Discovering dynamics from data

Viljem Križman, Matjaž Gams and Sašo Džeroski
Jožef Stefan Institute, Jamova 39, 61111 Ljubljana, Slovenia

Abstract

GOLDHORN is a machine discovery system intended to discover empirical laws that govern the behavior of dynamical systems. It upgrades its predecessor LAGRANGE with the ability to handle real world data. The main technique used in GOLDHORN is numerical integration and simulation of differential equations instead of derivation. Preprocessing of the input data with digital filters is also possible, as well as the discovery of difference equations. GOLDHORN was used to discover differential/difference equation models from measured data in several domains, including fluid dynamics and algal growth.

1 Introduction

While several sytems exist that discover empirical numerical laws from data, e.g., BACON [Langley, et al. 1987], ABACUS [Falkenheiner & Michalski, 1990] and FAHRENHEIT [Żytkow & Zhu, 1991], few have so far addressed the problem of discovering laws that govern dynamical systems. LAGRANGE [Džeroski & Todorovski 1993] and GPDD [Džeroski & Petrovski, 1994] were unique in this respect. The task addressed by these systems can be defined as follows. A set of real-valued system variables $X_1, \ldots, X_n$ is measured at regular intervals over a period of time, i.e., in the time points t_0, $t_0 + h$, ..., $t_0 + Nh$. Given such a behavior, system attempts to find a set of laws that describe the dynamics of the system. The laws to be discovered (also called a model) typically take the form of a set of ordinary differential equations, which is the task of dynamics discovery.

In addition to the input behavior, systems have to be provided with the values of several parameters: the order o of the dynamical system (the order of the highest derivative appearing in the dynamics equations), the maximum depth d of new terms introduced by combining old terms (variables), and the maximum number r of independent regression variables used for generating equations.

The LAGRANGE algorithm consists of three main stages. Taking the set of system variables, LAGRANGE first introduces their time derivatives (up to order o). It then introduces new variables (terms) by repeatedly applying multiplication to variables from S and their time derivatives. Finally, given the set V of all variables (terms), LAGRANGE generates and tests equations by using linear regression.

Roughly speaking, each subset of V sized at most $r + 1$ is used to generate a linear equation. The term with greatest depth (complexity) is chosen as the dependent variable and is expressed as a linear combination of the remaining ones. The constant coefficients in

the linear equation are calculated by applying linear regression. If the equation appears to be significant, it is added to the model. The significance of an equation is judged by the multiple correlation coefficient R and the normalized deviation S, calculated as $R^2 = 1 - E/\sum_{j=0}^{N}(y_j - \overline{y})^2$ and $S^2 = E/[(N+1)(y^2 + e^{-\overline{y}^2})]$, where E is the sum of squared errors of the dependent variable y. Smaller values of S and larger values of R correspond to more significant equations.

LAGRANGE has been applied to reconstruct the models of several dynamical systems [Džeroski & Todorovski 1993, Todorovski & Džeroski 1994], the most complex being the inverted pendulum, a standard benchmark problem for dynamic system control. However, the experiments have been performed on simulated data. Applications to modelling real dynamical systems have been hindered by the sensitivity of LAGRANGE to noise and other (less important) problems.

We have identified two main problems with LAGRANGE that are of statistical nature: the choice of the dependent variable for linear regression and the sensitivity to noise in the data. The latter is especially important, as numerical derivation is used. The problem of choosing the dependent variable is more serious in the presence of noise. First we describe GOLDHORN and the techniques it uses. The application of GOLDHORN to modelling two real dynamical systems from measured data is described next. One system from the area of fluid dynamics and algal growth in the Lagoon of Venice are successfully modelled. We conclude with a brief discussion.

2 GOLDHORN

This section describes the system GOLDHORN that upgrades LAGRANGE in several directions. First, it is implemented as spreadsheet application, so it allows extensive intraction with the user. Simple data analisis, as well as unlimited new variables introduction can be done while using spreadsheet. Second, it expresses the highest order derivatives explicitly as rational functions of the system variables and their lower order derivatives. By doing this, the system avoids the need to use the highest order numerical derivatives. To estimate equation coefficients, as well as the quality of the explicit equations, GOLDHORN uses numerical integration, rather than derivation. In addition, GOLDHORN allows the measured data to be pre-processed with filters that alleviate the effects of noise to a certain degree. Finally, GOLDHORN can be used to discover difference equations, thus avoiding numerical derivation.

