Bilinear Models from System Approach Justified for Classification, with Potential Applications to Bioinformatics

Richard Aló[1], François Modave[2], Vladik Kreinovich[2],
David Herrera[2], and Xiaojing Wang[2]

[1] Center for Computational Sciences & Advanced Distributed Simulation,
University of Houston-Downtown, One Main Street,
Houston, TX 77002, USA, RAlo@uh.edu
[2] Department of Computer Science, University of Texas at El Paso,
El Paso, TX 79968, USA, vladik@utep.edu

Abstract. When we do not know the dynamics of a complex system, it is natural to use common sense to get a reasonable first approximation – which turns out to be a bilinear dynamics. Surprisingly, for classification problems, a similar bilinear approximation turns out to be unexpectedly accurate. In this paper, we provide an explanation for this accuracy.

1 System Approach: In Brief

For many complex systems, e.g., for large-scale financial or biological systems, we do not know the exact equations that describe the system's evolution. In such situations, a reasonable idea is to use *system approach*, i.e., to describe the dynamics of the parameters $x_1, \ldots, x_n$ (that describe the system) by differential equations

$$\dot{x}_i = f_i(x_1, \ldots, x_n),$$

and build the expressions for f_i based on common sense [1, 2, 4, 5] (see also [3]).

For example, if an increase in x_i slows down the growth of x_j, then the expression f_j for $\dot{x}_j$ should include a term $-k \cdot x_i$ for some $k > 0$.

If the two factors x_j and x_k, when combined, produce an increased positive effect on the growth of x_i, then the expression f_i for $\dot{x}_i$ should include a bilinear term $+k \cdot x_j \cdot x_k$. The values of the corresponding parameters k can be determined empirically.

Common sense rarely goes beyond such simple interaction between parameters, so we end up with linear and bilinear expressions in $\dot{x}_i$.

Starting from the global economic and environmental models developed in the 1960s by the Club of Rome's Limits to Growth project [2], the resulting models provide a reasonable first *qualitative* approximation to the dynamics of complex systems – although, of course, to get good *quantitative* predictions, we need more sophisticated models.

2 Systems Approach in Classification

A similar approach can be used in classification and clustering, where we need to separate, e.g., stocks with a good growth potential from the risky ones, or cancerous cells from the normal ones.

To separate two classes based on the values of the parameters x_i, we can use a discrimination function $f(x_1, \ldots, x_n)$:

- objects for which $f > 0$ belong to the first class, and
- objects for which $f < 0$ belong to the second class.

3 Systems Approach in Classification Works Surprisingly Well: A Mystery and a Related Practical Question

By using common sense, we can also easily come up with a bilinear model for f, and we can empirically find the coefficients of the corresponding bilinear expression.

We applied this approach to bioinformatics data, and surprisingly, the resulting bilinear models provide not only a good qualitative fit, but a good quantitative fit as well. A natural question is: Is it a special feature of bioinformatics data, or we can hope to get a good quantitative fit for, e.g., financial data as well?

4 Mystery Explained

In this paper, we show that the success of bilinear models in classification is a general mathematical phenomenon – caused by the fact that we can change a discrimination function f and keep the same classes, as long as we do not change the set of value for which $f \geq 0$.

5 Approximations of Different Order of Accuracy: General Idea

In accordance with the above text, let us denote the actual (unknown) discrimination function by $f(x_1, \ldots, x_n)$. It is usually reasonable to assume that the function $f(x_1, \ldots, x_n)$ is smooth. Therefore, in the vicinity of every point $\widetilde{x} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$, we can expand this function into the Taylor series and keep only lower order terms in this expansion.

By using this approximate polynomial expression for the function $f(x_1, \ldots, x_n)$, we can get a good approximate description of the classification in the vicinity of the point $\widetilde{x}$.

Which starting points should we consider?

– If we consider a point $\widetilde{x}$ which is completely inside the first class, then all the points in its vicinity belong to the same class – and thus, in this vicinity, the classification problem can be trivially solved.
– Similarly, if a point $\widetilde{x}$ is completely inside the second class, then all the points in its vicinity belong to the same class – and thus, in this vicinity, the classification problem can also be trivially solved.

Thus, the localized classification problem is of interest only when the starting point $\widetilde{x} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$ is located on the border between the two classes, i.e., when

$$f(\widetilde{x}_1, \ldots, \widetilde{x}_n) = 0.$$

We can somewhat simplify the formulas if we introduce new coordinates $x_i \to x_i - \widetilde{x}_i$ in which the starting point takes the form $(0, \ldots, 0)$. In these new coordinates, the condition that the starting points lies on the border between the two classes takes the form

$$f(0, \ldots, 0) = 0.$$

Thus, in the following text, we will assume that the starting point is 0, and that $f(0, \ldots, 0) = 0$.

