24
Statistical Analysis of Complex Systems in Biomedicine

Harry B. Burke, M.D., Ph.D.

New York Medical College
Department of Medicine
Valhalla, NY 10595
914-347-2428 (voice)
914-347-2419 (fax)
burke@unr.edu

ABSTRACT
The future explanatory power in biomedicine will be at the molecular-genetic level of analysis (rather than the epidemiologic-demographic or anatomic-cellular levels). This is the level of complex systems. Complex systems are characterized by nonlinearity and complex interactions. It is difficult for traditional statistical methods to capture complex systems because traditional methods attempt to find the model that best fits the statistician's understanding of the phenomenon; complex systems are difficult to understand and therefore difficult to fit with a simple model. Artificial neural networks are nonparametric regression models. They can capture any phenomena, to any degree of accuracy (depending on the adequacy of the data and the power of the predictors), without prior knowledge of the phenomena. Further, artificial neural networks can be represented, not only as formulae, but also as graphical models. Graphical models can increase analytic power and flexibility. Artificial neural networks are a powerful method for capturing complex phenomena, but their use requires a paradigm shift, from exploratory analysis of the data to exploratory analysis of the model.

24.1 Introduction

In the past, most biomedical phenomena were analyzed at the demographic-epidemiologic or anatomic-cellular levels. Since phenomena at these levels is largely linear or nearly linear, traditional statistical models were very helpful. One result of these analyses is that, today, most biomedical variables are linear or nearly linear variables. But the future will not be like the past. The future explanatory power in biomedicine is at the molecular-genetic level of analysis. This level is characterized by complex systems, i.e., nonmonotonicity and complex interactions. Complex systems are difficult for traditional statistical models to capture because traditional methods require a priori information about the variables in order to represent the variables in the model. Thus, the traditional statistician must "explore" the data, and must explicitly model what is discovered. But exploration and explicit modeling is not always practical at the molecular-genetic level, where there can be twenty or more variables, and where the variables may interact in three-way and higher combinations.

There is evidence that cancer is a complex system and that future prognostic factors will be nonmonotonic and exhibit complex interactions. Cancer is primarily a genetic

[1] *Learning from Data: AI and Statistics V.* Edited by D. Fisher and H.-J. Lenz. ©1996 Springer-Verlag.

disease (Fearon [Fearon90]; Fishel et al. [Fisher93]; Leach et al. [Leach93]) and a complex system. Cancer genes do not act in isolation; oncogenes, suppressor genes, and genetic mutations cause cancer through the complex interaction of the genes and their products (Papadopoulos et al. [Papadopoulos94]; Steel [Steel93]). A cascade of genetic abnormalities is required to produce a cancer (Knudson [Knudson85]; Fearon and Vogelstein [Fearon90]). Thus, it cannot be assumed (1) that a gene or its product will be monotonic or that it will have an independent prognostic value before it is combined with other genes and/or their products, (2) that gene interactions are binary, or (3) that there will only be a few simple genetic interactions. Furthermore, it will probably not be possible to specify in advance of the analysis which complex genetic interactions exist. The need to capture nonmonotonicity and complex interactions exists because the prognostic value of the genetic changes and their products can depend on their nonmonotonic characteristics and interactions (Fearon and Vogelstein [Fearon90]).

24.2 Artificial Neural Networks

Artificial neural networks are a class of nonlinear regression and discrimination statistical methods, and they are of proven value in many areas of medicine (Westenskow et al. [Westenskow92]; Tourassi et al [Tourassi93]; Leong and Jabri [Leong92]; Palombo [Palombo92]; Gabor and Seyal [Gabor92]; Goldberg et al. [Goldberg92]; O'Leary et al. [O'Leary92]; Dawson et al. [Dawson91]; Wu et al. [Wu93]; Astin and Wilding [Astin92]; Weinstein et al. [Weinstein92]). In medical research, the most commonly used artificial neural networks are *feed-forward* networks of simple computing units that use backpropagation training. A feed-forward network is usually composed of three interconnected layers of nodes (computing unit): an input layer, a hidden layer, and an output layer. In our case, each input node corresponds to a patient variable. All nodes after the input layer sum the inputs to them and use a transfer function (also known as an activation function) when they send the information to the adjacent layer nodes. The transfer function is usually a sigmoid function such as the logistic. The connections between the nodes have adjustable weights that specify the extent to which the output of one node will be reflected in the activity of the adjacent layer nodes. These weights, along with the connections among the nodes, determine the output of the network.