LAGRANGE seeks equations of the form $F(X_1, \ldots, X_n, \dot{X}_1, \ldots, \dot{X}_n, \ldots, X_1^{(o)}, \ldots, X_n^{(o)}) = 0$, where $X_1, \ldots, X_n$ are the system variables, o the order of the dynamic system (i.e., of the highest order derivatives in the dynamics equations) and F is a polynomial of degree d or less. In GOLDHORN, we can restrict the search to equations that are linear in the highest order derivatives $X_i^{(o)}$ and can be rewritten as $X_i^{(o)} = F_i(X_1, \ldots, X_n, \dot{X}_1, \ldots, \dot{X}_n, \ldots, X_1^{(o-1)}, \ldots, X_n^{(o-1)})$, where F_i is a rational function. In this way, we obtain equations that are suitable for simulation of the dynamical system. As a bonus, the number of equations considered is greatly reduced. Namely, when introducing new variables with multiplication, only variables that contain at most one highest

order derivative are introduced. Furthermore, only equations where at least one term contains a highest order derivative are considered.

GOLDHORN first introduces all derivatives by numerical derivation. It then considers all implict equations that can be rewritten in explict form. The term with largest variance is chosen as the dependent variable and the initial coefficients are determined by linear regression. GOLDHORN then expresses $X_i^{(o)}$ explicitly, i.e., $X_i^{(o)} = F_i$, and uses nonlinear optimization and numerical integration to further fit the coefficents, i.e., to minimize the error defined as

$$E = \sum_{j=0}^{N} (X_i^{(o-1)}(t_0 + jh) - \int_{t=0}^{t_0+jh} F_i(t_0 + jh)dt)^2.$$

The downhill simplex method of nonlinear optimization [Press, et al .1986] is used. The quality of an equation is then judged by the quantity $A = E/\sum_{j=0}^{N} (X_i^{(o)}(t_0 + jh) - \overline{X_i^{(o)}})^2$. The lower A, the more significant the equation.

Digital filtering can be applied to measured signals (dynamical system behaviors) to selectively remove noise, e.g., a low-pass filter can be applied to remove high frequency noise. Originally, GOLDHORN included several linear filters with finite impulse response, but in the latest implementation the polinomial filters were added which better filter noise.

Finally, let us note that difference equations, instead of differential equations can be produced by GOLDHORN, thus avoiding the need for numerical derivation altogether. Instead of introducing $\dot{X}(t) = dX(t)/d(t)$, we introduce $X'(t) = X(t + h)$. In this case, numerical integration and derivation are not used. In addition to the correlation coefficient R and the normalized deviation S, the sum of squared error E can be used to estimate the significance of equations: the lower E, the more significant the equation.

3 Modelling real dynamical systems

Modelling water level oscillations in a surge tank

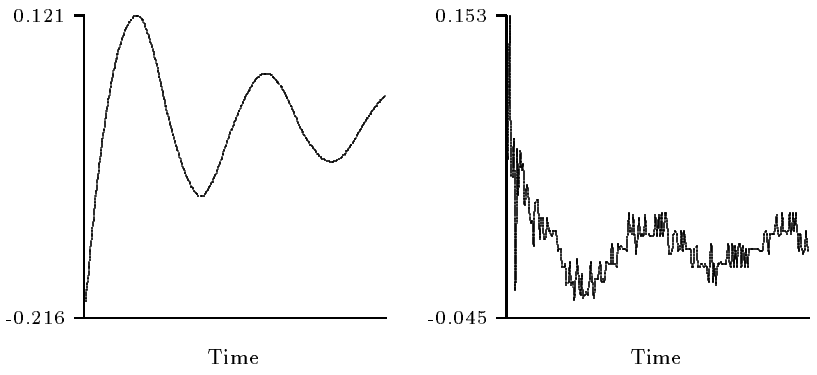

Figure 1: Water level (left) oscillations and the corresponding flow (right) in a surge tank.

