Soft Computing: Frontiers? A Case Study of Hyper-Spectral Satellite Imaging

Scott Starks
NASA Pan-American Center for
Environmental and Earth Studies (PACES)
University of Texas at El Paso
El Paso, TX 79968, USA
sstarks@utep.edu

Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
vladik@cs.utep.edu

Abstract

Soft computing methods such as fuzzy control, neural networks, etc., often require lots of computations even for small amounts of data. It is, therefore, sometimes believed that for larger amounts of data, the required amount of computations will be so large that we will reach the frontiers of soft computing.

In this paper, we show, on the example of hyper-spectral satellite imaging, that this belief is often too pessimistic. We should not be afraid to use (or at least to try to use) soft computing methods even for large amounts of data.

The problem: it looks like soft computing is approaching its frontiers

Often, soft computing requires lots of computations. Soft computing methods such as fuzzy control, neural networks, etc., often require lots of computations even for small amounts of data:

- When we use fuzzy control to describe a system with n input variables $x_1, \ldots, x_n$, then, even if we only use 2 different levels of each variable, we will still need 2^n rules. Even for reasonably small n, this is a huge number.
- Neural networks are also known to be slow to learn, even for small amounts of data. It is typical to have several thousand iterations to learn a simple dependence.

Pessimistic conclusions. If we simply extrapolation this already large amount of computation to the case when we have more input data, we will have to conclude that the required amount of computations will be so large that we will, very soon, reach the frontiers of soft computing.

What we are planning to do. In this paper, we show, on the example of hyper-spectral satellite imaging, that this belief is often too pessimistic. We should not be afraid to use (or at least to try to use) soft computing methods even for large amounts of data.

Case study

Satellite imaging. Nowadays, satellite imaging is one of the most important sources of geographical, geophysical, and environmental information. Satellite images can determine the amount and type of vegetation, the geological type of the underlying soils (and often, of the minerals below), etc.

However, with the current satellite images, it is sometimes difficult to decide what exactly we observe, because the existing Earth-sensing satellites, such as Landsat, only take the images at a few (≤ 7) frequencies.

An example of a problem in which a satellite image is currently not sufficient: kaolinite vs. dickite. Based on the (inevitably imprecise) measurements on the few frequencies, it is difficult, e.g., to distinguish between kaolinite and its rare amorphous but chemically similar forms such as dickite.

Kaolinite and dickite are the principle ingredients of *kaolin*, a soft white-clay mineral that is an essential ingredient in the manufacture of china and porcelain and is also widely used in the making of paper, rubber, paint, and many other products (see, e.g., (Patterson 1984)). It is also used in medicine: e.g., in the treatment of diarrhea, kaolin powder is the most widely used absorbent powder. Due to kaolin's importance, it is desirable to determine not only its *presence*, but its *type* as well.

Since crystal-based kaolinite and amorphous dickite are chemically similar, their spectra are very similar. Therefore, currently, in order to distinguish between these two minerals, we have to complement satellite images with geophysical and radar data (see, e.g., (Merényi et al. 1994) and (Merényi et al. 1997)).

Hyper-spectral satellite imaging. To produce more data, NASA is planning to launch the imaging satellites of the new generation, satellites that will have the ability to map the Earth on up to 500 optical frequencies. These coming satellites are nicknamed *Lewis* after the famous 19 century US geographer.

From the resulting multi-spectral images, it is, in principle, possible to determine many characteristics of soil and vegetation without using additional data; see (Moon and Merényi 1995) and (Merényi et al. 1996). For example, it is, in principle, possible to distinguish between kaolinite and dickite because from hyper-spectral images, we can extract spectra in each point, i.e., the dependence of its brightness $I(f)$ on the frequency f; the corresponding spectra, although similar, have different number of local maxima.

Perceived problems with the use of soft computing in hyper-spectral imaging. If we want to use fuzzy control methods for $n = 500$ inputs, we will need $2^{500} \approx 10^{150}$ rules, more rules than there are particles in the known Universe. This is completely unrealistic.

Similar estimates makes us believe that neural methods cannot be used.