Backpropagation consists of fitting the parameters (weights) of the model by a criterion function, usually square error or maximum likelihood, using a gradient optimization method. In feed-forward networks using backpropagation, the error between actual and expected network output units is propagated back from the output through the connections between nodes, in order to adjust the connection weights in the direction of minimum error.

The mathematical representation of a feed-forward network as described above can be viewed as a series of regression equations within a regression equation, where there can be as many regression equations as is necessary to fit the phenomenon. Thus,

$$\begin{aligned} h_j &= f(w^h_{j1}x_1 + w^h_{j2}x_2 + \cdots + w^h_{jn}x_n) \qquad (24.1) \\ o_k &= g(w^0_{k1}h_1 + w^0_{k2}h_2 + \cdots + w^0_{kn}h_n) \qquad (24.2) \end{aligned}$$

equation (24.1) specifies the output of each of the hidden nodes j, f is a nonlinear transfer function, w^h_{ji} is the weight from input unit i to hidden node j, and x_i is the value of an input variable. Equation (24.2) specifies the output or prediction, o_k, of the network, where g is a nonlinear transfer function, w^0_{kj} is the weight of the connection from hidden unit j to output node k, and h_j is the hidden node output. It should be noted that equation (24.2), without equation (24.1) input, is equivalent to logistic regression, where g is the logistic function, w is the beta coefficient, and h is the x covariate. Artificial neural networks with sufficient hidden units can approximate any continuous function to any degree of accuracy (Hornik et al. [Hornik89]; Leshno et al. [Leshno93]).

24.3 Clinical Example

We have compared the prognostic accuracy of the TNM staging system (Beahrs et al, 1992) and an artificial neural network according to five year cancer-specific survival.

Data: We have used three data sets (each of approximately 5,000 cases) in these analyses: two from the American College of Surgeons, the Patient Care Evaluation (PCE) breast cancer and colorectal cancer data sets; the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) breast cancer data set; and the Mayo Clinic prostate cancer data set. The variables in the PCE, SEER, and Mayo data sets are either binary or monotonic. The factors were selected in the past for collection because they were significant in a generalized linear model, e.g., logistic regression. There is no predictive model that can improve upon a generalized linear model when the predictor variables meet the assumptions of the model and there are no interactions.

Accuracy: There are three components to predictive accuracy: the quality of the data, the predictive power of the prognostic factors, and the prognostic method's ability to capture the power of the prognostic factors. This work focuses on the third component. Comparative accuracy is assessed by the area under the receiver operating characteristic (ROC) curve (Hanley and McNeil [Hanley82]). The receiver operating characteristic area varies from zero to one. When the prognostic score is unrelated to survival, the score is .5, indicating chance accuracy. The farther the score is from .5 the stronger the prediction model. Specifically, the TNM staging system's predictive accuracy is determined by comparing (using the area under the ROC curve) its prediction for each individual patient, where the prediction is the fraction of all the patients in that stage who survive, to each patient's true outcome.

Model: The artificial neural network results reported in this paper are based on backpropagation training which uses the maximum likelihood criterion function and the gradient descent optimization method and the "NevProp" software implementation for training. Significant differences in the receiver operating characteristic areas between the TNM staging system and the artificial neural network are tested following Hanley and McNeil [Hanley82]. The training data set (approximately 3,500 cases) is divided into training (approximately 2,000 cases) and stop-training (approximately 1,500 cases) subsets. Training is stopped when accuracy starts to decline on the stop-training data subset. All analyses employ the same training and testing (validation) data sets, and all results are based on the one time use of the testing data sets.

Results: A comparison of the accuracy of the TNM staging system and the artificial neural network (Table 24.1) using the PCE breast cancer data set, which examines breast cancer-specific five-year survival accuracy for only the TNM variables, demonstrates that

TABLE 24.1. Comparison of TNM staging system and artificial neural networks (all comparisons are for five year, cancer-specific survival).

DATA SETS	TNM	ANN
PCE 1983 Breast Cancer - TNM v	.720	.770
PCE 1983 Breast Cancer - 54 v	.720	.784
SEER 1977 Breast Cancer - TNM v, 10 yr	.692	.730
PCE 1983 Colorectal Cancer - TNM v	.737	.815
PCE 1983 Colorectal Cancer - 87 v	.737	.869
Mayo Clinic Prostate Cancer	.563	.811

the artificial neural network's predictions are significantly more accurate (TNM .720 vs. ANN .770, $p < .001$). Adding 51 commonly collected variables to the TNM variables further increases the accuracy of the artificial neural network (.784). Extending these results to the SEER breast cancer data set, with a breast cancer-specific ten year survival endpoint; using only the TNM variables, the artificial neural network's prognostic accuracy is significantly greater than the TNM staging system (TNM .692 vs. ANN .730, $p < .01$).