A surge tank is a container that is usually connected to the pressure pipes that conduct water to the turbines of a hydro-power plant. If a sudden change of the flow through the

turbines occurs, a pressure surge is generated in the pipeline. The surge can have serious consequences to the pipeline (explosion or implosion), so a surge tank is situated as close as possible to the place of formation of the surge, i.e., the turbine valve. Surge pressure is transformed to water movement in the surge tank, resulting in an increase or decrease of the steady-state water level in the tank. The rest of the pipeline is thus not exposed to the pressure shocks.

In this experiment, GOLDHORN was used to model the water level oscillations in a laboratory replica of a surge tank. Only one variable is measured, i.e., the water level H. The measurements and the numerical derivative of the water level are depicted in Figure 1. The system is a second order one. In addition to the water flow $\dot{H}$, the magnitude of the flow $|\dot{H}|$ was provided.

The maximum depth of terms was set to two and the number of independent regression variables to three. Several filters were applied to the data. The following equation was the best according to the A measure $\ddot{H} = -0.057H - 3.629\dot{H}|\dot{H}|$. When simulated, this equation reproduces the observed behavior almost exactly (no visible difference). On the nonfiltered data, GOLDHORN produced the equation $\ddot{H} = -0.059H - 2.737\dot{H}|\dot{H}|$. For comparison, LAGRANGE produced the equation $\ddot{H} = 0.239 - 0.184H - 23.9777\dot{H}|\dot{H}|$ from the nonfiltered data, and $\ddot{H} = -0.058H - 3.829\dot{H}|\dot{H}|$ from filtered data.

Modelling algal growth in the Lagoon of Venice

The Lagoon of Venice measures 550 km^2, but is very shallow, with an average depth of less than 1m. It is heavily influenced by anthropogenic inflow of nutrients - 7 mio kg/year of nitrogen and 1.4 mio kg/year of phosphorus [Bendoricchio, et al. 1994]. These loads (mainly nitrogen) are above the Lagoon's admissible trophic limit and generate its dystrophic behavior, which is characterized by excessive growth of algae, mainly Ulva rigida.

Four sets of measured data were available [Coffaro, et al. 1993]. The data were sampled weekly for slightly more than one year at four different locations in the Lagoon. Location 0 was sampled in 1985/86, locations 1, 2, and 3 in 1990/91. The sampled quantities are nitrogen in ammonia NH_3, nitrogen in nitrate NO_3, phosphorus in orthophosphate PO_4 (all in μg/l), dissolved oxygen DO (in % of saturation), temperature T (degrees C), and algal biomass B (dry weight in g/m^2). In addition to the measured variables, GOLDHORN was provided with the growth μ and mortality ω rates, which are known quantities in ecology and can be calculated from the measured variables.

We applied GOLDHORN to model algal growth at Station 0. Difference equations were sought that express $B(t+1)$, i.e., the algal biomass at week $t+1$, in terms of the measured variables and the growth/mortality rates at week t, i.e., $NH_3(t)$, $NO_3(t)$, $PO_4(t)$, $DO(t)$, $T(t)$, $B(t)$, $\mu(t)$, and $\omega(t)$. The depth of variables was set to two and the number of independent regression variables to eight. According to the sum of squared errors E, the best equation was $B(t+1) = -0.611/\omega(t) - 2077\omega(t) + 0.653DO(t) + 0.662B(t) + 7.490T(t)$.

Figure 2 depicts the measured biomass and the biomass predicted by simulating the above equation. While the fit is not perfect, one should take into account that measure-

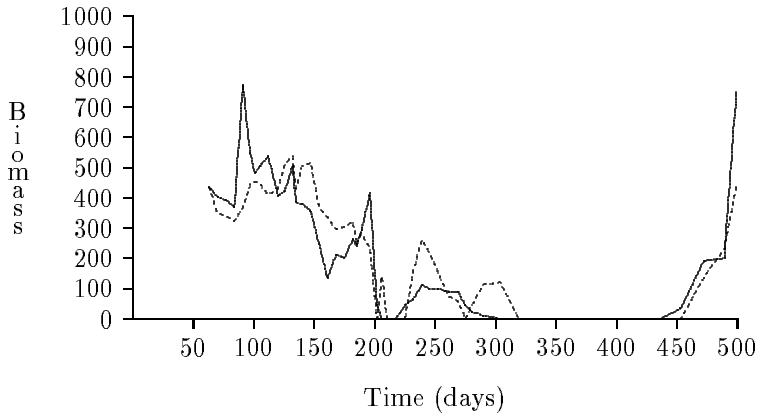

Figure 2: Algal biomass in the Venice lagoon as measured (solid line) and predicted by the equation discovered by GOLDHORN (dashed line).

