論文の内容の要旨

論文題目　Causal Machine Learning from Small Data: A Data Augmentation Approach
（小データからの因果的機械学習： データ拡張によるアプローチ）

氏　　名　手嶋毅志

The twenty-first century has seen a rapid and widespread development of automated intelligent systems, such as computer-assisted diagnosis, object detection, speech recognition, and automated translation. Many of such systems are powered by *statistical machine learning*, a paradigm in which various methods for learning from data have been developed. The designs of statistical machine learning parallel *inductive inference* in logical reasoning: learning from observations — the *training data* — and drawing conclusions about the unobserved — the *testing instances*.

For inductive reasoning to be helpful, some "uniformity of nature" principle is required. In statistical machine learning, or more generally in statistics, analogues of such a *uniformity principle* are embedded in one form or another, such as an assumption that the data is an *independent and identically distributed* sample of a probability distribution or that one should be a rational decision-maker. Corresponding to each of such different premises, reasonable inferential rules — the *learning methods* — have been derived.

Causality is a form of such a *uniformity principle*, which is distinctive in human's course of thinking and perception of the world. What is causality? Some philosophical theories of causation emphasize that causality is about *difference-making*: without the cause, the effect would not have happened. In particular, some view causal relations as potential routes by which the world can be manipulated or controlled, i.e., difference-making about what *potential outcome* is realized by some intervention. Others emphasize that causality is about *production*: causes *bring about* their effects. In particular, some appeal to the concept of *causal mechanisms*, i.e., complex systems producing some behaviors through invariant direct interaction of a number of parts, as is often considered in the explanatory practices of the special sciences.

Founded by these two viewpoints of philosophical theories of causality, namely *interventions* and *mechanisms*, the *statistical frameworks of causal modeling* have been developed since the end of the 20th century. Such frameworks enable natural formulations of causality-related quantities based on probability theory, such as the average difference made by an intervention.

The pragmatic utility of acquiring the knowledge of such intervention-related quantities is rather apparent: one can use it for making informed decisions about the actions, e.g., answering questions such as what intervention to perform to get a favorable result. From this viewpoint, the knowledge of detailed mechanisms is only a *means* to infer the consequences of our interventions. On the other hand, our intellectual curiosity to learn about the causality of nature seems to go beyond the pragmatic utility of knowing interventionistic quantities. Indeed, elucidating a mechanism has been a gold standard for explanations in scientific practice, even when making interventions is not necessarily an immediate target

in such fields.

Then, a natural question arises: are there pragmatic motivations for finding out the detailed causal mechanisms when the knowledge may not be relevant to any interventions we can implement? This dissertation provides a partial but concrete answer: the knowledge of causal mechanisms can facilitate *learning from small data* in *statistical machine learning* for predictive modeling. We provide the answer by designing the methods to incorporate the knowledge encoded in statistical causal models into the learning process.

Learning from small data, despite the rapid progress in the methodology of machine learning, remains an essential challenge in the field. When data is limited in quantity, it is important to incorporate appropriate prior knowledge about the nature of the data in order to learn an accurate predictor. In this dissertation, we approach the small-data learning problem from the perspective of exploiting known or acquired causal knowledge. The general idea is to incorporate the *statistical independence* relations implied by the statistical causal models into the machine learning procedures by developing *data augmentation* strategies.

The following is the chapter organization. Chapter 1 provides the conceptual background and declares the central statement of this dissertation, and it is followed by Chapter 2 where we review the *structural causal framework* of statistical causal modeling. Specifically, we review two interrelated formulations, namely the structural causal models (SCMs) and graphical causal models (GCMs). The two types of models are in a hierarchical relation: SCMs capture the quantitative knowledge of the data generating mechanisms expressed using deterministic functions, and GCMs retain only the coarser qualitative knowledge of the dependency relations in the data generating mechanisms expressed using a graphical representation.

Following these introductory chapters, in the main Chapters 3 and 4, we develop the proposals for exploiting the knowledge of the causal models for supervised machine learning. Concretely, in Chapter 3, we consider the case that the graphical representation of a GCM is either estimable or known thanks to domain experts, and in Chapter 4, the case that partial knowledge of the deterministic functions of an SCM is estimable from the data of a relevant domain. When the GCMs or SCMs characterizing the data generating mechanisms are (partially) known, we can infer some properties of the probability distribution of the data, namely certain statistical independence relations. However, it is not straightforward to incorporate such knowledge into predictive modeling. Therefore, in these chapters, we introduce *data augmentation* methods that allow us to exploit the knowledge encoded in the causal models for supervised machine learning in a manner that is independent of the predictor's model class which we use. The proposed methods enjoy theoretical guarantees of *excess risk bounds* indicating that the proposed methods suppress overfitting by reducing the apparent complexity of the predictor hypothesis class. Using real-world data conforming to the problem setups, we also provide numerical experiments showing that the proposed method is effective in improving the prediction accuracy, especially in the small-data regime.

We dedicate Chapter 5 to presenting a theoretical result that reinforces the justification of the method of Chapter 4, which relies on the modeling technique called *invertible neural networks* (INNs). As a recently emerged function approximation model, the INNs had not been given a theoretical guarantee of their representation power, i.e., whether the model class theoretically has sufficient flexibility to approximate various complex functions. This was a critical concern that could undermine the applicability of the proposed method of Chapter 4 to a broad range of applications. The results in Chapter 5 are affirmative: the INNs used in Chapter 4 enjoy a theoretical representation power guarantee, namely that they are *universal approximators* for a fairly large class of smooth invertible maps. We use Chapter 5 to discuss this result in length because it is also an interesting theoretical result in its own right whose scope is not limited to supplementing Chapter 4.

Finally, in Chapter 6, we summarize the overall conclusion of the dissertation and discuss further the possibilities of future research directions. To summarize, the chapters of this dissertation jointly provide the affirmation of the thesis that the causal knowledge captured by statistical causal models can be helpful in tackling the small-data learning problems in statistical machine learning. The proposed strategy for exploiting the causal knowledge is based on data augmentation, and thus the proposed methods can be readily combined with virtually any supervised learning method for learning a predictor.