Chapter 2
Data

ABSTRACT

The initial step for a data scientist when addressing a business question is to identify the data type, as not all types can be employed in data mining analyses. Accordingly, the data scientist must select a suitable data type that corresponds to the data mining technique and classify the data into categorical and continuous types, regardless of the source of the data. Quality control is a significant factor for the data scientist, particularly if data collection was poorly administered or designed, leading to issues like missing values. Once the data scientist has acquired a relevant dataset, they should inspect the outliers associated with each feature to make sure the data is suitable for analysis. Observing outliers through data visualizations, such as scatter plots, is a common practice among data scientists, highlighting the crucial role of data type determination.

INTRODUCTION

Though the Administrators or Domain Experts in each industry see how necessary to use data analysis to solve problems or to drive the organization, though they and are aware of the advances in data analysis using data mining techniques, they cannot link the existing data to individual data processing techniques. In many cases, it was found that the development of data visualization is sufficient for data analysis to answer questions without the need for data mining techniques. On contrary in some cases, the organizations do not have enough data to be analyzed to respond to the determined questions.

The main reason is that administrators or domain experts in the organization lack understanding of which data can be used and how the questions should be asked in order to gain useful data. Many organizations have developed their domain experts to become data scientists and trained them in a wide range of data analysis. The positions of data scientists are necessary as they can use the existing data to analyze and solve problems in a timely manner. For example, in the medical industry, patient data can be used to analyze the treatment. The use of data in each form of data analysis is therefore an important point for data science students.

DOI: 10.4018/978-1-6684-4730-7.ch002

ATTRIBUTE

The attributes of the data are used to define the scope of the Data Object. For instance, the data attribute is gender, and data object is male, female, or genderless. Data attributes are also used in different contexts (Inmon and Lindstedt, 2015). When data scientists write programs, data attributes are viewed as variables and considered as features. Data scientist collects data in a dataset consisting of a feature that stores multiple records. In data science, these records are called "Instance". The structure to store Feature, Data Objects, and Instances, is as follows (Gru¨tter, 2019: Angiulli & Fassetti, 2021).

Table 1. The example of feature, data object, and instance

First Name	Last Name
Jirapon	Sunkpho
Sarawut	Ramjan
Kom	Campiranon

In Table 1, there are totally 2 features consisting of first name and last name. The first name features are Jirapon, Sarawut and Com while the last name features are Sunkpho, Ramjan and Campiranon, respectively. When data objects from multiple features are combined, they become the Instances. From Table 1, there are 3 Instances consisting of Jirapon Sunkpho, Sarawut Ramjan and Kom Campiranon, respectively.

Data Dimensionality: Is the number of features within a dataset (Li, Horiguchi and Sawaragi, 2020). In data science, it focuses on analyzing the data in various fields without focusing on storing data from the beginning. The number of dimensions is an issue that the data scientist must consider in order to select only the features that can support the analysis. Therefore, the data scientists need to reduce the number of dimensions to have the features necessary for data analysis. This allows data scientists to reduce the time and digital resources required to process massive instances of datasets.

Resolution: Is a measure used to store Data Object in each Feature. In Data Science (Wang et al., 2019), when data scientists collect datasets from multiple sources, there is often a problem with storing the same data across multiple measures. For example, Feature height from dataset A is collected in centimeters, whereas Feature height from dataset B is stored in inches. Therefore, before implementing those datasets, data scientists must perform Attribute Transformation or changing the shape of the Data Object to be in one form or another

DATA TYPE

Different data mining and visualization techniques use different types of data. It may use only one type of data or a combination of different types. It is a great challenge for Data Scientists to analyze the data to identify whether it is consistent with the data mining techniques and to answer the determined questions. The data can be divided into 2 types.