Data Mining and Business Intelligence

ITA5007

PROF. E.P.EPHZIBAH

Topic

STEPS IN DATA MINING

Steps in data mining

Some of the most serious errors in data analysis result from a poor understanding of the problem—an understanding that must be developed before we get into the details of the algorithms to be used.

Here is a list of steps to be taken in a typical data mining effort:

Steps in Data Mining (contd..)

1. Develop an understanding of the purpose of the data mining project- select a project that would be beneficial to the society
   - Women
   - Children
   - Poor people
   - Farmers
   - Physically challenged
   - Mentally challenged
   - Students (School or College)

Steps in Data Mining (contd..)

2. Obtain the dataset to be used in the analysis:

This often involves random sampling from a large database to capture records to be used in an analysis. While data mining deals with very large databases, usually the analysis to be done requires only thousands or tens of thousands of records.

3. Explore, clean, and preprocess the data:

This involves verifying that the data are in reasonable condition. How should missing data be handled? Are the values in a reasonable range, given what you would expect for each variable? Are there obvious outliers? We also need to ensure consistency in the definitions of fields, units of measurement, time periods, and so on.

Steps in data mining (contd..)

4. Reduce the data, if necessary, and separate them into training, validation, and test datasets:

This can involve operations such as eliminating unneeded variables, transforming variables (e.g., turning “money spent” into “spent > $100” vs. “spent ≤ $100”), and creating new variables. Make sure that you know what each variable means and whether it is sensible to include it in the model.

5. Determine the data mining task (classification, prediction, clustering, etc.)

Steps in data mining (contd..)

6. Choose the data mining techniques to be used (regression, neural nets, hierarchical clustering, etc.).

7. Use algorithms to perform the task:

This is typically an iterative process—trying multiple variants, and often using multiple variants of the same algorithm (choosing different variables or settings within the algorithm). Where appropriate, feedback from the algorithm's performance on validation data is used to refine the settings.

Steps in data mining (contd..)

8. Interpret the results of the algorithms:

This involves making a choice as to the best algorithm to deploy, and where possible, testing the final choice on the test data to get an idea as to how well it will perform.

9. Deploy the model.

This involves integrating the model into operational systems and running it on real records to produce decisions or actions.

The steps in SEMMA, a methodology developed by SAS (Software and Services):

Sample	Take a sample from the dataset; partition into training, validation, and test datasets.
Explore	Examine the dataset statistically and graphically.
Modify	Transform the variables and impute missing values.
Model	Fit predictive models (e.g., regression tree, collaborative filtering).
Assess	Compare models using a validation dataset.