Using the PCE colorectal data sets, the predictive accuracy of the two methods can be compared. For only the TNM variables, the artificial neural network's prognostic accuracy is significantly greater than the TNM stage model in predicting colorectal-specific five year survival (TNM .737 vs. ANN .815, $p < .001$). Adding 84 commonly collected factors to the TNM variables further increases the accuracy of the artificial neural network (.869).

The Mayo Clinic data set demonstrates that, for prostate cancers represented in its corpus, the TNM staging system has a low prognostic accuracy (.563), and that the artificial neural network, with other commonly collected variables, is significantly more accurate (TNM .563 vs ANN .811, $p < .001$).

24.4 The Future of Artificial Neural Networks

To demonstrate the power of the artificial neural network to capture unanticipated nonmonotonicities and complex interactions, a constructed nonmonotonic variable is added to the 54 PCE breast cancer variables. The constructed nonmonotonic variable consists of two normal distributions centered at zero, one having a standard deviation of 1 for patients who are alive at five years, the other having a standard deviation of 10 for patients who are dead by five years. If the artificial neural network cannot capture nonmonotonicity without a priori specification of the phenomena, then its accuracy should remain at .770 with the TNM variables and .784 with the 54 variables, on the test set. The artificial neural network does capture the predictive power of the nonmonotonic factor, and its accuracy increases to .948 with the TNM variables and to .961 with the 54 variables, on the test set (Table 24.2)

A constructed complex three-way interaction is added to the 54 PCE breast cancer variables. The artificial neural network captures the informative three-way interaction, from among the 29,260 possible three-way interactions, its accuracy increases from .784 to .942 on the test data set. It is the case that anticipated nonmonotonicity has, with varying degrees of success, been modeled by classical prediction models. Although it is computationally intensive, classical prediction models can test for a predictive three-way interaction among 29,260 possibilities, but it is not clear how they would discover four-way and higher interactions of nonlinear variables. It can be concluded that artificial neural networks are powerful models; they can capture the explanatory power inherent in complex systems.

TABLE 24.2. PCE 1983 Breast Cancer Data: 5 year Survival Prediction Accuracy, nonmonotonic variable added or three-way interaction added.

PREDICTION MODEL	nonmonotonic variable TNM variables accuracy*	 55 variable accuracy*	three way 57 variable accuracy*
pTNM Stages	.720	.720	.720
Stepwise Logistic Regression	.762	.776	.776
Backpropagation ANN	.948	.961	.942

* The area under the curve of the receiver operating characteristic.

At the present time, using variables selected by traditional statistical methods, it is not required that an artificial neural network be more accurate than traditional statistical methods in order for it to be an appropriate statistical method for cancer prediction. Artificial neural networks can be recommended for cancer prediction because: (1) they are as accurate as the best traditional statistical methods (results not presented), (2) they are able to capture complex phenomena (e.g., nonmonotonicity and complex interactions) without a priori knowledge, and (3) they are a general regression method, therefore, if the phenomenon is not complex, so that accuracy can be maintained using a simpler model, artificial neural networks can be reduced to simpler models resulting in simpler representations.

There are several possible objections to artificial neural networks, including: (1) they require an analysis of the model. They capture phenomena without requiring prior exploration of the data, but they require exploration of the model. More will be said regarding model analysis later in the paper. (2) Some believe that artificial neural networks are overparameterized because they can have a large number of weights. Overfitting can be prevented by keeping the weights small, thereby reducing the effective number of degrees of freedom. This can be accomplished by penalizing large weights, or stopping the iterative fitting algorithm before the weights have grown to their full size. It is often the case that, when one of these methods is used, predictive accuracy is better than it would be if we used a smaller model and fit the data without restriction. When a method is used that reduces the weights that are not being increased by the input variables, the weights to the hidden layer shrink, and when there are only linear relationships present, as the hidden layer weights approach zero the neural network approximates a generalized linear model.

(3) It is thought that artificial neural networks are less "transparent" (the importance of the variables is less obvious), than traditional statistical models. This view of transparency fundamentally misunderstands the situation. Artificial neural networks are as complex as is necessary to capture the phenomenon. Generally, if the phenomenon is complex, the model must be complex. If the phenomenon is simple enough to be captured by

simple models, then artificial neural networks can be reduced to a simple model, and the importance of covariates is easily observed. For example, if the phenomenon is linear, then a two layer (no hidden layer) artificial neural network with linear transfer functions, is mathematically identical to linear regression, and the weights of the artificial neural network are identical to the beta coefficients of the linear regression model. Therefore, model transparency (i.e., ease of variable interpretation) is properly understood as a function of complexity and accuracy. For a simple phenomena, a properly chosen simple model is easily interpretable. For a complex phenomenon (e.g., complex interactions) and a properly chosen model, increases in model complexity result in increases in accuracy if overfitting is avoided. Increases in model complexity reduce the transparency of both traditional statistical models and artificial neural network statistical models.

