6

PYTHON AND DATA PRE-PROCESSING

Mohd Hilmi Hasan

6.1 BIG DATA

In the Fourth Industrial Revolution (IR 4.0) era, data has become an important asset to an organisation. The term Big Data is formally defined by Gartner Inc. [66] as "...high-volume, and high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation". Simply put, big data means data that is so huge, fast, and varied that it is impossible to process using traditional methods.

Big data is also commonly defined based on its characteristics of "Vs" Despite many definition variations (4Vs, 5Vs, 6Vs, 10Vs etc.), for the purpose of learning, Figure 6.1 shows containing 10Vs definition of big data.

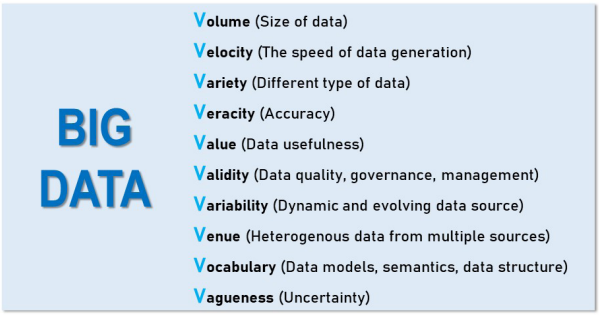

Figure 6.1 10 Vs of big data and the definitions

Big data can be of various types. Table 6.1 shows examples of big data types and their related processing. In this book, the examples are related to the numerical data and for learning purposes, their size is small so that it is suitable, yet adequate enough, to be tested on a normal computer.

Table 6.1 Example type of big data and processing

Type of Data	Common Processing Involved
Image	Transformation, coding, compression, segmentation, description, noise removal, recognition
Video	Splitting, time profiling, decoding, encoding, noise removal
Audio	Noise removal, recognition, feature analysis
Numerical	Cleaning, feature selection, dimensionality reduction, transformation

6.1.1 Data Dimensionality

Dimensionality refers to the number of attributes a dataset has. For example, equipment data contains a vast number of variables such as temperature, speed, pressure, etc. Each of these variables are represented as columns in the dataset. A high dimensional data is where the data contains an astoundingly high dimension that makes the calculations become extremely difficult.

The more variables added to a dataset, the more difficult it becomes to carry out the prediction of certain elements. This means that adding more variables will reduce the predictive power of a dataset. There is a process known as dimensionality reduction, which becomes part of the processes in data pre-processing. Dimensionality reduction simplifies the data while maintaining its integrity.

Some examples of techniques used for dimensionality reduction are feature selection (filter, wrapper), feature extraction and Principal Component Analysis (PCA). In conclusion, dimensionality reduction techniques aim at reducing the number of variables under consideration.

6.2.1 Data Augmentation

Although too many variables reduce the prediction power, as discussed above, the performance of machine learning models depends on the quantity and diversity of data. This means that, data scarcity may lead to inaccurate predictions. The data scarcity issue may be due to many reasons, among others including data confidentiality that restricts sharing, lack of data collection/generation devices such as sensors, data collection is difficult/time consuming, and manual data collection by the engineers that is carried out occasionally.

One of the solutions for data scarcity is augmentation. Data augmentation is a process that increases the amount of data. The process takes the existing data and performs some changes/modifications to generate new data. Data augmentation is commonly used in the computer vision including techniques such as scaling, cropping, padding and colour augmentation.