In reality, neural methods are well applicable. Surprisingly, neural network methods work very well for hyper-spectral images; see, e.g., (Merényi et al. 1996) and (Moon and Merényi 1995)).

Moreover, surprisingly, even *linear* methods, that were not applicable at all for 7-frequency images, work pretty well for hyper-spectral ones.

The question is: Why?

There are two possible answers:

- A *pessimistic* answer: we accidentally run into a problem that is unusually simple, but in general, problems of this type are tough.
- An *optimistic* answer: in general, the above pessimistic estimates were indeed too pessimistic, and soft computing is not yet near the frontier.

In this text, we will give two arguments in favor of the optimistic viewpoint.

First argument: for several reasonable classes of functions, neural networks stay feasible even for large number of inputs

In the last decade, there have been several research papers that study the approximation of functions from R^n to R by (feedforward) neural networks. Several very general universal approximation results has been proven; see, e.g., (Mhaskar and Micchelli 1992), (Park and Sandberg 1993).

However, if we do not restrict the class of the approximated functions, the worst-case estimates on the number of hidden units, that guarantee a given accuracy of approximation, grows exponentially with the number n of variables of the function $f(x_1, \ldots, x_n)$ to be approximated; see, e.g., (Mhaskar and Micchelli 1992) and (Kůrková 1992). A general result from (DeVore, Howard, and Micchelli 1992) confirms that there is no hope for a better estimate when the class of multivariable functions being approximated is defined in terms of the bounds of partial derivatives and parameters of approximating networks are chosen continuously.

On the other hand, in applications, functions of hundreds of variables are approximated sufficiently well by neural networks with only moderately many hidden units; see, e.g., e.g., (Sejnowski and Yuhas 1989).

To explain this result, Jones (Jones 1992) and Barron (Barron 1992), (Barron 1993) showed that it is possible to approximate any function satisfying a certain condition with a given accuracy $\varepsilon > 0$ by a neural network with $O(\varepsilon^{-1})$ neurons. These results were further generalized in (Kůrková, Kainen, and Kreinovich 1995) and (Kůrková, Kainen, and Kreinovich 1997).

Second argument: linear methods are indeed helpful

In this section, we will explain why linear methods work well. The surprise efficiency of linear methods is a very general phenomenon that occurs not only in the new field of hyperspectral image processing, but also in such well-established areas as geophysics, economics, etc. Let us therefore describe a general background.

Indirect measurements. In many real-life situations, we must estimate the value of a physical quantity y that is difficult to measure directly. So, to avoid direct measurements, we measure y *indirectly:* namely, we measure whatever variables we can, and then based on the measured value $x_1, \ldots, x_n$, we try to estimate y.

These situations are very frequent in geophysics, when it is very costly to measure the properties of the deep layers, and much cheaper to measure the waves reflected from those layers.

In many cases, we do not know the exact relation between x_i and y. In some situations, we know the relationship between x_i and y, so we can use this known model to estimate y based on x_i. In geophysics, this relationship is usually highly non-linear.

In many situations, however, this dependency between x_i and y is not a priori known, and has to be determined experimentally.

There exist many methods and algorithms that extract the dependency from the experimental data. There exist many statistical methods that help us to discover such a dependency (see, e.g., (Deming 1964)). To apply them, we must have several situations in which we know both y and x_i. So, we have the values $x_{11}, \ldots, x_{n1}, y_1$ that correspond to the first situation, the values $x_{21}, \ldots, x_{2n}, y_2$ that were measured in the second situation, ..., and, finally, the values $x_{N1}, \ldots, x_{Nn}, y_N$ (here, by N we denoted the total number of such situations). From these data, we extract a function f such that $y_k = f(x_{k1}, \ldots, x_{kn})$ for $1 \le k \le N$.

When this function is known, we can use it to estimate y from x_i for the new measuring situations.

The simplest of these methods (called *linear regression*) uncovers linear dependencies, i.e., dependencies of the type $y = a_0 + a_1 x_1 + \ldots + a_n x_n$ for some constant coefficients a_i. The corresponding statistical software packages have built-in checks on whether the data is really consistent with linear dependency (e.g., χ^2-method).