24.5 Domain Knowledge and Model Knowledge

Domain knowledge is information regarding the phenomena acquired by examining the data, and model knowledge is information regarding the phenomena acquired by examining the empirically derived model. Both are required for understanding phenomena, but their relative importance in the overall analysis can differ. In traditional statistics domain knowledge is the dominant approach. The statistician talks to the researcher, who suggests where the statistician can explore the data. The statistician then examines the data, finds the best fitting model, and performs inferential calculations to determine variable significance and importance. But this is not the only possible approach to understanding phenomena. Selection of the best (most accurate predictions) model can be based on either knowing the relationships between the predictors and the relationship of the predictors to the phenomenon, and selecting the model that best captures these relationships, or, if the relationships are not known a priori, selecting a model that is capable of capturing any relationships. This latter approach, selecting the model that can capture any phenomena, is very different from the traditional approach. It requires that the model be explored rather than the data. The relationships are captured in the model, and one decomposes the model in order to discover the phenomena.

It should be noted that, in terms of models, there is no difference between prediction and classification. Prediction and classification differ in the questions being asked, i.e., the character of the data and the type of outcome. Thus robotic control, from a model-theoretic perspective, does not differ from cancer-outcome prediction: vision is classification and movement is prediction.

There are some aspects of the analysis of a phenomena that are domain specific, and some that are model specific. For example, the number of hidden layer nodes (i.e., subregression equations) is domain specific. The number of hidden layer units cannot be determined, a priori, by any analytic method because the number of units depends on the complexity of the phenomena; a simple phenomena requires no hidden units, a complex may require five or ten hidden units.

An example of model analysis is the determination of whether the phenomenon exhibits nonmonotonicities or complex interactions. The approach is to compare the results of a two-layer neural network with those of a three-layer neural network. If the two-layer is significantly less accurate than the three-layer neural network, then there are nonmonotonic relationships, interactions, or both. If there is no difference between the two models, then the model can be simplified. If there is a difference, then a complex (three-layer) model must be used to capture the nonlinearities and interactions.

For simple phenomena, e.g., phenomena that do not require the use of a hidden layer in an artificial neural network, artificial neural networks are as transparent as other statistical models. For complex models sensitivity analysis can determine the contribution of input variables in the artificial neural network prediction (Intrator [Intrator93]). But sensitivity analysis is not adequate because complex relationships, represented by complex mathematical equations, are not easily understood. To understand these complex relationships visual models are needed. Buntine [Buntine94] points out "Graphical operations manipulate the underlying structure of a problem unhindered by the fine detail of the connecting functional and distributional equations. This structuring process is important in the same way that a high-level programming language leads to higher productivity over assembly language."(p 160) The ability to represent and manipulate artificial neural networks, in terms of graphical models, provides power and flexibility in model analysis.

24.6 References

[Astin92] Astin ML, & Wilding P. (1992) Application of neural networks to the interpretation of laboratory data in cancer diagnosis. Clin Chem 1992;38:34-38.

[Beahrs92] Beahrs OH, Henson DE, Hutter RVP, & Kennedy BJ. (1992) Manual for staging of cancer, 4th ed. Philadelphia: JB Lippincott, 1992.

[Buntine94] Buntine WL. (1994) Operations for learning with graphical models. J Art Intell Res 1994;2:159-225.

[Dawson91] Dawson AE, Austin RE, & Weinberg DS. (1991) Nuclear grading of breast carcinoma by image analysis. J Clin Pahol 1991;95(Suppl):S29-S37.

[Fearon90] Fearon ER, & Vogelstein B. (1990) A genetic model for colorectal cancer. Cell 1990;61:759-67.

[Fisher93] Fishel R, Lescoe MK, & Rao MRS et al. (1993) The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. Cell 1993;75:1027-36.

[Gabor92] Gabor AJ, & Seyal M. (1992) Automated interictal EEG spike detection using artificial neural networks. Electroencephalogr Clin Neurophysiol 1992;83:271-80.

[Goldberg92] Goldberg V, Manduca A, & Ewert DL. (1992) Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence. Med Phys 1992;19:1275-81.

[Hanley82] Hanley JA, & McNeil BJ. (1982) The meaning of the use of the area under the receiver operating characteristic (ROC) curve. Radiology 1982; 143:29-36.