6 First Approximation: Linear Separation

In the first approximation, we can approximate an arbitrary smooth function $f(x_1, \ldots, x_n)$ in the vicinity of 0 by a linear expression:

$$f(x_1, \ldots, x_n) = a_0 + \sum_{i=1}^{n} a_i \cdot x_i.$$

Since our interest is in classification, we are only interested in functions for which $f(0, \ldots, 0) = 0$. Substituting $x_1 = \ldots = x_n = 0$ into the above general expression for a linear function and equating the result to 0, we conclude that $a_0 = 0$.

Thus, in the first approximation, the general classification-related discrimination function takes the form

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{n} a_i \cdot x_i.$$

7 Second Approximation: Quadratic Separation

In the next (second) approximation, we can approximate an arbitrary smooth function $f(x_1, \ldots, x_n)$ in the vicinity of 0 by a quadratic expression:

$$f(x_1, \ldots, x_n) = a_0 + \sum_{i=1}^{n} a_i \cdot x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \cdot x_i \cdot x_j.$$

Since we need $f(0,\ldots,0)=0$, we conclude that $a_0=0$ and thus, that

$$f(x_1,\ldots,x_n)=\sum_{i=1}^{n} a_i\cdot x_i+\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\cdot x_i\cdot x_j.$$

This general expression is much more general than bilinear:

- it has linear terms $a_i\cdot x_i$;
- it has bilinear terms $a_{ij}\cdot x_i\cdot x_j$ for $i\neq j$;
- however, the general expression also has purely quadratic (not bilinear) terms $a_{ii}\cdot x_i^2$ (corresponding to $i=j$).

8 For System Dynamics, Bilinear Functions Provide a Rather Crude Approximation

In the original application of systems theory, the function $f(x_1,\ldots,x_n)$ describes the dynamics of the system, i.e., the (first) time derivative of the corresponding coordinates. In such applications, the function $f(x_1,\ldots,x_n)$ can be determined by observing the dynamics of the system, i.e., by observing how the values of the parameters x_i change with time. Since in principle, we can have an arbitrary dynamical system, we can therefore have arbitrary functions $f(x_1,\ldots,x_n)$.

When we approximate a general function $f(x_1,\ldots,x_n)$ by a linear expression, we thus ignore quadratic and higher order terms in the expansion of this function $f(x_1,\ldots,x_n)$. Thus, the approximation error of this approximation is quadratic (in x_i).

Similarly, when we approximate a general function $f(x_1,\ldots,x_n)$ by a quadratic expression, we thus ignore cubic and higher order terms in the expansion of this function $f(x_1,\ldots,x_n)$. Thus, the approximation error of this approximation is cubic (in x_i).

In principle, we can consider systems-theory bilinear approximation, in which we only keep linear and bilinear terms in the quadratic expansion but we ignore purely quadratic terms $a_{ii}\cdot x_i^2$. This approximation is intermediate between the linear and the quadratic ones.

What is the accuracy of this intermediate approximation? Since we ignore some quadratic terms, the accuracy of this approximation is quadratic is x_i. In other words, this intermediate approximation is asymptotically of the same order of accuracy as the much simpler linear approximation.

Comment. Of course, the accuracy of the bilinear model is somewhat better than the accuracy of the linear approximation – since in the intermediate approximation, we keep some quadratic terms. For example, in systems in which all the variables x_i are highly related, it is reasonable assume that all the terms $a_{ij}\cdot x_i\cdot x_j$ are of the same order of magnitude. In this case,

- in the linear approximation, we ignore n^2 terms $a_{ij}\cdot x_i\cdot x_j$ corresponding to all possible pairs (i,j), $i=1,2,\ldots,n$, $j=1,2,\ldots,n$; so, the approximation error is of order $n^2\cdot\delta$, where δ is the average value of each term $a_{ij}\cdot x_i\cdot x_j$;

– in the bilinear approximation, we only ignore n quadratic terms $a_{ii} \cdot x_i^2$ corresponding to $i = 1, 2, \ldots, n$; so, the approximation error is of order $n \cdot \delta$.

Thus, the approximation error of the bilinear model is $\approx n$ times smaller than for the linear model – i.e., for large n, drastically smaller.

9 The Fact that We Are Interested in Classification Applications Allows Further Simplifications

In dynamics applications, the function $f(x_1, \ldots, x_n)$ can be determined from observations. In the classification applications, however, the only think that we can infer from observations is for which points $x = (x_1, \ldots, x_n)$, the function $f(x_1, \ldots, x_n)$ has positive values and for which points $x = (x_1, \ldots, x_n)$, the function $f(x_, \ldots, x_n)$ has negative values.

