CHAPTER I
INTRODUCTION

[about data]

Nowadays, data plays an important role in both academia and business. They can find some business opportunities from their data. Then, the field of study that unifies engineering, mathematics, statistics, and computer science emerges, called data science. Data science is broadly used in business and industry to find the business opportunity hidden in data. Chatterjee et al. (2021) [**?**] concluded that using data science in an enterprise increases accuracy in innovative decisions and the potentiality of competition in business.

[tell that missing data is a problem often met]

However, not only mining or modeling data but also data preparation is a task that machine learning practitioners have to do. They spend much time in cleaning and organizing data. One of the common challenges is missing features which is the problem of absence for some values in the given datasets. This problem is relatively common in almost all research. It can significantly affect results obtained from the data [**?**] in sense of learning.

[causes of missing data]

There are many causes of missing data. First, the problem of measuring instruments is a cause that often occurs in a production line. Second, some missing data can emerge from the nature of data, such as different factors corrected between males and females in a clinical survey or the disappearance of a patient during the treatment process. Third, the lack of knowledge of the sample can also result in missing data. There may be the case of the sample giving a "don't know" answer to an observer in some complicated question.

[significance of missing data]

Why is the missingness problem noteworthy, especially in machine learning and deep learning? The most trivial answer to this question is about computation of almost machine learning algorithms. Including deep learning emerging the state-of-the-art in the current era, they calculate the fixed size of the representation vector, sometimes called embedding. However, without preprocessing, missing data leads to the problem that we cannot feed the data directly into a model because of the absence of some values inside them.

[Imputation: the most basic solution]

One of the simple way is to basically impute these missing values by a constant value, mean or interpolation. However, these imputation may not reflect the actual dataset. Actually, we do not even know the actual dataset. In addition, some missing behavior is also informative such as unrating the product from a customer which may mean dissatisfaction but the customer avoids giving a reason directly. So imputation seems that we do not concern the hidden information about missing data. Moreover, Ma, et al. (2018) [**?**] commented that the imputation method ignores uncertainties of data which means as the same as hidden information said earlier.

[More a bit Imputation but with learning from data]

Not only basic imputation methods, but many work also investigated imputation based methods by learning from data. However, it still relies on the assumption that all missing values can be replaced as a number. But not all missing scenario can be imputed, for example, unrating the product. Also the case of some non measurable feature depending on others, the value 0 and a missing value are not equivalent. Moreover, these methods is independent from the learning process. In other words, the same model may be lead to different way by different imputation methods. The only way is to cross test all possible imputation methods to see whether which method is the best one for the desired model. This still requires much effort and time-consumption in the process of feature engineering.

[Tabular data: a format of data often used and also meet missing]
Tabular data is a format of data that displays in rows and columns. One of them corresponds to the collected features of each data sample. It is a ubiquitous data format used in practical applications in many domains such as enterprise operations, manufacturing, clinics, and surveys. For example, doctors or medical technologists may want to predict the chance of a specific disease for a patient from the patient's historical records that are stored in the form of a table. Missing data problem is also a challenge in modeling tabular data. Even though tabular data has its own structure represented by feature vector, missing values can eliminates the structure of tabular data so that model cannot learn or compute on such data.

[It's time to tell about graphs and GNNs: graph first]
So we need a learning algorithm or a data structure that is more flexible in computation rather than using the fixed length of feature vector fed into algorithm directly. This work concentrates on graphs data structures and graph neural network (GNN) algorithms. As we have ever seen in data structure, graphs are flexible in how to storage data via edges and nodes. Moreover, they are suitable to be flexibly computed by using edges.

[It's time to tell about graphs and GNNs: GNNs next]
GNNs are a class of neural based methods designed to perform inference on graph data. It has emerged as a promising approach for modeling structured and semi-structured data, such as social networks [**?**], molecular structures [**?**], and knowledge graphs [**?**]. Particularly when dealing with high-dimensional or incomplete datasets, GNNs have demonstrated superior performance in various applications, as they effectively capture complex dependencies among nodes and edges in a graph [**?** , Kipf & Welling 2017]. Moreover, GNNs have advantages over traditional machine learning methods as they do not rely on assumptions about independence of instances and features. This makes GNNs are more flexible in learning with respect to both data they use and computation they do.

[Conclude to our expectation to use graphs and GNNs to address missing data problem]

Therefore, the objective of this work is to investigate the new method to learn representations or embeddings of tabular data containing missing values by using GNNs. We also aim to find a new way to represent tabular data by a graph. The proposed algorithm will allow users to directly feed a data point from tabular data containing missing values without any preprocessing process. Moreover, the model can be trained in an end-to-end fashion.

1.1 Research Objectives

Problem Statement

Missing data, which often occurs in many domains of tabular data, can lead to biased or inaccurate models if not handled properly. Current methods for handling missing data, such as imputation, can introduce additional assumptions and may not capture the underlying structure of the data. Moreover, doing so in the process of feature engineering takes much time and effort of practitioners. We seek to develop an approach that can effectively handle missing data in datasets without relying on process of data preparation for missing values such as imputation.

Objectives

The primary objective of this research is to investigate a GNN-based framework that can address the challenge of missing data. The further detail of our objectives is as follows:

1. develop an algorithm to construct graph structures for the representation of data that can be used for computation of the corresponding GNN prediction algorithm,
2. develop a GNN algorithm that can be used for complete data, and

3. develop a GNN algorithm that can be trained by datasets containing missing values without any preprocessing, and also can predict the corresponding label of the given data that may contain missing values.

1.2 Outlines

In Chapter 2, we explain more detail related to our work including graph representation, GNNs and related work. Readers can find deep detail and also formal definition in the section preliminary of this chapter. And then, we describe how we perform this research in Chapter 3: methodology.