Data Mining notes

Marco Natali

INDICE

ELENCO DELLE FIGURE

1 INTRODUCTION TO DATA MINING

In the course of Data Mining are introduced and analyzed methods and models to use to analyze large amount of data, so we start giving the definition of Data Mining as

Def. Data Mining is the use of efficient techniques for the analysis of very large collections of data and the extraction of useful and possibly unexpected patterns in data (hidden knowledge)

Enormous data growth in both commercial and scientific databases, due to advances in data generation and collection technologies but also there is also a mantra that says to gather whatever data you can whenever and wherever possible.

Lots of data is being collected and warehoused, like for example Web data, where Yahoo has Peta Bytes of web data or also Facebook has billions of active users, also Amazon handles millions of visits each day and this also explains why know how to process and find useful information from huge amount of data will be very important.

Data are divided in two useful categories:

PRIMARY DATA: original data that has been collected for a specific purpose and they are not altered by humans.

SECONDARY DATA: data that has been already collected and made available for other purposes and may be obtained from many sources.

The process of Knowledge discovered in database (KDD) can be described by figure 1, where we have input data that will have a preprocessing phase and then we applied to a data mining modelation and we obtain the information gained.

We have so discovered that Data Mining is an phase of KDD process and we descrived all phases that can be viewed on 2, that will all analyzed during the course:

DATA INTEGRATION: involves the process of data understanding, data cleaning, merging data coming from multiple sources and transfoming them to load them into a Data Warehouse Databases.

DATA WAREHOUSE: is a database targeted to answer specific business questions.

DATA SELECTION: relevant data to analysis tasks are retrieved from data.

DATA TRANSFORMATION: transform data into appropriate form for mining (summary, aggregation, etc.)

Figura 1: Phases on a Data Mining approach

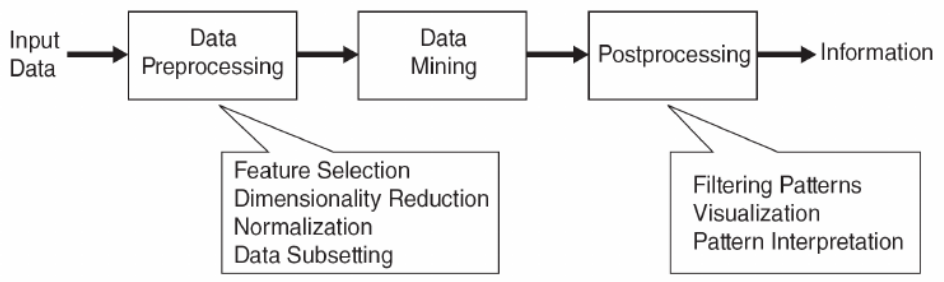
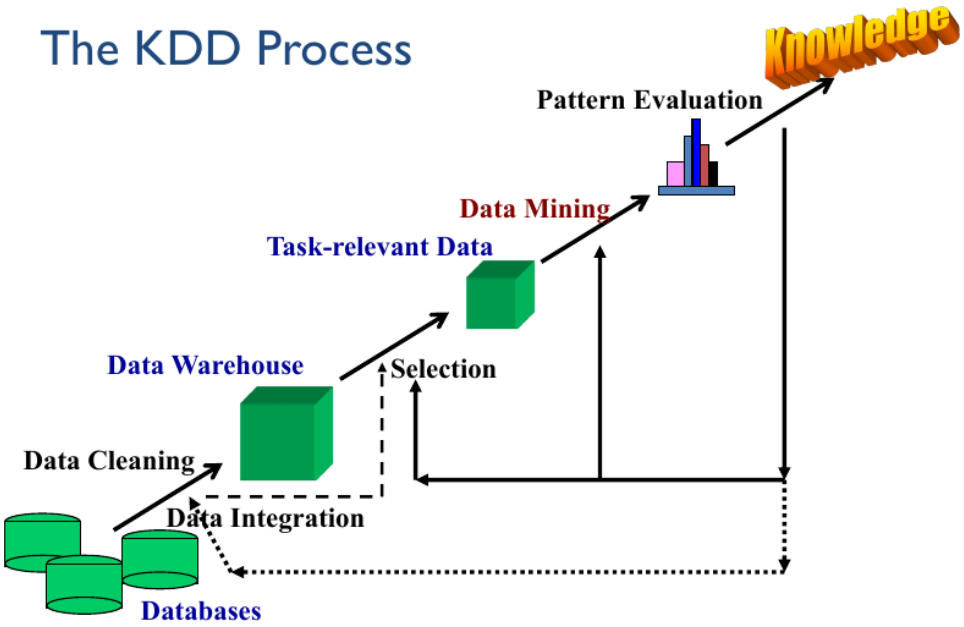


Figura 2: KDD process phases



PATTERN EVALUATION: Identify truly interesting patterns

KNOWLEDGE REPRESENTATION: use visualization and knowledge representation tools to present the mined data to the user

Data Mining approches can be divided in two different tasks:

PREDICTION METHODS: we use some variables to predict unknown or future values of other variables.

DESCRIPTION METHODS: the purpose is to find human-interpretable patterns that describe the data.

We now descrive the 4 modellation task that are developed and used on Data Mining:

CLASSIFICATION AND REGRESSION: refers to the task of building a model for the target variable as a function of explanatory variable and can be a *classification* (model for discrete class attribute) or a *regression* where we continuious data where we would like to predict as a function of the values of other attributes.

CLUSTERING: finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

ASSOCIATION RULES: Given a set of records each of which contain some number of items from a given collection, produce dependency rules which will predict occurrence of an item based on occurrences of other items.
It is used in Market-basket analysis to optimize sales promotion and also in medical informatics to find combination of patient symptoms and test results associated with certain diseases.

ANOMALY DETECTION: detect significant deviations from normal behavior and it can be used in Credit Card fraud detection or in detect network intrusions.

Traditional techniques have often encountered practical difficulties in meeting the challenges posed by big data application in particular in scalability of peta/etabytes of data, in high dimension of attributes, in heterogeneous and complex data and

also in ownership of data so this is why was introduced and also used data mining approach.