Introduction to the Course: Data Mining

Nischal Regmi

Everest Engineering College

2021

Preamble

- The advent of information technology has made it possible to accumulate ever increasing volume and variety of data from diverse sources. The term 'big data' is presently in fashion.
- A proper analysis of such big data makes it easier for business firms and other institutions to take better decision.
- Because of ever existing competition, business firms have no alternative to accumulate and analyze such big data.
- These requirements led to the development of a methodology to store and process large and diverse data, which we call as data warehousing.

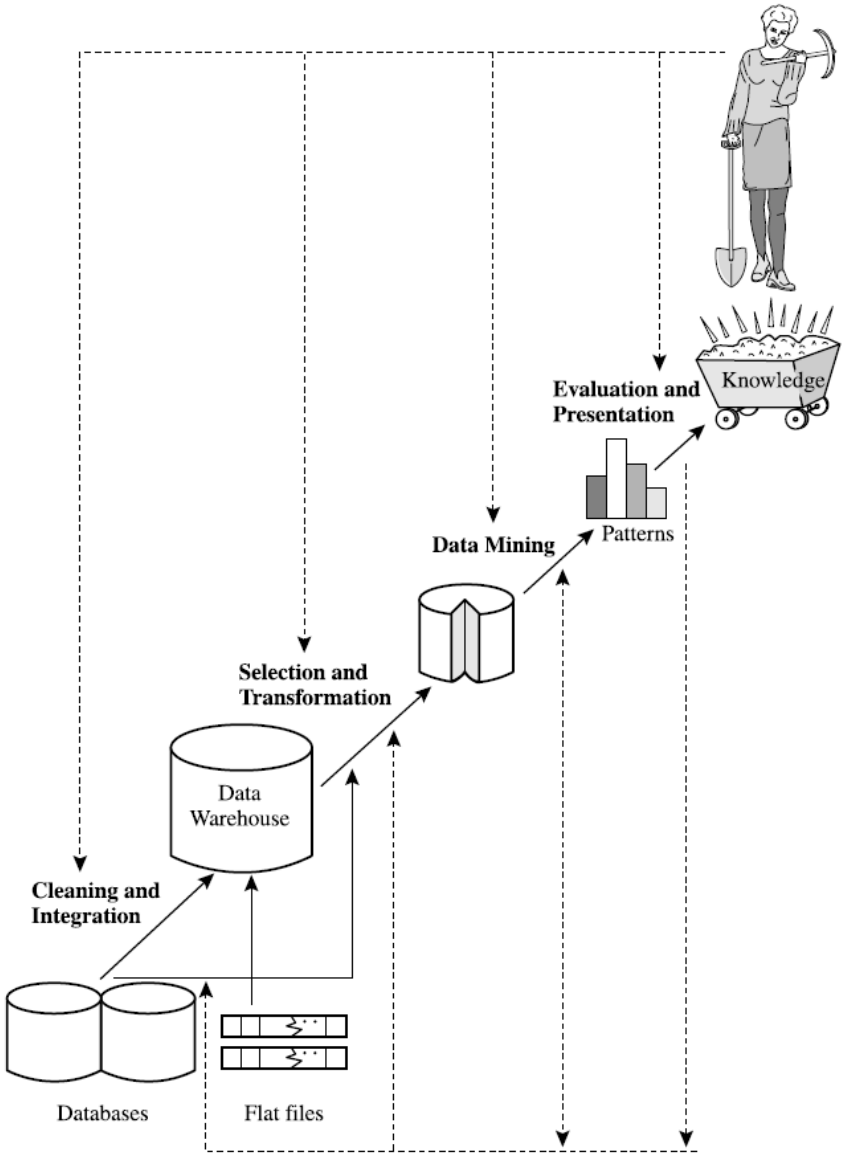

Figure: The *Knowledge Discovery from Data* (KDD) process. Data mining is one step in the process. (fig. source – Han and Kamber, 2006)

Data Mining Example Problem 1: Structured data

From a large data set that consists values of points $X = (x_1, x_2, \ldots, x_n)$, we may be interested in pattern such as

- Which variable is the cause and which is the effect? For example, x_1 could be number of cigarettes consumed per day, x_2 is a variable that measures the level of depression, and x_3 the person's income. The question then is to identify if depression leads to smoking habit, low income leads to depression and so on. This problem is called *causal mining*.
- How can the points be grouped according to their similarity? For example, x_1 could be household income, x_2 the highest level of education of the householders, and x_3 the size of household. The question is then to group the household according to their socioeconomic status. This problem is called *clustering*

Data Mining Example Problem 2: Unstructured Data

Suppose that you are given a large collection of documents (collection of newspaper articles, novels, policy documents, whatever.) We may be interested in the following questions

- What topics or themes the documents talk about? (e.g. in a policy document collection, themes could be 'development', 'poverty alleviation', etc.)
- What influence is the content of one document making on the content of another document in the collection? (e.g. education policies could influence information technology policies)
- What jargon/terminologies do the documents highlight? (e.g. Information technology policy documents of the past highlighted the slogan '*vikaas ko laagi sanchar*')

Data Mining? Data Science? Data Analytics?

- These all are buzzwords with overlapping and contended meanings.
- But Data Mining differs from Artificial Intelligence and Machine Learning
  - AI aims to make *'intelligent' machines*
  - Data Mining aims to extract insights from available data for *human analysts* like scientists, sociologists and managers.
  - Machine Learning is a subject area that studies *algorithms* for extracting relationships in data.
  - AI and Data Mining both use Machine Learning (and other techniques too)
- Data mining is an interdisciplinary subject that uses methods from machine learning, statistics and artificial intelligence

Teaching and Evaluation Plan

Marking Breakdown (for Internal Assessment)

Assignment (with viva,contains practical part too): 40

Essay: 20

Theory exam: 40

Attendance: 0 marks, but 80% mandatory

Total: 100

Lab

- My focus will be more on correct way of applying data mining techniques than on the question of efficient implementation of the algorithms. Regarding the topics, I will focus on the following
  1. Basics
  2. Classification
  3. Regression
  4. Clustering
- I will also provide extra lab exercises the enthusiasts – but submission of the report will not be mandatory for the extra lab exercises.

A Note of Caution

- Because of the glamour of "data science" in the present times, many student select data mining for their elective.
- Though the application of existing data mining algorithm is easy, but I must warn that from the engineering education perspective, data mining could be challenging.
  - Engineers are supposed to be 'problem solvers', not just software application/library users.
  - Solving a new problem requires a thorough understanding of the fundamentals.
- Students should have a strong background in mathematics, mainly statistics, linear algebra, and calculus.
- In addition, a good knowledge of programming; a prior experience of Python or R is desirable but not mandatory.

Reference Books

I will be referring the following as books.

1. Data Mining Concepts and Techniques – J. Han and M Kamber Second Edition, publisher: Morgan Kaufmann, ISBN: 978-1-55860-901-3
2. Introduction to Data Mining – Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, publisher: Pearson

The textbook by Pang-Ning Tan et al. is, in my opinion, more lucid and easy-read for the students. So, most of the examples, notations and figures in my lecture will be from the second book.