1. Introduction

Francis Bacon said that ”Knowledge is power”. Is it so? If it is, where is the power in knowledge? Power is the ability to control, or at least to influence, events. Control implies taking an action that produces a known result. Thus, the power in knowledge is in knowing what to do in order to get what we want – knowing which actions yield which results and how and when to take them. Where does knowledge come from?

We need observations, measurements, data from the surrounding world.

To get knowledge, information and understanding, we make data mining.

Data mining is the analysis of often large observational data sets to find (unsuspected) relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

Data mining involves several different algorithms to accomplish different tasks. All such algorithms attempt to fit a *model* to the data. They examine the data and determine a model that is the closest to the charcateristics of the data being reviewing. Data mining algorithms can be seen to consist of three parts:

- *Model*: The purpose of the algorithm is to fit a model to the data.
- Preference: Some criteria must be used to fit one model over another.
- *Search*: All algorithms require some technique to search the data.

Let us look at an example.

Example

The data of a credit card company are modeled as divided into four classes: authorize, ask for further identification before authorization, do not authorize and do not authorize but contact police.

The data mining tasks are twofold. The historical data have to be used to determine how the data fit into the four classes. Then the problem is to apply this model to each new purchase.

Here the search requires examining past data about credit card purchases and their outcome to determine what criteria should be used.

The preference will be given to criteria that seem to fit data best. For instance, we do not authorize the use of a credit card reported to be stolen.

As seen from Fig. 1.1, the model created can be either *predictive* or *descriptive* in nature. Under these some of the most common data mining tasks are shown using the model type mentioned.

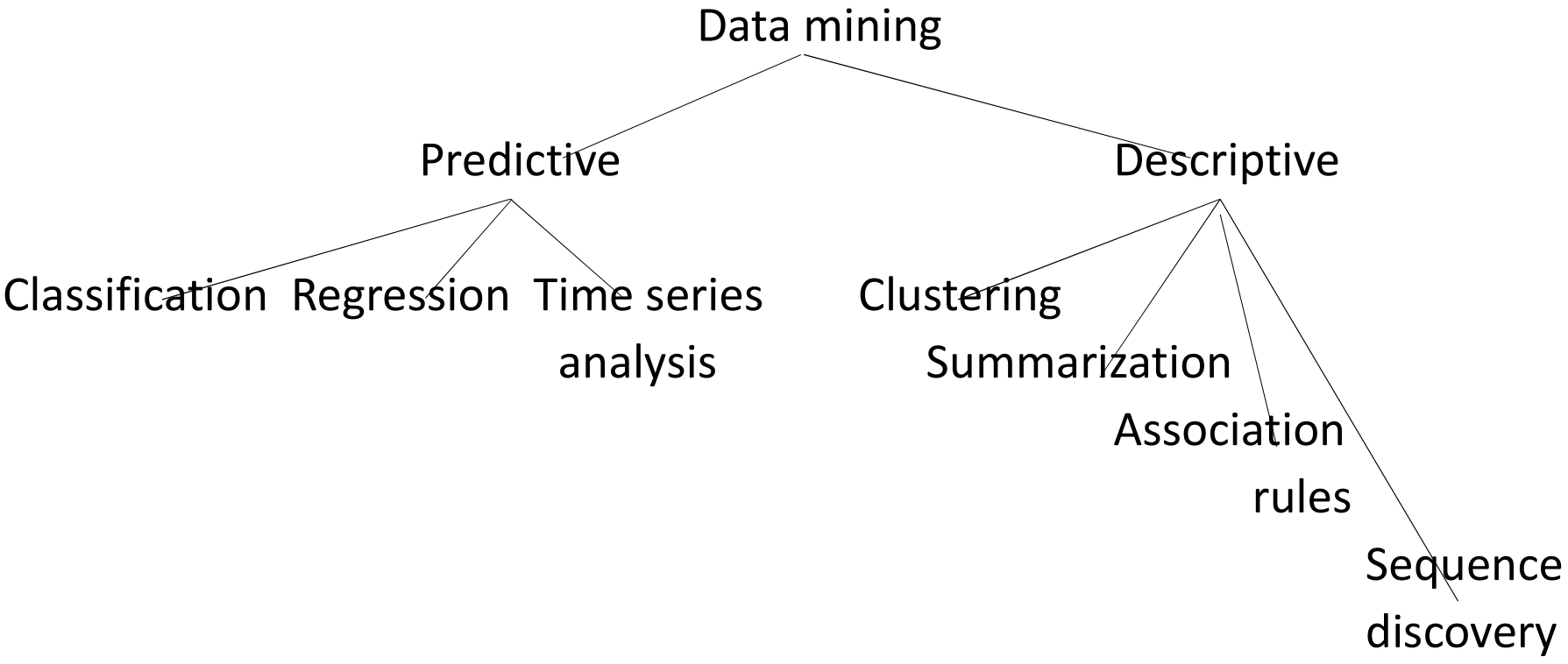


Fig. 1.1. Data mining models and tasks (some of them to be considered later in detail).

