Handling Missing Data with Graph Representation Learning

Jiaxuan You

Abstract

Machine learning with missing data has been approached in two different ways, including feature imputation where missing feature values are estimated based on observed values, and label prediction where downstream labels are learned directly from incomplete data. However, existing imputation models tend to have strong prior assumptions and cannot learn from downstream tasks, while models targeting label prediction often involve heuristics and can encounter scalability issues. Here we propose GRAPE, a graph-based framework for feature imputation as well as label prediction. GRAPE tackles the missing data problem using a graph representation, where the observations and features are viewed as two types of nodes in a bipartite graph, and the observed feature values as edges. Under the GRAPE framework, the feature imputation is formulated as an edge-level prediction task and the label prediction as a node-level prediction task. These tasks are then solved with Graph Neural Networks. Experimental results on nine benchmark datasets show that GRAPE yields 20% lower mean absolute error for imputation tasks and 10% lower for label prediction tasks, compared with existing state-of-the-art methods.