"बेटी बचाओ, बेटी पढ़ाओ"

JAYOTI VIDYAPEETH WOMEN'S UNIVERSITY, JAIPUR

(Format for Preparing E Notes)

Faculty of Education & Methodology

Faculty Name- **JV'n Dr. Satish Chandra Pandey** **(Associate Professor)**

Program- MCA-I Semester / Year

Course Name - MCA

Session No. & Name – 1.2 (Process of Data Mining)

Academic Day starts with –

- Greeting with saying **'Namaste'** by joining Hands together following by 2-3 Minutes Happy session, Celebrating birthday of any student of respective class and **National Anthem**.

Lecture Starts with-

Review of previous Session- Introduction of Data Mining

- Topic to be discussed today- Today We will discuss about Process of Data Mining

Data Mining Process:

Data Mining is a process of discovering various models, summaries, and derived values from agiven collection of data.

The general experimental procedure adapted to data-mining problems involves the followingsteps:

1. ***State the problem and formulate the hypothesis***

Most data-based modeling studies are performed in a particular application domain. Hence, domain-specific knowledge and experience are usually necessary in order to come up with a meaningful problem statement. Unfortunately, many application studies tend to focus on the data-mining technique at the expense of a clear problem statement. In this step, a modeler usually specifies a set of variables for the unknown dependency and, if possible, a general form of this dependency as an initial hypothesis. There may be severalhypotheses formulated for a single problem at this stage. The first step requires the combined expertise of an application domain and a data-mining model. In practice, it usually means a close interaction between the data-mining expert and the application expert. In successful data-mining applications, this cooperation does not stop in the initialphase; it continues during the entire data-mining process.

2. ***Collect the data***

This step is concerned with how the data are generated and collected. In general, there aretwo distinct possibilities. The first is when the data-generation process is under thecontrol of an expert (modeler): this approach is known as a designed experiment. The second possibility is when the expert cannot influence the data-generation process: this isknown as the observational approach. An observational setting, namely, random data generation, is assumed in most data-mining applications. Typically, the sampling

distribution is completely unknown after data are collected, or it is partially and implicitlygiven in the data-collection procedure. It is very important, however, to understand how data collection affects its theoretical distribution, since such a priori knowledge can be very useful for modeling and, later, for the final interpretation of results. Also, it is important to make sure that the data used for estimating a model and the data used later for testing and applying a model come from the same, unknown, sampling distribution. Ifthis is not the case, the estimated model cannot be successfully used in a final application of the results.

3. Preprocessing the data

In the observational setting, data are usually "collected" from the existing databses, data warehouses, and data marts. Data preprocessing usually includes at least two common tasks:

1. **Outlier detection (and removal)** – Outliers are unusual data values that are not consistent with most observations. Commonly, outliers result from measurement errors, coding and recording errors, and, sometimes, are natural, abnormal values. Such nonrepresentative samples can seriously affect the model produced later. There are two strategies for dealing with outliers:

a. Detect and eventually remove outliers as a part of the preprocessing phase, or
b. Develop robust modeling methods that are insensitive to outliers.

2. Scaling, encoding, and selecting features – Data preprocessing includes several steps such as variable scaling and different types of encoding. For example, one feature with the range [0, 1] and the other with the range [−100, 1000] will not have the same weights in the applied technique; they will also influence the final data-mining results differently. Therefore, it is recommended to scale them and bring both features to the same weight for further analysis. Also, application-specific encoding methods usually achieve

dimensionality reduction by providing a smaller number of informative features for subsequent data modeling.

These two classes of preprocessing tasks are only illustrative examples of a large spectrum of preprocessing activities in a data-mining process.

Data-preprocessing steps should not be considered completely independent from other data-mining phases. In every iteration of the data-mining process, all activities, together, could define new and improved data sets for subsequent iterations. Generally, a good preprocessing method provides an optimal representation for a data-mining technique by incorporating a priori knowledge in the form of application-specific scaling and encoding.

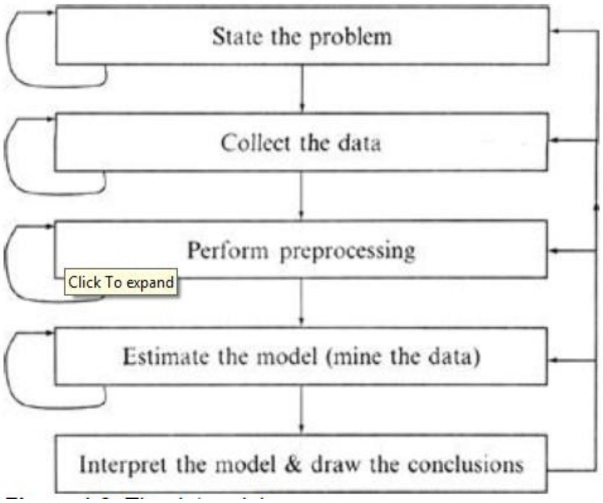
4. *Estimate the model*

The selection and implementation of the appropriate data-mining technique is the main task in this phase. This process is not straightforward; usually, in practice, the implementation is based on several models, and selecting the best one is an additional task. The basic principles of learning and discovery from data are given in Chapter 4 of this book. Later, Chapter 5 through 13 explain and analyze specific techniques that are applied to perform a successful learning process from data and to develop an appropriate model.

5. *Interpret the model and draw conclusions*

In most cases, data-mining models should help in decision making. Hence, such models need to be interpretable in order to be useful because humans are not likely to base their decisions on complex "black-box" models. Note that the goals of accuracy of the model and accuracy of its interpretation are somewhat contradictory. Usually, simple models are more interpretable, but they are also less accurate. Modern data-mining methods are expected to yield highly accurate results using highdimensional models. The problem of interpreting these models, also very important, is considered a separate task, with specific

techniques to validate the results. A user does not want hundreds of pages of numeric results. He does not understand them; he cannot summarize, interpret, and use them for successful decision making.


State the problem
Collect the data
Perform preprocessing
Click To expand
Estimate the model (mine the data)
Interpret the model & draw the conclusions

The Data mining Process

Classification of Data mining Systems:

The data mining system can be classified according to the following criteria:

- Database
- Technology
- Statistics
- Machine
- Learning
- Information
- Science
- Visualization
- Other Disciplines

Some Other Classification Criteria:

- Classification according to kind of databases mined
- Classification according to kind of knowledge mined
- Classification according to kinds of techniques utilized
- Classification according to applications adapted

Classification according to kind of databases mined

We can classify the data mining system according to kind of databases mined. Database system can be classified according to different criteria such as data models, types of data etc. And the data mining system can be classified accordingly. For example if we classify the database according to data model then we may have a relational, transactional, object-relational, or data warehouse mining system.

Classification according to kind of knowledge mined

We can classify the data mining system according to kind of knowledge mined. It is means datamining system are classified on the basis of functionalities such as:

- Characterization
- Discrimination
- Association and Correlation Analysis Classification
- Prediction Clustering
- Outlier Analysis
- Evolution Analysis

Classification according to kinds of techniques utilized

We can classify the data mining system according to kind of techniques used. We can describesthese techniques according to degree of user interaction involved or the methods of analysis employed.

Classification according to applications adapted

We can classify the data mining system according to application adapted. These applications areas follows:

- Finance
- Telecommunications
- DNA
- Stock Markets

- E-mail

University Library Reference-

- Data Mining by A K Pujari
- Academic Day ends with- National song**' Vande Mataram'**