1.1 Basic data mining tasks
Classification

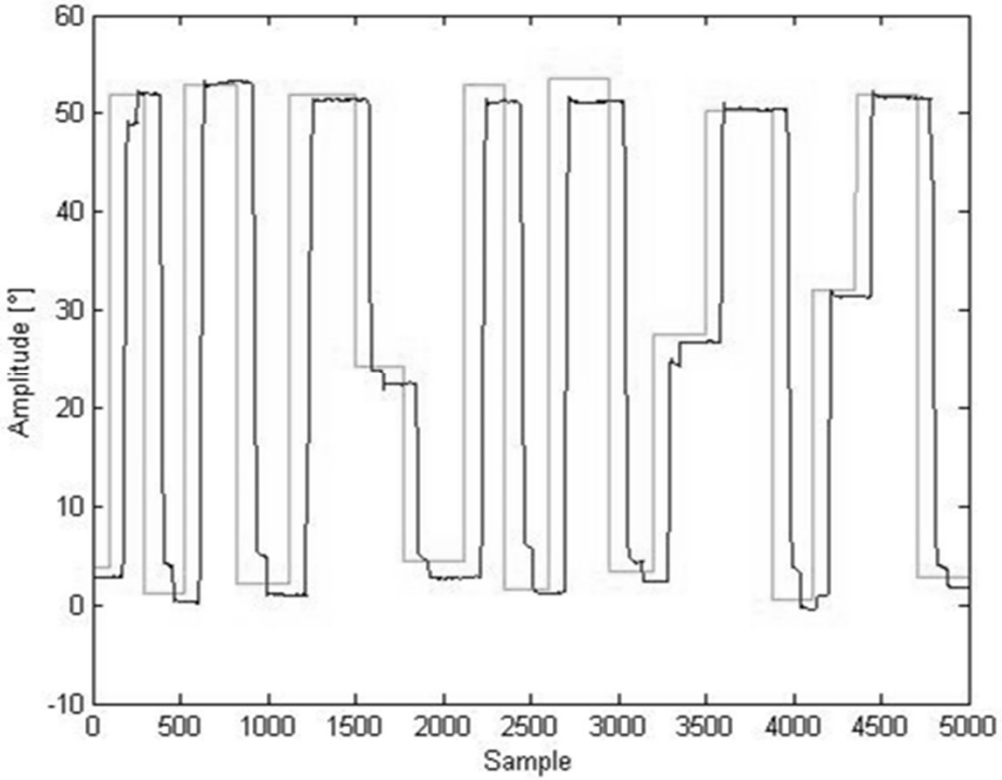
Classification maps data into predefined groups called classes. It is typically called supervised learning because the classes are known or determined before mining the data. For instance, a physician has to diagnose whether a subject is sick or not, or which disease one has from some given alternatives (being possible within a certain medical specialty and with known symptoms and other information).

Regression

Regression is to use a data item to map to real valued prediction variable, whereas classification concerned discrete values, classes. Regression involves the learning of the function that does this mapping. It assumes that the target data fit into some known type of the function, e.g., linear or logistic, and determines the best function of this type that models the data given. Some error analysis is employed to determine which is ”best”. For example, regression analysis could be tried to the data in Fig. 1.2.

Using regression one might predict how a signal would continue after the data given.

Fig. 1.2. A (blue) saccadic eye movement signal sampled at 250 Hz where the saccades of a subject followed the virtually noiseless (green) stimulation signal, i.e., the subject followed the stimulation by the gaze. The down direction on the vertical axis corresponds to an eye movement to the left, and the up direction to the right. The length of the signal was 20 s. The stimulation was a coherent light dot jumping from the left to the right or vice versa after varying intervals.



Time series analysis

Fig. 1.2 also presents a data item for *time series analysis* or *signal analysis*.

It is examined how some phenomenon varies over time. It is not necessary to look at samples measured with a constant interval as in Fig. 1.2. This example could consider various intervals such as minutes, hours, days, weeks, month, years etc. In fact, we repeated these eye movement test series with such intervals for biometrics purpose to verify or identify individuals[1].

