# Common sources of erroneous, inaccurate or misleading results in data mining

Wilhelmiina Hämäläinen

November 28, 2023

Figure 1 presents the knowledge discovery process with an additional step 0 (gathering data). Step 0 is usually not included to the process, since the data miner is typically invited when the data has been gathered. However, many problems could be avoided if the data gathering had been done carefully.

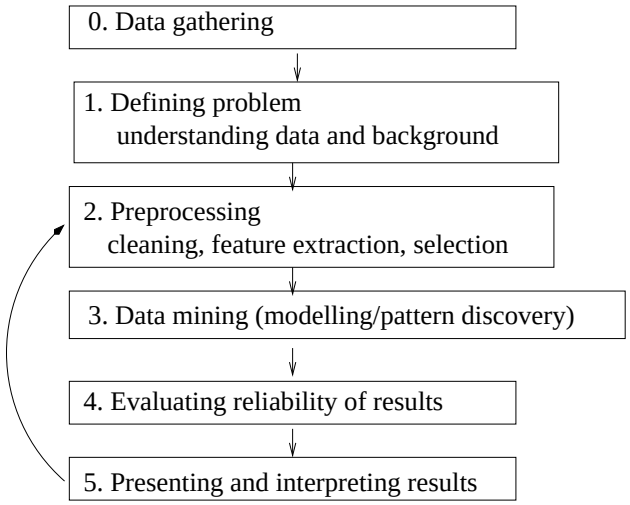


Figure 1: Steps of the KDD process extended with step 0, gathering data.

Here are important problem sources at each phase:

0. Gathering data: This step is the ultimately the main source of errors and inaccuracies in data, stemming from, e.g., faulty or inaccurate devices, human respondents giving incorrect/missing information for private reasons, and human errors in manual recording of data. The data

may also be too sparse or not representative to the true population, which leads to many problems – the model parameters cannot be estimated accurately, the discoveries will lack statistical validity (may not hold in future data) and the results may be misleading (describe some exceptional subpopulation instead of the assumed target population).

1. Defining the problem: The computational problem (and related assumptions) may be misspecified, because the data miner and client do not understand each other (people in different fields have their own terminology and may even use the same term for different meanings). In the worst case, the data miner solves a wrong problem, e.g., searches frequent co-occurrence associations when the client wanted statistical associations. Another problem is misspecification of **data types**, which can lead to serious errors at later steps (remember: everything that looks numbers is not ratio-scale numerical data – even categorical values can be coded by numbers and there are numerical datatypes, like circular variables, that require special treatment).

2. Preprocessing: Real world data contains nearly always some anomalies and data cleaning is seldom perfect, leaving possibly erroneous values or incorrectly imputed missing values. Outliers can bias results seriously, depending on the modelling technique (e.g., Pearson correlation coefficients are inaccurate, $K$-means cannot detect real clusters), unless detected and dealt appropriately.

   Feature extraction and selection have a crucial effect on results, leading to e.g., missed and trivial patterns or suboptimal clusterings. One common error is to ignore the data types and apply PCA or SVD dimension reduction on non-numerical or circular features. Similar problems occur if the curse of dimensionality is ignored, because it tends to produce misleading similarity and distance evaluations and emphasize irrelevant (possibly erroneous) features.

3. Data mining: At this step, the main source of errors is ignorance of methods and their assumptions. One may choose a method that doesn't solve the specified problem (e.g., search only condensed presentations of frequent associations instead of significant statistical associations) or whose assumptions don't fit the data (e.g., certain shape of clusters, absence of outliers). Many errors stem from programs and they should be tested carefully and one should also prepare for exceptions (like data formatting errors). One common source of errors is to use ready made library functions without checking what they actually calculate or how

the parameters should be set. One should also remember that existing programs do contain bugs or may not work correctly in pathological cases. Here one could also mention accuracy problems related to presentation of floating point numbers – small inaccuracies can accumulate and produce seriously erroneous results.

4. Evaluating reliability of results: Validation of results is extremely important, because the methods tend to return something even from random data. One would need some guarantees that the patterns are not due to chance, but likely to hold also in future data. One problem is that the validation method itself may be wrongly selected or have unrealistic assumption (e.g., on the distribution of data). Evaluating quality of clustering is especially difficult, because the validation method may assume a different clustering objective than intended. E.g., many famous internal validation indices cannot detect good spectral clustering, because the clusters are arbitrarily shaped.

5. Presenting and interpreting results: The results can be misinterpreted, because of misleading presentation. Visualizations are often inaccurate or even misleading (e.g., 2-D presentations of multidimensional data may not show real clustering). One should also supply all information that is needed for interpretation (e.g., obtaining excellent clustering index value is misleading, if all data points formed singleton clusters). One should also understand what conclusions can be drawn from models and patterns. A classical error is to assume that statistical dependence means causality, sometimes with serious consequences (vs. Simpson's paradox).