Statistica Sinica **37** (2027), 1-20
doi:https://doi.org/10.5705/ss.202023.0202

ADDRESSING LABEL NOISE IN CAUSATION CLASSIFICATION VIA KERNEL MEAN EMBEDDINGS

Pingbo Hu and Grace Y. Yi*

University of Western Ontario

Abstract: A basic task in causal inference is to determine whether a cause-effect relationship exists between two sets of variables, akin to a binary classification problem. Given a sequence of independent and identically distributed paired vectors, one can use the *kernel mean embedding of probability distributions* to map empirical distributions into a reproducing kernel Hilbert space and then train a classifier in that feature space to predict the presence of a causal relationship for future pairs drawn from the same distribution as the training pairs. This strategy, however, is vulnerable to label noise (mislabeling), a common issue in causation studies. In this paper, we analyze and quantify mislabeling effects. We develop a valid learning method that explicitly accounts for label noise and establish theoretical results accordingly.

Key words and phrases: Causation learning, classification, kernel mean embedding, label noise.

1. Introduction

Learning cause-effect relationships has attracted extensive attention in both statistical and machine learning communities. The *potential outcome* framework, originating from Splawa-Neyman, Dabrowska and Speed (1990), is a popular statistical approach to infer causality (Rubin, 1974). Alternatively, learning causal relationships among variables can be framed as a supervised classification problem. In principle, all relevant variables, including causes, outcomes, and confounding factors, can be organized into ordered pairs (X, W), where X is a candidate cause and W is a candidate effect. One may exhaustively enumerate distinct (X, W) pairs and assign a binary label, 1 or -1, to indicate whether X is the cause of W using a study-specific labeling rule. The resulting task then becomes a binary classification problem with the binary label, denoted l, as the output and (X, W) as the input.

A motivating example is the SUP3 dataset from the Kaggle competition (Guyon, 2013), consisting of $n \triangleq 162$ variable pairs across diverse domains, such as chemistry, climatology, ecology, economics, engineering, epidemiology, genomics, medicine, physics, and sociology. Each pair is labeled as either 1 or -1, indicating the presence or absence of a causal relationship within the pair.

*Corresponding author. E-mail: gyi5@uwo.ca