ment errors for the biomass are of the order 20-50 %. The equation predicts correctly most of the peaks and crashes, both in time and to a certain degree in magnitude. These quantites are more important to ecologists than the degree of fit.

4 Discussion

We have presented the GOLDHORN system for machine discovery of empirical laws that govern dynamical systems. As compared to its predecessor LAGRANGE, GOLDHORN has the important ability to handle noisy data. The particular techniques used are digital filtering, numerical integration, and the discovery of difference equations. While the discovery of difference equations has already been proposed [Todorovski & Džeroski, 1994], it has not been applied to measured data.

We applied GOLDHORN to model two real dynamical systems from measured data. Good models, with almost perfect fit in one case and a good qualitative fit in the other, were obtained. One of the two systems was a laboratory replica of a real systems and was measured under controlled conditions. The other process, namely algal growth in the Lagoon of Venice, takes place in a real dynamical system. It has all the characteristics that make automated modelling difficult: measurement errors for biomass are between 20 and 50 %, important factors that influence the measured biomass are not taken into account, e.g., winds and tidal currents, etc. Nevertheless, the model constructed by GOLDHORN captures important properties of algal growth as it is able to predict correctly most of the algal blooms (peaks) and crashes: it is the blooms and crashes that ecologists are most worried about.

References

[1] Bendoricchio, G., Coffaro, G., and De Marchi, C. (1994). A trophic model for Ulva Rigida in the Lagoon of Venice. *Ecological Modelling*, 75/76: 485–496.

[2] Coffaro, G., Carrer, G., and Bendoricchio, G. (1993). *Model for Ulva Rigida Growth in the Lagoon of Venice.* Report UNESCO MURST Research Project Venice Lagoon Ecosystem. University of Padova.

[3] Džeroski, S. and Petrovski, I. (1994). Discovering dynamics with genetic programming. In *Proc. Seventh European Conference on Machine Learning.* Springer, Berlin. To appear.

[4] Džeroski, S. and Todorovski, L. (1993). Discovering dynamics. In *Proc. Tenth International Conference on Machine Learning*, pages 97–103. Morgan Kaufmann, San Mateo, CA, 1993.

[5] Falkenheiner, B. and Michalski, R. (1990). Integrating quantitative and qualitative discovery in the ABACUS system. In Kodratoff, Y. and Michalski, R., editors, *Machine Learning: An Artificial Intelligence Approach*, pages 153–190. Morgan Kaufmann, San Mateo, CA.

[6] Križman, V. (1994) Handling noisy data in automated modeling of dynamical systems. MSc Thesis, Faculty of Electrical and Computer Engineering, University of Ljubljana, Slovenia.

[7] Križman, V. (1998) Automated structure identification of dynamic systems models PhD Thesis, Faculty of Electrical and Computer Engineering, University of Ljubljana, Slovenia.

[8] Langley, P., Simon, H., Bradshaw, G., and Żytkow, J. (1987). *Scientific discovery.* MIT Press, Cambridge, MA.

[9] Moulet, M. (1994). Iterative model construction with regression. In *Proc. Eleventh European Conference on Artificial Intelligence*, pages 448–452. John Wiley & Sons, Chichester.

[10] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1986). *Numerical Recipes.* Cambridge University Press, Cambridge, MA.

[11] Todorovski, L. and Džeroski, S. (1994). Modeling dynamic systems with machine discovery. *Electrotechnical Review*, 61(1-2): 55–64. In Slovenian.

[12] Żytkow, J. and Zhu, J. (1991). Application of empirical discovery in knowledge acquisition. In *Proc. Fifth European Working Session on Learning*, pages 101–117. Springer, Berlin.