potential outcome of a policy is of high importance in Cognitive Computing due to the high-cost of randomized controlled trials. Deep learning models are well-known for their superb performance based on large amount of training data. However, they are also known to be vulnerable to copy or amplify the biases in training data. This problem could be more risky when we do not have the access of the training data and the downstream decision making tasks are purely dependent on the pre-trained models. For instance, models learned from training set cannot perform well on data collected from other sources, e.g. another group of people, which is partly caused by the reason that the model learns spurious features in the data instead of the true features that indicate the semantic label. Two future paths of studies are raised based on this concern: first, I would like to develop methods that can connect the performance of the model on collected test data, to the data generating process behind the training data, so that we are able to understand more about the modeling process, and how much we can trust the model's output in the new environment. This study is challenging because the lack of information about the potential shifts between data collected from different environment, hence, domain knowledge or extra supervision should be considered to gain more information. In the second path, I would like to explore the connection between the decision making process, and the data shifts. In some situation, shifts in the data are not misled by spurious correlations, but the strategic modification which could influence the decision making process. This problem raises specific concern in scenarios such as financial credit justification, and the challenges are also derived from the unknown data generating process. Potential causal analysis by domain knowledge and extra supervision are needed to develop solutions for this problem, such as the descriptive attributes.

Hence, instead of focusing on developing new models, one of my future research is to focus on the data. I would like to develop methods to analyze the potential biases that the data could introduce to the model, and what the outcome would be biased according to the characteristics of the data. Based on that, I would like to study different data augmentation methods, such as generative model, to generate more data by debiasing algorithms.

In another direction, I would like to explore the uncertain quantification methods to let the model report the confidence level for the prediction they made, and to propose method that can tackle the confidence mismatch problem during prediction. The aim is to build a system that can answer 'I'm not sure' to the out-of-distribution samples, and let the human take over when it is necessary. By doing this, a safe autonomous / policy making system can be built to improve the efficiency and reduce the risk.