First, the structure of line or curve is examined to determine its behavior. Second, distance function can be used to determine the similarity between time series. Third, historical data can be used to predict future values.

[1] Y. Zhang J. Laurikkala and M. Juhola, Biometric verification with eye movements: results from a long-term recording series, IET Biometrics, 4(3), 162-168, 2015

Prediction

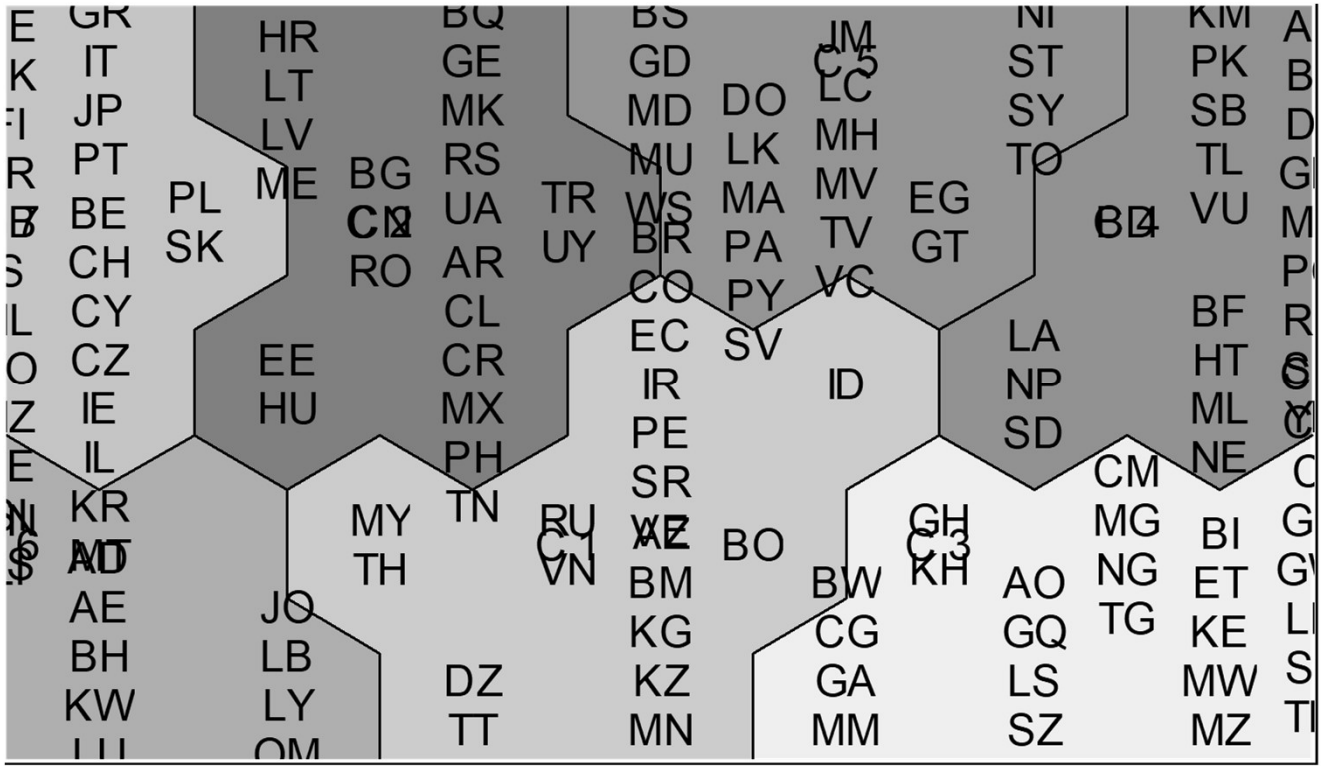
Frequently, real-world data mining applications are seen as predicting future data states on the basis of past and current data. *Prediction* is viewed as a type of classification. The difference is that prediction deals with a future state rather than a current state. Weather forecast is an example. Prediction is not always doing something with time series, but other approaches are used as well.

Clustering

Clustering is similar to classification except that the groups called clusters are not predefined. It is also called *unsupervised learning* or sometimes segmentation, while classification is of *supervised learning or training*. Clustering is thought of as partitioning or segmenting the data into clusters being or not being disjointed. Clustering is typically executed by determining the similarity among the data of predefined variables (attributes or features). The most similar data cases are grouped into clusters.

In Fig. 1.3 there is an example of countries clustered with Kohonen self-organizing map (SOM) according to a set of variables.