Non-linear curve-fitting methods also exist.

For the cases when we are not sure whether the relationship is linear or not, traditional statistical methodology requires that we first try simple linear regression methods, and then, if linear methods do not work, try more complicated non-linear methods.

In geophysics, we expect non-linearity, but surprisingly, linear regression methods work fine. The equations that describe the geophysical phenomena are *highly non-linear.* So, we would expect that in the majority of cases, linear regression will fail, and we will have to use non-linear methods.

Unexpectedly, in many cases, linear methods succeed! Namely, they generate reasonable linear dependencies that fit the experimental data perfectly (fits in the sense that built-in checks of linearity confirm that linearity is possible). This phenomenon occurs in all areas of geophysics: in seismology (Thurber 1983), (Baker and Doser 1988), in electrical methods (Pous, Marcuello, and Queralt 1987), (Boerner and Holladay 1990), in general geophysics (Menke 1984), (Baker 1988).

The same strange phenomenon occurs in economics: when we, e.g., analyze the dependency of the workers-per-manager ratio on the parameters that characterize the business (Litov 1981), (Kreinovich 1989), we also get a pretty good fit for linear regression in an evidently non-linear situation.

Why? Why are linear methods working so well in non-linear situations? In this paper, we present an answer to this question. This answer will consist of two parts:

- First, we will show that in many real-life situations, the variables x_i are not independent, they are functionally dependent on each other.
- Second, we will show that these dependencies enables us to use linear regression.

If there is a model with m parameters, then, in principle, it is sufficient to measure m quantities x_i. Before we start the general argument, let us first consider the case when we know the model of the object that we are analyzing.

In many situations, there is a model that describes the analyzed phenomena. Usually, a model has several (unknown) parameters. For example, a geophysical area is often described in terms of 3 or more layers, with about 3 parameters to describe the properties of each layer.

If we know the model, then, of course, we know the number m of parameters $p_1, \ldots, p_m$ of that model. By saying that we have a model we mean that we know how the desired value y depends on these parameters, i.e., we know a function g that computes y from p_j: $y = g(p_1, \ldots, p_m)$. We also know how all other measurable quantities x_i depend on p_j: $x_i = h_i(p_1, \ldots, p_m)$. Suppose that we measured m quantities. Then, we have m equations $h_i(p_1, \ldots, p_m) = x_i$, $1 \leq i \leq m$, with the known right-hand sides and m unknowns $p_1, \ldots, p_m$.

In general, when the number of equations coincides with the number of unknowns, we have a unique solution. Therefore, from x_i, we can uniquely determine the parameters p_j. And as soon as we know the parameters, we can compute the value of $y = g(p_1, \ldots, p_m)$. So, from $x_1, \ldots, x_m$, we can (in principle) uniquely determine y.

What if we do not know the model, but we do know that the object is uniquely determined by m parameters? In this case, we do not know how to compute y from $x_1, \ldots, x_m$, but we still know that in principle, we can uniquely determine y from $x_i, 1 \leq i \leq m$. This is the case when regression methods have to be applied. So, we make several measurements of x_i and y in different situations, and try to find the dependency $y = f(x_1, \ldots, x_m)$.

If we use the smallest possible number of measurements, then we cannot apply linear methods. If we use exactly m measurements, and the actual function f is non-linear, then, of course, we cannot use linear regression methods.

In real life, we usually perform more measurements to increase precision. Since measurements are usually not ideally precise, the estimates that we get from the smallest possible amount of measurements $x_1, \ldots, x_m$ are also not precise. To make the estimates more precise, we perform additional measurements.

As a result, the number n of variables x_i that we measure is much greater than the smallest possible number m.

As a result, the variables x_i are interdependent. The same logic that showed that y is uniquely determined by m parameters $x_1, x_2, \ldots, x_m$, shows that any other physical characteristic of our object is also uniquely determined by $x_1, \ldots, x_m$. In particular, it is true for the quantities $x_{m+1}, x_{m+2}, \ldots, x_n$.