Thus, if we replace the original function $f(x_1, \ldots, x_n)$ with an "equivalent" new function $f'(x_1, \ldots, x_n)$ – i.e., a new function that leads to the same subdivision of points into the first class and the second class – then both function will be consistent with the same observations. Thus, from observations, we will not be able to tell whether the actual discrimination function is $f(x_1, \ldots, x_n)$ or $f'(x_1, \ldots, x_n)$.

In this paper, we will use this non-uniqueness of the discrimination function $f(x_1, \ldots, x_n)$ to simplify the general quadratic expression for f.

In precise terms, the two functions $f(x_1, \ldots, x_n)$ and $f'(x_1, \ldots, x_n)$ are equivalent if the following two conditions are satisfied:

$$f(x_1, \ldots, x_n) > 0 \text{ if and only if } f'(x_1, \ldots, x_n) > 0$$

and

$$f(x_1, \ldots, x_n) > 0 \text{ if and only if } f'(x_1, \ldots, x_n) > 0.$$

10 Explanation of Bilinear Functions

As the new function $f'(x_1, \ldots, x_n)$, we will consider a function of the type

$$f'(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) \cdot \left(1 + \sum_{j=1}^{n} b_j \cdot x_j\right).$$

In the small vicinity of 0, $\left|\sum\limits_{i=j}^{n} b_j \cdot x_j\right| \ll 1$, hence

$$1 + \sum_{i=1}^{n} b_j \cdot x_j > 0.$$

Thus, in this vicinity, the values $f(x_1, \ldots, x_n)$ and $f'(x_1, \ldots, x_n)$ have the same sign – i.e., these functions are indeed equivalent.

Let us show that for a generic quadratic function

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{n} a_i \cdot x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \cdot x_i \cdot x_j,$$

we can select the values b_j in such a way that the resulting product $f'(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) \cdot \left(1 + \sum\limits_{j=1}^{n} b_j \cdot x_j\right)$ (or, to be more precise, the quadratic approximation to this product) does not have purely quadratic terms and is, thus, purely bilinear.

Indeed, in this product $f'(x_1, \ldots, x_n)$ the terms coming from multiplying $\sum a_{ij} \cdot x_i \cdot x_j$ and $\sum b_j \cdot x_j$ are cubic and can be thus, in this approximation, safely ignored. Thus, in our quadratic approximation, only the product of $\sum a_i \cdot x_i$ and $\sum b_j \cdot x_j$ must be added. So, in the quadratic approximation, the product function has the following form:

$$f'(x_1, \ldots, x_n) = \sum_{i=1}^{n} a_i \cdot x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \cdot x_i \cdot x_j + \left(\sum_{i=1}^{n} a_i \cdot x_i\right) \cdot \left(\sum_{j=1}^{n} b_j \cdot x_j\right).$$

What are the purely quadratic terms in this expression? For each i, we have two terms proportional to x_i^2: the original term $a_{ii} \cdot x_i^2$ and the new term $a_i \cdot b_i \cdot x_i^2$. Thus, if $a_{ii} + a_i \cdot b_i = 0$ for all $i = 1, 2, \ldots, n$, the resulting expression $f'(x_1, \ldots, x_n)$ will be free of purely quadratic terms – i.e., bilinear.

Thus, it is sufficient to take

$$b_i = -\frac{a_{ii}}{a_i}.$$

Of course, this division is only possible when $a_i \neq 0$ for all i – but this is what is happening in the *generic* case, and this is what we planned to prove – that in the generic case, it is possible to use a purely bilinear discrimination function.

The result is proven.

11 Discussion: A Similar Simplification Is Not Always Possible for Higher Order Models

We have shown that in the second approximation, we can reduce an arbitrary discrimination function to an equivalent bilinear one.

Bilinear terms come from common sense analysis of the system. For higher order terms, a similar common sense analysis enable us to come up with trilinear terms $a_{ijk} \cdot x_i \cdot x_j \cdot x_k$, where i, j, and k are different variables. A natural question is: can we reduce a general cubic expression to a trilinear one?

More generally, we can have multi-linear functions, i.e., functions $f(x_1, \ldots, x_n)$ which are linear in each of their variables x_i. A natural general

question is: can we always reduce an arbitrary discrimination function to a multi-linear one?

The answer to this general equation is "no". Indeed, for any fixed number of variables $x_1, \ldots, x_n$, we only have finite many possible multi-linear terms:

- for a single variable x_1, we can only have one term $a_1 \cdot x_1$;
- for two variables x_1 and x_2, we can only have three term $a_1 \cdot x_1 + a_2 \cdot x_2 + a_{12} \cdot x_1 \cdot x_2$;
- etc.