Fig. 1.3. Using 61 demographic, social and economic variables, and specific variable, annual homicide rate (per 100 000 inhabitants), 181 countries were grouped to 7 clusters with SOM[2]. Variables were, e.g., adult literacy rate, employment of population (≥25 years) ratio, expected years of schooling of children, GDP per capita $ and Internet users %. For instance, FI is in C7 where there are many developed wealthy countries. Data source was from UN Development Program. As to variable values, similar countries are close to each other.



[2] X Li, H. Joutsijoki, J. Laurikkala M. Siermala and M. Juhola, Homicide and its social context: Analysis using the Self-Organizing Map, Applied Artificial Intelligence, 29,382-401, 2015.

Summarization, association rules and sequence discovery

Summarization or characterization maps data into subsets associated with simple descriptions. It extracts or derives representative information about a data set. Often numeric results such as means, medians and standard deviations are computed for the variables of the data.

Association rules can be understood as IF-THEN-ELSE expressions. They are models that identify specific types of data associations. For example, let us imagine that if 60% of time of custom visits cucumbers are sold, then 70% of time tomatoes are also sold in a supermarket. That is why, they are located next to each other.

Sequence discovery or analysis is used to determine sequential patterns in data. For example, 60% of the users of page A of XYZ Corp. follow page sequences of {A,B,C}, {A,D,B,C} or {A,E,B,C}. The webmaster could add a link directly from A to C.

[Note! The two latter are not considered in the current course, but in others.]

1.2 Approaches and tasks for data mining processes

To explore and model the data set, we build a framework for data mining so that appropriate tools are applied to appropriate data that is properly prepared to solve key problems and deliver solutions needed. Ten golden rules are given as follows:

1. Select clearly defined problems that will yield tangible benefits.
2. Specify the required solution.
3. Define how the solution delivered is going to be used.
4. Understand as much as possible about the problem and the data set (domain).

5. Let the problem drive modeling, i.e., tool selection, data preparation and other actions.
6. Stipulate assumptions.
7. Refine the model iteratively.
8. Make a model as simple as possible - but no simpler.
9. Define instability in the model (critical areas where change in output is drastically different for a small change in inputs).
10. Define uncertainty in the model (critical areas and ranges in the data set where the model produces low confidence predictions or insights).

Rules 1-3 are the first three stages of the data exploration process. Rule 4 captures the insight that if one knows what is doing, success is more likely. Rule 5 advises to find the best means for the process, not just a job one can do with the tool. Rule 6 advises not just to assume, but rather to tell someone. Rule 7 says to keep trying different things until the model seems as good as it is going to get. Rule 8 means KISS (Keep it Sufficiently Simple). Rules 9 and 10 mean state what works, what does not, and where one is not sure.

Phases of data mining

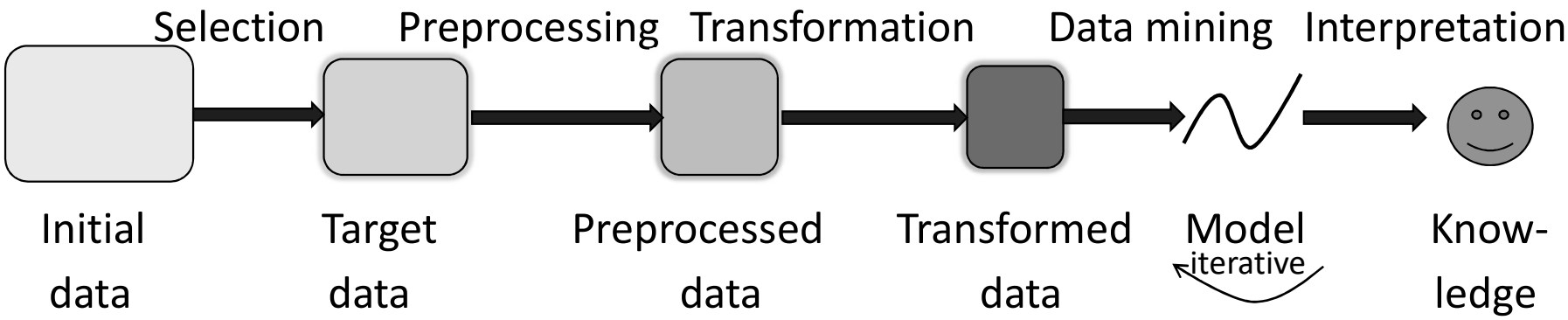


Fig. 1.4. Data are selected, perhaps from several sources. Variables have to be selected and their erroneous values corrected or removed and missing values perhaps imputed (substituted). In transformation, if necessary, the data are transformed to a more usable form. Modeling then follows and, finally, the interpretation or evaluation of results is made.

