Математическое и компьютерное моделирование:
материалы III Международной научной конференции
(Омск, 12 ноября 2015 г.). Омск, 2015. С. 137–139. UDC 004.89

***G. Ramirez*[1], *C. Tweedie*[2], *J. Carlsson*[2], *V. Kreinovich*[2]**

[1] *New Mexico State University, Las Cruces, USA*
[2] *University of Texas at El Paso, El Paso, USA*

HOW TO MODIFY DATA PROCESSING ALGORITHMS SO THAT THEY DETECT ONLY DEPENDENCIES WHICH MAKE SENSE TO DOMAIN EXPERTS

Formulation of the problem: need to fit the data to a model in which only expert-determined pairs of variables are directly related. In many areas of physics, we know the equations relating different quantities. However, in many other application areas, e.g., in environmental sciences, we do not know these equations. So, we have to determined the dependence between different quantities based on observations.

There are many techniques to determining dependencies from data: linear and nonlinear regression, neural networks, etc. When we apply these techniques to environmental data, we get dependencies which are in reasonably good accordance with the data. However, ecologists are reluctant to accept the resulting models, since these models seem to indicate a direct relation between (almost) all pairs of quantities – including pairs for which there is no good physical reason to expect such a dependence.

It is therefore desirable to come up with models in which there direct dependence is limited to pairs of quantities for which such a dependence makes sense to domain experts.

Our result: it is always possible to find such a dependence. Let us first consider the simplest case, when we have a continuous observation of several quantities, i.e., when we know the values $x_i(t)$ of different quantities i in different moments of time.

Experts select pairs for which there may be a direct dependence. Thus, we can form a graph in which vertices are quantities, and an edge indicates a possible dependence. The transitive closure of this graph should be the set of all quantities – since otherwise we would have two sets of completely independent quantities, while in environmental sciences, all the quantities are usually dependent – directly or indirectly.

This means that for every two quantities x_i and x_j, there exists a connecting chain of dependent pairs: $x_i \sim c_{i_1} \sim x_{i_2} \sim \ldots \sim x_j$. How do we describe the dependence between the dependent pairs $x \sim y$?

In this case, as we will show, all the observations are consistent with the assumption that all the dependencies are linear. Indeed, a general shift-invariant linear dependence has the form

$$y(t) = \int k(t-s) \cdot x(s)\, ds$$

for some function $k(t)$. In Fourier transform, this leads to

$$\hat{y}(\omega) = \hat{k}(\omega) \cdot \hat{x}(\omega).$$

Thus, we can always take

$$\hat{k}(\omega) = \frac{\hat{y}(\omega)}{\hat{x}(\omega)}.$$

In the case when we consider observations in several regions α, in general, linear dependence is no longer possible, since by applying the above formula to observations corresponding to different regions, we can get different values of $k(t)$. In this case, we can consider non-linear dependencies, e.g., quadratic ones. A general shift-invariant quadratic dependence takes the form

$$y_\alpha(t) = \int k(t-s) \cdot x_\alpha(s)\, ds +$$

$$\int k_2(t-s_1, t-s_2) \cdot x_\alpha(s_1) \cdot x_\alpha(s_2)\, ds_1\, ds_2.$$

Once we know the observations $x_\alpha(t)$ and $y_\alpha(t)$ corresponding to different regions α, we get a system of linear equations for determining the unknown values $k(t)$ and $k_2(t_1, t_2)$. For T moments of time and R regions, we have $T \cdot R$ equations with T^2 unknowns. Usually, $T \gg R$, so we have more variables than equations. Thus, in general, this system has a solution – which means that we can indeed always describe the data by a model in which only expert-determined pairs of quantities are directly related.

How can we actually find these dependencies? For each quantity i, we can use, e.g., a neural network to find the desired dependence of this quantity – almost like it is usually done. The only difference is that as inputs, we only allow quantities which, according to the expert, can influence this quantity.

Acknowledgments. This work was supported in part by the National Science Foundation grants HRD-0734825, HRD-1242122, and DUE-0926721.