Multi-Level Approaches for Data Mining and Data Processing: Two Examples

Multi-level approaches become widely applicable to complex tasks of data analysis, wherein complexity may refer to what we try to learn from data, to technical details that prevent users from understanding data mining algorithms, or to data volumes that are too large to keep using standard algorithms. Levels (or layers) can be understood in various ways, as in artificial neural networks, ensembles of classifiers, granular computing, layered learning, et cetera. In this talk, we attempt to categorize the meanings of levels in data mining and processing approaches. Then we focus on two examples: 1) execution of basic database operations that stop being basic for terabytes of data, and 2) extraction of data dependencies in a way that is clear to the users and, in the same time, remains useful while constructing classifiers.

The first example relates to the database technology developed by Infobright, where SQL queries can be executed over large data volumes stored on a standard machine, and with a need of neither advanced database tuning nor administration. Performance is achieved using a two-level model of data storage and processing, wherein, at the higher layer, we operate with rough rows, each corresponding to a set of 2^16 of original rows. Rough rows are automatically labelled with compact information related to the values of corresponding rows. This way, we create a new information system, where objects correspond to rough rows and attributes – to various types of compact information. Data operations are supported at such a rough level, with an access to original rows still possible whenever compact information turns out insufficient to continue query execution. We implement a number of algorithms that use compact information to minimize and optimize an access to original data stored in a compressed form on disk.

The second example relates to a two-level methodology for extracting multi-attribute dependencies approximately holding in data, wherein the higher layer corresponds to their intuitive representation, while the lower layer hides away the details of how the degrees of their satisfaction in data are actually computed. Extraction of multi-attribute dependencies is an important phase in data mining, e.g., in feature selection, classification, or construction of mechanisms for reasoning about data. On the other hand, the users usually want to interpret such dependencies without a need of understanding in what mathematical sense and to what specific degree they hold in data. In this talk, we outline a framework for representing, extracting and reasoning about multi-attribute dependencies in a way that is the same in case of expressing their degrees of satisfaction in data using, e.g., statistical estimates, information measures, rough sets, fuzzy sets, etc., referring to some common mathematical properties of all those approaches. We believe that the proposed framework is both convenient for the users and efficient in adjusting lower-level technical details for particular data types and particular goals of data analysis.