Therefore, $x_{m+1} = f_{m+1}(x_1, \ldots, x_m)$ for some function f_{m+1}, $x_{m+2} = f_{m+2}(x_1, \ldots, x_m)$ for some other function f_{m+2}, etc.

In other words, the variables x_i are *not* independent: there is a functional dependency between them.

Example. As an example, let us consider the simplest case when one parameter p is sufficient to determine the values of all the physical quantities. In this case, $y = g(p)$, and $x_i = h_i(p)$ for some functions g and h_i.

Since $m = 1$, we need only one measurement to determine p uniquely. If x_1 is known, then we can determine p as the solution of the equation $h_1(p) = x_1$. So, $p = h_1^{-1}(x_1)$, where by h_1^{-1}, we denoted a function that is inverse to h_1. Therefore, $y = g(p) = g(f^{-1}(x_1))$. Similarly, $x_2 = h_2(p) = h_2(h_1^{-1}(x_1))$, i.e., x_1 and x_2 are functionally dependent on each other. Likewise, x_1 and x_3, x_1 and x_4, etc, are mutually dependent.

Part II: If the values x_i are inter-dependent, then for sufficiently large n, we can apply linear regression. Let us see what happens when the values x_i are inter-dependent.

First case: only one parameter p. Let's first consider the case when everything depends on only one parameter p.

As long as the dependency of y and x_i on p is smooth (and it usually is), we can expand the functions $y = g(p)$ and $x_i = h_i(p)$ into the Taylor series:

$$y = a^{(0)} + a^{(1)}p + \ldots + a^{(l)}p^l + \ldots$$

$$x_i = a_i^{(0)} + a_i^{(1)}p + \ldots + a_i^{(l)}p^l + \ldots$$

We measure both x_i and y with a certain precision ε. Therefore, if we get the value $\tilde{x}_i$ as a result of the measurement, it means that an actual value of x_i belongs to an interval $[\tilde{x}_i - \varepsilon, \tilde{x}_i + \varepsilon]$. Because of this imprecision, we can use the approximate formula for $x_i(p)$ as long as the approximation error does not exceed ε. So, instead of taking all Taylor terms, we can retain only those that guarantee the precision ε. As a result, we get the following formulas:

$$y = a^{(0)} + a^{(1)}p + \ldots + a^{(l)}p^l,$$

$$x_i = a_i^{(0)} + a_i^{(1)}p + \ldots + a_i^{(l)}p^l.$$

If we denote by N the total number of situations that we can use to determine the dependency of y on x_i, by $p_{(k)}$ the (unknown) value of the parameter p in $k-$th experiment ($1 \le k \le N$), then we get the following formulas:

$$y_k = a^{(0)} + a^{(1)}p_{(k)} + \ldots + a^{(l)}p_{(k)}^l,$$

$$x_{1k} = a_1^{(0)} + a_1^{(1)}p_{(k)} + \ldots + a_1^{(l)}p_{(k)}^l,$$

$$x_{2k} = a_2^{(0)} + a_2^{(1)}p_{(k)} + \ldots + a_2^{(l)}p_{(k)}^l,$$

$$\ldots$$

$$x_{nk} = a_n^{(0)} + a_n^{(1)}p_{(k)} + \ldots + a_n^{(l)}p_{(k)}^l$$

for all $k = 1, 2, \ldots, N$.

We can form $N-$dimensional vectors $\vec{x}_i = (x_{i1}, \ldots, x_{iN})$ and $\vec{y} = (y_1, \ldots, y_N)$. The above equalities can be reformulated in terms of these vectors as follows:

$$\vec{x}_i = a_i^{(0)}\vec{1} + a_i^{(1)}\vec{p} + \ldots + a_i^{(l)}\vec{p}^{\,l},$$

$$\vec{y} = a^{(0)}\vec{1} + a^{(1)}\vec{p} + \ldots + a^{(l)}\vec{p}^{\,l},$$

where we denoted

$$\vec{1} = (1, 1, \ldots, 1),$$

$$\vec{p} = (p_{(1)}, \ldots, p_{(N)}),$$

$$\ldots$$

$$\vec{p}^{\,j} = (p_{(1)}^j, \ldots, p_{(N)}^j),$$

$$\ldots$$

So, $n + 1$ vectors $\vec{x}_1, \ldots \vec{x}_n, \vec{y}$ belong to a $(k + 1)-$dimensional space L: namely, to the linear space generated by $k + 1$ vectors $\vec{1}$, $\vec{p}$, $\vec{p}^{\,2}$, ..., $\vec{p}^{\,l}$.

