Model Identification through Data Mining

Diego Liberati
Consiglio Nazionale delle Ricerche, Italy

INTRODUCTION

In many fields of research as in everyday life, one has to face a huge amount of data, often not completely homogeneous and many times without an immediate grasp of an underlying simple structure. A typical example is the growing field of bio-informatics, where new technologies, like the so-called micro-arrays, provide thousands of gene expression data on a single cell in a simple and quickly integrated way. Further, the everyday consumer is involved in a process not so different from a logical point of view, when the data associated with the consumer's identity contributes to a large database of many customers whose underlying consumer trends are of interest to the distribution market.

The large number of variables (i.e., gene expressions, goods) for so many records (i.e., patients, customers) usually are collected with the help of wrapping or warehousing approaches in order to mediate among different repositories, as described elsewhere in the encyclopedia.

Then, the problem arises of reconstructing a synthetic mathematical model, capturing the most important relations between variables. This will be the focus of the present contribution.

A possible approach to blindly building a simple linear approximating model is to resort to piece-wise affine (PWA) identification (Ferrari-Trecate et al., 2003). The rationale for this model will be explored in the second part of the methodological section of this article.

In order to reduce the dimensionality of the problem, thus simplifying both the computation and the subsequent understanding of the solution, the critical problems of selecting the most salient variables must be solved. A possible approach is to resort to a rule induction method, like the one described in Muselli and Liberati (2002) and recalled in the first methodological part of this contribution. Such a strategy offers the advantage of extracting underlying rules and implying conjunctions and/or disjunctions between such salient variables. The first idea of their non-linear relationships is provided as a first step to designing a representative model,using the selected variables.

The joint use of the two approaches recalled in this article, starting from data without known background information about their relationships, first allows a reduction in dimensionality without significant loss in information, and later to infer logical relationships. Finally, it allows the identification of a simple input-output model of the involved process that also could be used for controlling purposes.

BACKGROUND

The two tasks of selecting salient variables and identifying their relationships from data may be sequentially accomplished with various degrees of success in a variety of ways. Principal components order the variables from the most salient to the least, but only under a linear framework. Partial least squares allow an extension to non-linear models, provided that one has prior information on the structure of the involved non-linearity; in fact, the regression equation needs to be written before identifying its parameters. Clustering may operate in an unsupervised way without the a priori correct classification of a training set (Booley, 1998). Neural networks are known to use embedded rules with the indirect possibility (Taha & Ghosh, 1999)of making rules explicit or to underline the salient variables. Decision trees (Quinlan, 1994), a popular framework, can provide a satisfactory answer to both questions.

MAIN THRUST

Recently, a different approach has been suggested—Hamming clustering. This approach is related to the classical theory exploited in minimizing the size of electronic circuits, with the additional ability to obtain a final function able to generalize everything from the training dataset to the most likely framework describing the actual properties of the data. In fact, the Hamming metric tends to cluster samples with code that is less distant. This is likely to be natural if variables are redundantly coded via thermometer (for numeric variables) or used for only-one (for logical variables) code (Muselli & Liberati, 2000). The Hamming clustering approach reflects the following remarkable properties:

- It is fast, exploiting (after the mentioned binary coding) just logical operations instead of floating point multiplications.

- It directly provides a logical understandable expression (Muselli & Liberati, 2002), which is the final synthesized function directly expressed as the OR of ANDs of the salient variables, possibly negated.

When the variables are selected, a mathematical model of the underlying generating framework still must be produced. At this point, a first hypothesis of linearity may be investigated (usually being only a very rough approximation) where the values of the variables are not close to the functioning point around which the linear approximation is computed.

Building a non-linear model is far from easy; the structure of the non-linearity needs to be a priori knowledge, which is not usually the case. A typical approach consists of exploiting a priori knowledge, when available, to define a tentative structure, then refining and modifying it on the training subset of data, and finally retaining the structure that best fits a cross-validation on the testing subset of data. The problem is even more complex when the collected data exhibit hybrid dynamics (i.e., their evolution in time is a sequence of smooth behaviors and abrupt changes).

An alternative approach is to infer the model directly from the data without a priori knowledge via an identification algorithm capable of reconstructing a very general class of a piece-wise affine model (Ferrari-Trecate et al., 2003). This method also can be exploited for the data-driven modeling of hybrid dynamic systems, where logic phenomena interact with the evolution of continuous-valued variables. Such an approach will be described concisely later, after a more detailed drawing of the rules-oriented mining procedure, and some applications will be discussed briefly.

Binary Rule Generation and Variable Selection While Mining Data

The approach followed by Hamming clustering in mining the available data to select the salient variables and to build the desired set of rules consists of the three steps in Table 1.

- **Step 1:** A critical issue is the partition of a possibly continuous range in intervals, whose number and limits may affect the final result. The thermometer code then may be used to preserve ordering and distance (in the case of nominal input variables, for which a natural ordering cannot be defined, instead adopting the only-one). The simple metric used is the Hamming distance, computed as the number of different bits between binary strings. In this way, the training process does not require floating point computation but rather basic logic operations. This is one reason for the algorithm speed and for its insensitivity to precision.
- **Step 2:** Classical techniques of logical synthesis are specifically designed to obtain the simplest AND-OR expression able to satisfy all the available input-output pairs without an explicit attitude to generalization. To generalize and infer the underlying rules at every iteration, by Hamming clustering groups together in a competitive way, binary strings have the same output and are close to each other. A final pruning phase does simplify the resulting expression, further improving its generalization ability. Moreover, the minimization of the involved variables intrinsically excludes the redundant ones, thus enhancing the very salient variables for the investigated problem. The low (quadratic) computational cost allows for managing quite large datasets.
- **Step 3:** Each logical product directly provides an intelligible rule, synthesizing a relevant aspect of the underlying system that is believed to generate the available samples.

Table 1. The three steps executed by Hamming clustering to build the set of rules embedded in the mined data

1. The input variables are converted into binary strings via a coding designed to preserve distance and, if relevant, ordering. 2. The 'OR of ANDs' expression of a logical function is derived from the training examples coded in the binary form of step 1. 3. In the OR final expression, each logical AND provides intelligible conjunctions or disjunctions of the involved variables, ruling the analyzed problem.

Identification of Piece-wise Affine Systems Through a Clustering Technique

Once the salient variables have been selected, it may be of interest to capture a model of their dynamical interaction. Piece-wise affine identification exploits K-means clustering that associates data points in multivariable space in such a way that jointly determines a sequence of linear submodels and their respective regions of operation without even imposing continuity at each change in the derivative. In order to obtain such a result, the five steps reported in Table 2 are executed.

- **Step 1:** The model is locally linear; small sets of data points close to each other likely belong to the same submodel. For each data point, a local set is built,