[Hornik89] Hornik K, Stinchcombe M, & White H. Multilayer feedforward networks are universal approximators. Neural Networks 1989;2:359-66.

[Intrator93] Intrator O, Intrator N. (1993) Neural networks for interpretation of nonlinear models. Proceedings of the Statistical Computing Section, American Statistical Society, San Francisco CA; 1993:244-9.

[Knudson85] Knudson AG Jr. (1985) Hereditary cancer, oncogenes, and antioncogenes. Cancer Res 1985; 45: 1437-43.

[Leach93] Leach FS, Nicolaides NC, & Papadopoulos N et al. (1993) Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. Cell 1993;76:1215-25.

[Leong92] Leong PH, & Jabri MA. (1992) MATIC - an intracardiac tachycardia classification system. PACE 1992;15:1317-31.

[Leshno93] Leshno M, Lin VY, Pinkus A, & Schocken S. (1993) Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Networks 1993;6:861-67.

[O'Leary92] O'Leary TJ, Mikel UV, & Becker RL. (1992) Computer-assisted image interpretation: use of a neural network to differentiate tubular carcinoma from sclerosing adenosis. Modern Pathol 1992;5:402-5.

[Palombo92] Palombo SR. (1992) Connectivity and condensation in dreaming. J Am Psychoanal Assoc 1992;40:1139-59.

[Papadopoulos94] Papadopoulos N, Nicolaides NC, Wei Y, & et al. Mutation of a mutL homolog in hereditary colon cancer. Science 1994; 263:1625-29.

[Steel93] Steel M. (1993) Cancer genes: complexes and complexities. Lancet 1993;342:754-5.

[Tourassi93] Tourassi GD, Floyd CE, Sostman HD, & Coleman RE. (1993) Acute pulmonary embolism: artificial neural network approach for diagnosis. Radiology 1993;189:555-58.

[Weinstein92] Weinstein JN, Kohn KW, Grever MR, Viswanadham VN, Rubenstein LV, & Monks AP. (1992) Neural computing in cancer drug development: predicting mechanism of action. Science 1992;258:447-51.

[Westenskow92] Westenskow DR, Orr JA, & Simon FH. (1992) Intelligent alarms reduce anesthesiologist's response time to critical faults. Anesthesiology 1992;77:1074-9.

[Wu93] Wu Y, Giger ML, Doi K, Vyborny CJ, Schmidt RA, & Metz CE. (1993) Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. Radiology 1993;187:81-87.

25

Learning in Hybrid Noise Environments Using Statistical Queries

Scott E. Decatur

Aiken Computation Laboratory
Division of Applied Sciences
Harvard University
Cambridge, MA 02138

ABSTRACT We consider formal models of learning from noisy data. Specifically, we focus on learning in the *probability approximately correct* model as defined by Valiant. Two of the most widely studied models of noise in this setting have been *classification noise* and *malicious errors*. However, a more realistic model combining the two types of noise has not been formalized. We define a learning environment based on a natural combination of these two noise models. We first show that hypothesis testing is possible in this model. We next describe a simple technique for learning in this model, and then describe a more powerful technique based on statistical query learning. We show that the noise tolerance of this improved technique is roughly optimal with respect to the desired learning accuracy and that it provides a smooth tradeoff between the tolerable amounts of the two types of noise. Finally, we show that statistical query simulation yields learning algorithms for other combinations of noise models, thus demonstrating that statistical query specification truly captures the generic fault tolerance of a learning algorithm.

25.1 Introduction

An important goal of research in machine learning is to determine which tasks can be automated, and for those which can, to determine their information and computation requirements. One way to answer these questions is through the development and investigation of formal models of machine learning which capture the task of learning under plausible assumptions.

In this work, we consider the formal model of learning from examples called "probably approximately correct" (PAC) learning as defined by Valiant [Val84]. In this setting, a learner attempts to approximate an unknown target concept simply by viewing positive and negative examples of the concept. An adversary chooses, from some specified function class, a hidden $\{0,1\}$-valued target function defined over some specified domain of examples and chooses a probability distribution over this domain. The goal of the learner is to output in both polynomial time and with high probability, an hypothesis which is "close" to the target function with respect to the distribution of examples. The learner gains information about the target function and distribution by interacting with an example oracle. At each request by the learner, this oracle draws an example randomly according to the hidden distribution, labels it according to the hidden target function, and returns the labelled example to the learner. A class of functions $\mathcal{F}$ is said to be PAC learnable if

[1] *Learning from Data: AI and Statistics V*. Edited by D. Fisher and H.-J. Lenz. ©1996 Springer-Verlag.