When $n > l$, these vectors cannot be all linearly independent. Therefore, they are linearly dependent, i.e., there exist values c_i such that

$$c_1\vec{x}_1 + c_2\vec{x}_2 + \ldots + c_n\vec{x}_n + c_{n+1}\vec{y} = 0.$$

In general, all these coefficients c_i are different from 0. If we exclude the degenerate case when $c_{n+1} = 0$, then we can divide this equality by c_{n+1} (so that the coefficient at $\vec{y}$ will become equal to 1), and move all the terms except $\vec{y}$ to the other side of the equation. As a result, we get the formula

$$\vec{y} = a_1\vec{x}_1 + a_2\vec{x}_2 + \ldots + a_n\vec{x}_n,$$

where $a_i = -c_i/c_{n+1}$. If we now recall that each of these vectors $\vec{x}_i$ and $\vec{y}$ is actually formed by the measured values of the corresponding quantities in N situations, then from the above vector equality we can conclude that for all measurements $k = 1, 2, \ldots, N$, we have

$$y_k = a_1x_{1k} + a_2x_{2k} + \ldots + a_nx_{nk}.$$

In other words, for every measurement, the following equality holds:

$$y = a_1x_1 + a_2x_2 + \ldots + a_nx_n.$$

But this means exactly that *linear regression works.*

Example. Let us consider the case when $y = p + p^2$, $x_1 = p$, and $x_2 = p - p^2$. In this case, the variable y can be explicitly expressed in terms of x_1: $y = x_1 + x_1^2$. This expression is clearly non-linear, so, if we will try to apply linear regression to determine the dependency of y on x_1, the linear regression method will indicate failure.

However, if we try to represent y in terms of both x_1 *and* x_2, then we have a linear expression: $y = 2x_1 - x_2$. Indeed, $2p - (p - p^2) = p + p^2$.

General case. In the general case, we have m parameters $p_1, \ldots, p_m$. So, when we restrict the Taylor expansion to the terms $p_1^{d_1}p_2^{d_2}\ldots p_m^{d_m}$ of power $d_j \le l$, we get $(l+1)^m$ different terms (because each of d_j can take any value from 0 to l). Therefore, we need $(l+1)^m$ coefficients to describe an approximation.

In this case, if we similarly introduce the vectors, we will have $n + 1$ vectors $\vec{x}_i$ and $\vec{y}$ that are linear combinations of $(l+1)^m$ vectors $\vec{p}^{\,d_1,d_2,\ldots,d_m}$ with components $p_{1j}^{d_1}p_{2j}^{d_2}\ldots p_{mj}^{d_m}$, $1 \le j \le N$.

So, if the number of variables n is sufficiently large (in this case, if $n+1 > (l+1)^m$), then we can likewise conclude that the vectors $\vec{x}_i$ and $\vec{y}$ are linearly dependent, and therefore, *linear regression methods can be applied.*

Conclusion. If linear regression works in a non-linear situation, one does not need to search for an error. Moreover, if this is the situation, then we can be sure that the variables x_i are functionally interdependent, so we can look for the dependencies between them.

Acknowledgments

This work was partially supported by NSF Grants No. EEC-9322370 and DUE-9750858, by NASA under cooperative agreement NCCW-0089, and by the Future Aerospace Science and Technology Program (FAST) Center for Structural Integrity of Aerospace Systems, effort sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, under grant number F49620-95-1-0518.

We are very thankful to Brian Penn for helpful discussions.

References

Baker, M. R. 1988. Quantitative interpretation of geological and geophysical well data. Ph.D. Dissertation, Department of Geological Sciences, University of Texas at El Paso.

Baker, M. R., and D. I. Doser, D. I. 1988. Joint inversion of regional and teleseismic earthquake waveforms. *J. Geophys. Res.* 93: 2037–2046.

