Knowledge Discovery - A Control Theory Perspective

Bjarne A. Foss

*Department of Engineering Cybernetics
Norwegian University of Science and Technology
N-7034 Trondheim, NORWAY
Fax: + 47 73 59 43 99 E-mail: Bjarne.Foss@itk.ntnu.no*

May 5, 1997

Abstract

Knowledge discovery in databases is a label for an activity performed in a wide variety of application domains within the science and business communities as well as for pleasure, see Fayyad, Haussler & Stolorz (1996) and Imielinski & Mannila (1996). The activity uses a large and heterogeneous data-set as a basis for synthesizing new and relevant knowledge. The knowledge is new because hidden relationships within the data are explicated and/or data is combined with prior knowledge to elucidate a given problem. The term relevant is used to emphasize that knowledge discovery is a goal-driven process in which knowledge is constructed to facilitate the solution of a problem.

Knowledge discovery may be viewed as a process containing many tasks as discussed in Fayyad, Piatetsky-Shapiro & Smyth (1996). Some of these tasks are well understood while others depend on human judgement in an implicit matter. Further, the process is characterized by heavy iterations between the tasks. This is very similar to many creative engineering processes, eg. the development of dynamic models as discussed in the industrial field study by Foss, Lohmann & Marquardt (1997). In this reference mechanistic, or first principles based, models are emphasized and the tasks involved in model development are defined by:

1. *Initial data collection and problem formulation.* The initial data are collected and some more or less precise formulation of the modeling problem is developed.
2. *Tools selection.* The software tools to support modeling and allow simulation are selected.
3. *Conceptual modeling.* The system to be modeled e.g. a chemical reactor, a power generator, or a marine vessel, is abstracted at first. The essential compartments and the dominant phenomena occurring are identified and documented for later reuse.
4. *Model representation.* A representation of the system model is generated. Often, equations are used, however, a graphical block diagram (or any other formalism) may alternatively be used depending on the modeling tools selected above.

5. *Implementation.* The model representation is implemented using the means provided by the modeling system of the software employed. These may range from general programming languages, to equation-based modeling languages or graphical block-oriented interfaces.
6. *Verification.* The model implementation is verified to really capture the intent of the modeler. No simulations for the actual problem to be solved are carried out for this purpose.
7. *Initialization.* Reasonable initial values are provided or computed, the numerical solution process is debugged.
8. *Validation.* The results of the simulation are validated against some reference, ideally against experimental data.
9. *Documentation.* The modeling process, the model, and the simulation results during validation and application of the model are documented.
10. *Model application.* The model is used in some model-based process engineering problem solving task.

For other model types like neural network models where data-driven knowledge is utilized, the modeling process will be somewhat different. Some of the tasks, like the conceptual modeling phase, will vanish. In this paper both black-box (or data-driven) models (Ljung 1987), (Ljung 1991), and mechanistic models (Marquardt 1996) will be discussed.

Typical application areas for dynamic models are control, prediction, planning and fault detection and diagnosis. A major deficiency of today's methods is the lack of ability to utilize a wide variety of knowledge. As an example a black-box model structure has very limited abilities to utilize first principles knowledge on a problem. This has provided a basis for developing different hybrid schemes. Two hybrid schemes will highlight the discussion. First, it will be shown how a mechanistic model can be combined with a black-box model to represent a pH neutralization system efficiently (Johansen & Foss 1993). Second, the combination of continuous and discrete control inputs is considered utilizing a two-tank example as case. Different approaches to handle this heterogeneous case are considered (Slupphaug, Vada & Foss 1997).

The hybrid approach may be viewed as a means to integrate different types of knowledge, i.e. being able to utilize a heterogeneous knowledge base to derive a model. Standard practice today is that methods and software can treat large homogeneous data-sets. A typical example of a homogeneous data-set is time-series data from some system, e.g. temperature, pressure and compositions measurements over some time frame provided by the instrumentation and control system of a chemical reactor. If textual information of a qualitative nature is provided by plant personnel the data becomes heterogeneous.

The above discussion will form the basis for analyzing the interaction between knowledge discovery, and modeling and identification of dynamic models. In particular we will be interested in identifying how concepts from knowledge discovery can enrich state-of-the-art within control, prediction, planning and fault detection and diagnosis of dynamic systems.

References

- Fayyad, U., Haussler, D. & Stolorz, P. (1996). Mining scientific data, *Comm. of the ACM* **39**(1): 51–57.
- Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases, *AI Magazine* **17**(3): 37–54.
- Foss, B. A., Lohmann, B. & Marquardt, W. (1997). A field study of the industrial modeling process, *Proc. ADCHEM'97, Banff, Canada*. Keynote lecture.
- Imielinski, T. & Mannila, H. (1996). A database perspective on knowledge discovery, *Comm. of the ACM* **39**(11): 58–64.
- Johansen, T. A. & Foss, B. A. (1993). State-space modeling using operating regime decomposition and local models, *Preprints 12th IFAC World Congress, Sydney, Australia*, Vol. 1, pp. 431–434.
- Ljung, L. (1987). *System Identification: Theory for the User*, Prentice-Hall, Inc., Englewood Cliffs, NJ.
- Ljung, L. (1991). Issues in system identification, *IEEE Control Systems Magazine* **11**(1): 25–29.
- Marquardt, W. (1996). Trends in computer-aided process modeling, *Comp. Chem. Engnn.* **20**: 591–609.
- Slupphaug, O., Vada, J. & Foss, B. A. (1997). Mpc with mixed-discrete control inputs, *Proc. American Control Conference, Albuquerque, NM*. Accepted.