Visualization

Vizualization refers to the visual presentation of data and results.

Graphical: Graph structures include bar charts, pie charts, histograms, lines and curves.

Geometric: These include box plot and scatter diagrams.

Pixel-based: Different colors can be used.

Icon-based: There are figures, colors and other forms.

Hierarchical: Regions are divided based on data values.

Hybrid: The preceding ways can be combined.

Historical perspective of data mining

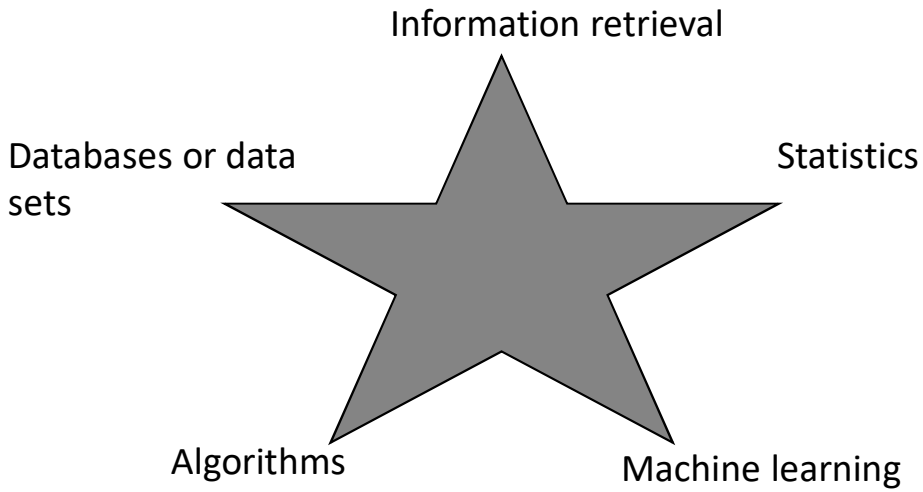


Fig. 1.5. Data mining uses many statistical tools originated even from the late 1700s to the 1970s (*k*-means clustering). Databases since the 1960s are needed to store and retrieve data. Information retrieval usually concerns digital documents. Algorithms and machine learning from artificial intelligence and pattern recognition are the ”core” of data mining computation.

1.3 Data mining issues

There are some crucial implementation issues associated with data mining:

Human interactions: Since data mining problems are often not precisely stated, both application domain and data mining experts are needed. Training data and results desired are defined. Interpretation of results is important to do carefully.

Overfitting, overtraining or overlearning: Overfitting occurs when a model is built to be too detailed or strictly fit the data given. Thus, it may lose its generalization ability and is not valid for future data.

Outliers: There are sometimes many data entries that do not fit nicely into the derived model. They may be erroneous values or otherwise exceptional that are best to remove. For instance, the age of 0 year for patient data is such.

Interpretation of results: This may require experts to correctly interpret the results obtained.

Visualization: To easily view and understand input data and results visualization is helpful.

Large data sets: Data sets may be massive which creates problems to handle such. Sampling and parallelization are effective to attack these problems.

High dimensionality: There are often great numbers of variables (columns in a table or matrix) and data cases or items (rows). Sometimes some variables may even interfere with the correct completetion of a data mining task. It is important to find such and remove. A great number of variables can make the complexity of a problem too huge. This *dimensionality curse* may require *dimensionality reduce.* We have to select which variables are left out.

Multimedia data: Usually data mining methods are targeted to traditional data types, i.e., numeric, characters and text. They are not always suitable for multimedia, e.g., geographic data (GIS).

Missing data: There may be missing variable values, incomplete data. Some algorithms require complete data. That is why, missing values have to be estimated or variables with very frequent missing values perhaps to be removed.

Irrelevant data: Some variables may be useless. If all values of a variable are constant, it is called *dead* and can be removed. If almost all values are constant, it is not straightforward whether it can be removed. Those very rare values could be essential in some situations.

Noisy data: Some values might be invalid or incorrect. A user or a measuring equipment has given a false value. These are corrected or deleted, but first they have to be found.

Changing data: Data cannot be assumed to be static even if mostly we start from this thought. Therefore, algorithms must be rerun from time to time.

Integration: It is good if data mining proceses can be integrated to the whole of other data processing.

Application: Determining the intended use for the information obtained from the data mining function is a challenge: how domain experts are able to effectively use the results of data mining.

The preceding 14 issues should be adressed by data mining algorithms and processes.

To say it simple, a data mining process typically consists of the following phases:

1. State the problem.
2. Choose the tool.
3. Get some data.
4. Make a model.
5. Apply the model.
6. Evaluate the results.