Barron, A. R. 1992. Neural net approximation, In Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems, 69–72.

Barron, A. R. 1993. Universal approximation bounds for superposition of a sigmoidal function. *IEEE Transactions on Information Theory* 39: 930–945.

Boerner, D. E., and Holladay, J. S. 1990. Approximate Frechet derivatives in inductive electromagnetic soundings. *Geophysics* 55: 1589–1595.

Deming, W. E. 1964. *Statistical adjustment of data.* N.Y.: Dover.

DeVore, R., Howard, R., and Micchelli, C. 1992. Optimal nonlinear approximation. *Manuscripta Mathematica* 63: 469–478.

Jones, L. K. 1992. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics* 20: 601–613.

Klir, G., and Yuan, B. 1995. *Fuzzy sets and fuzzy logic: theory and applications.* Upper Saddle River, NJ: Prentice Hall.

Kreinovich, V. How to explain the efficiency of linear regression for non-linear economic problems, Technical Report, Leningrad Technological Center Informatika, Leningrad (in Russian).

Kůrková, K. 1992. Kolmogorov's theorem and multilayer neural networks. *Neural Networks* 5: 501–506.

Kůrková, V., Kainen, P. C., and Kreinovich, V. 1995. Dimension-independent rates of approximation by neural networks and variation with respect to half-spaces. In Proceedings of World Congress on Neural Networks, WCNN'95, Washington, DC, July 1995, Vol. I, 54–57. NJ: INNS Press.

Kůrková, V., Kainen, P. C., and Kreinovich, V. 1997. Estimates of the Number of Hidden Units and Variation with Respect to Half-spaces. *Neural Networks* (to appear).

Litov, B. A. 1981. Analysis of the personnel structure of the design institutions. Theory and Methodology. Ph.D. Dissertation. Leningrad (in Russian).

Menke, W. 1984. *Geophysical Data Analysis: Discrete Inverse Theory.* N.Y.: Academic Press.

Merényi, E., Csathó, B., Bodrogi, M., and Gulyás, Á. 1994. Utilization of Landsat images for mapping natural resources and for enviromnetal protection in Hungary, In Proc. Tenth Thematic Conference on Geologic Remote Sensing, San Antonio, TX, May 9–12, 1994, Vol. II, 491–502.

Merényi, E., Csathó, B., Bodrogi, M., and Gulyás, Á. 1997. Integration of Landsat images, geophysical and radar data for mapping soil composition in temperate climate environment, Hungary. Submitted to *Remote Sens. Environ.*

Merényi, E., Taranik, J. V., Minor, T. B., and W. H. Farrand, W. H. 1996. Quantitative comparison of neural networks and conventional classifiers for hyperspectral imagiery. In: Green, R. O. ed. Summaries of the Sixth Annual JPL Airborne Earth Science Workshop. Pasadena, CA, March 4–8, 1996, Vol. 1.

Mhaskar, H. N., and Micchelli, C. A. 1992. Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied Mathematics* 13: 350–373.

Moon, T., and Merényi, E. 1995. Classification of hyperspectral images using wavelet transforms and neural networks. In Proceedings of the Annual SPIE Conference, 2569.

Park, J., and Sandberg, I. W. 1993. Approximation and radial-basis-function networks. *Neural Computation* 5: 305–316.

Patterson, S. H. 1984. *Kaolin, refractory clay, ball clay, and halloysite in North America, Hawaii, and the Caribbean region*, Alexandria, VA: U.S. Department of the Interior, Geological Survey.

Pous, J., Marcuello, A., and Queralt, P. 1987. Resistivity inversion with "a priori" information. *Geophysical Prospecting* 35: 590–603.

Sejnowski, T. J., and Yuhas, B. P. 1989. Mapping between high-dimensional representations of acoustic and speech signal, In *Computation and Cognition*, SIAM, Philadelphia: SIAM, 52–68.

Thurber, C. H. 1983. Earthquake locations and three-dimensional structure in the Coyote Lake area, central California. *J. Geophys. Res.* 88: 8226–8236.