Thus, for each n, we have a finite-parametric family of multi-linear discrimination functions, i.e., a family that can be characterized by finitely many parameters.

On the other hand, the set of possible class is infinitely-parametric: indeed, each class can be described by a smooth separating line $x_n = F(x_1, \ldots, x_{n-1})$, and a general separating line can be described by a general Taylor expansion and thus, require infinitely many parameters.

So, for general higher order approximations, multi-linear functions are not sufficient.

Are multi-linear functions sufficient at least for cubic terms? The answer is again "no", even for $n = 3$ variables. Indeed, a general trilinear discrimination function of 3 variables has:

- three parameters a_1, a_2, and a_3 describing the linear terms;
- three parameters a_{12}, a_{13}, and a_{23} describing the bilinear terms; and
- a single parameter a_{123} describing the only possible trilinear term

$$a_{123} \cdot x_1 \cdot x_2 \cdot x_3.$$

to the total of $3 + 3 + 1 = 7$ parameters. On the other hand, a general cubic discriminating curve $x_3 = F(x_1, x_2)$, with a general cubic dependence

$$F(x_1, x_2) = b_1 \cdot x_1 + b_2 \cdot x_2 + b_{11} \cdot x_1^2 + b_{12} \cdot x_1 \cdot x_2 + b_{12} \cdot x_2^2 +$$

$$b_{111} \cdot x_1^3 + b_{112} \cdot x_1^2 \cdot x_2 + b_{122} \cdot x_1 \cdot x_2^2 + b_{222} \cdot x_2^3$$

requires $9 > 7$ parameters. Thus, it is not possible to describe a general third-order classification by a trilinear discrimination function.

Comment. A similar analysis can answer the following natural question: we have reduced a general quadratic discrimination function to a bilinear one; can we reduce it further, to some class with even fewer parameters?

To answer this question, let us count how many parameters we need to describe a general bilinear function of n variables, and how many parameters we need to describe a general class in the quadratic approximation.

To describe a general bilinear function of n variables, we need to describe:

- n coefficients a_i, $i = 1, 2, \ldots, n$;
- $\dfrac{n \cdot (n-1)}{2}$ coefficients a_{ij} corresponding to all possible pairs (i, j), $i \neq j$.

The total number of parameters needed for this approximation is

$$n + \frac{n \cdot (n-1)}{2} = \frac{1}{2} \cdot (2n + n^2 - n) = \frac{1}{2} \cdot (n^2 + n).$$

A generic second-order classification can be described by a quadratic expression

$$F(x_1, \ldots, x_{n_1}) = \sum_{i=1}^{n} b_i \cdot x_i + \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} b_{ij} \cdot x_i \cdot x_j$$

describing the separating curve $x_n = F(x_1, \ldots, x_{n-1})$. To describe this general quadratic function, we need:

- $n-1$ parameters b_i, $i = 1, 2, \ldots, n-1$;
- $\frac{(n-1) \cdot (n-2)}{2}$ parameters b_{ij} corresponding to all possible pairs (i, j), $i \neq j$; and
- $n-1$ parameters b_{ii}, $i = 1, 2, \ldots, n-1$.

The total number of parameters needed for this approximation is

$$n - 1 + \frac{(n-1) \cdot (n-2)}{2} + n - 1 = \frac{1}{2} \cdot (2(n-1) + (n^2 - 3n + 2) + 2(n-1)) =$$

$$\frac{1}{2} \cdot (n^2 + n - 2) = \frac{1}{2} \cdot (n^2 + n) - 1.$$

By comparing these two values, we can see that there is only one extra parameter in the bilinear expression – and it can be reduced by setting, e.g., $a_1 = \pm 1$. This reduction can be achieved if we divide the original function $f(x_1, \ldots, x_n)$ by a positive number $|a_1|$ – this division does not change the sign of $f(x_1, \ldots, x_n)$ and thus, leads to an equivalent discrimination function. After this simple reduction, we have exactly as many parameters as we need to describe a general quadratic expression – and thus, no further reduction is possible.

Acknowledgments

This work was supported in part by NASA under cooperative agreement NCC5-209, NSF grants EAR-0225670 and DMS-0532645, Star Award from the University of Texas System, and Texas Department of Transportation grant No. 0-5453.

References

1. Forrester, J. W.: *Principles of Systems*, Pegasus Communications, 1968.
2. Meadows, D. H.: *Limits to Growth*, Signet Publ., 1972.
3. Nguyen, H. T., Kreinovich, V.: *Applications of Continuous Mathematics in Computer Science*, Kluwer, Dordrecht, 1997.
4. von Bertalanffy, L.: *Perspectives on General System Theory*. G. Braziller, New York, 1975.
5. von Bertalanffy, L.: *General System Theory*. G. Braziller, New York, 1984.