Application of Connectionist Learning Methods to Manufacturing Process Monitoring

Judy A. Franklin
Richard S. Sutton
Charles W. Anderson

GTE Laboratories Incorporated
40 Sylvan Road
Waltham, Ma. 02254

1 Introduction

One of the projects in the Machine Learning Department at GTE Laboratories involves studying quality control of a fluorescent bulb manufacturing line. All manufacturing processes are subject to incompletely understood changes due to variations in raw materials, environmental factors such as weather, wearing and aging of the machinery, and changes in operators. This can result in marked variations in yield, quality, and rejection rates. These problems can occur even in the most mature and established manufacturing processes.

Our goal is *Computer Integrated Manufacturing (CIM)*, the autonomous computer-based monitoring of plant behavior, determination of causal influences, and, ultimately, adaptive control of the plant process. The first stage is process monitoring with respect to yield and other performance measurements. By using past experience to find correlations between approximately one hundred sensory measurements, we will determine which process variables most affect quality.

Two approaches are being compared. One employs standard statistical procedures to find correlations between sensor measurements and quality. The sensor data from the production line are collected over a period of time and correlations are made *off-line* at infrequent intervals using analyses such as linear regression. The second approach is to estimate the correlations incrementally, as the data are collected, on-line and in real-time. The estimates are updated incrementally using *connectionist learning procedures*.

2 Connectionist Learning Methods

The exploration and development of incremental connectionist learning methods is the focus of our work on this project. Connectionist models consist of "neuron-like" processing "elements (units)" that interact and form a network via weighted connections. The "state" or "activity level" of each unit is determined by the input received from the other units through the connections and from inputs received from the environment. One goal of connectionist research is to discover efficient learning procedures that allow multi-layered networks of these units to construct an internal representation of the environment (Hinton, 1987 [4]; Barto and Anderson, 1985 [2]). For example, in our application, a network of units could be used to construct a representation of the manufacturing process that is conducive to finding dependencies between sensor values and quality measurements. For the initial experiments reported here, we focus on one-layer networks, reserving multi-layered networks for later study.

Connectionist learning procedures can be viewed as *incremental* methods for computing standard statistical quantities. Figure 1 shows a one-layer *ADALINE* network [10] (see section 4) that computes essentially the same quantities as a linear regression; that is, independent variables (the sensor measurements) are numerically related to dependent variables (factory yield and other performance and quality measures). In the limit, both techniques produce exactly the same numbers (Widrow and Stearns, 1985 [11]). The difference is that the adaptive network processes the data *incrementally*, whereas linear regression is a batch process. In this paper, we present results comparing the performance if these two methods. Conventional CIM approaches suffer from several problems that can potentially be solved by connectionist approaches. The primary problems are computational complexity and a limited ability to deal efficiently with nonlinearities.

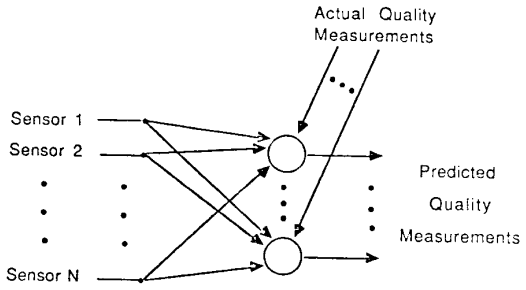

Figure 1: A one-layer network of ADALINES used in manufacturing process identification.

The processing time of the ADALINE network increases only *linearly* with the number of independent variables (sensors); the processing required per time step is $O(n)$, where n is the number of sensors. The total processing required by linear regression is, on the other hand, approximately $O(n^{\log 7}) \approx O(n^{2.81})$.[1] As will be discussed below, even the most incremental implementation of linear regression requires at least $O(n^2)$ processing per time step. The learning network thus offers a savings of at least one factor of n. We expect n to eventually be about 150 in our application. The manufacturing line with which we are working generates a new set of independent and dependent sensor measurements every 5 minutes.

An additional advantage of the ADALINE network is that it can process each 5 minutes worth of data as it is generated, while conventional regression techniques require all the data to be collected for days or weeks, and then processed all at once as a batch. By completely processing all data as it arrives, there is no accumulating buildup of past data in need of processing, a major computational stumbling block of "batch" linear regression. For hundreds of sensors, these computational differences can have a tremendous effect: the network could be implemented on a much smaller computer, or it could be used with many more sensors, or more frequently sampled sensors.

The reduced computational complexity of the connectionist network approach also allows more freedom in the choice of the model used to predict outcomes and correlations. With conventional methods, nonlinear relationships among measurables would require a prohibitive amount of additional processing or human intervention to select a small number of such relationships for consideration.

[1] This is based on using the method of V. Strassen (Strassen, 1969, or see, e.g., Aho et al., 1974), the best known method for problems of this size.

0-8186-2012-9/89/0000/0709$01.00 © 1989 IEEE

Connectionist networks have been shown to be able to learn complex nonlinear functions (Rumelhart, Hinton, Williams, 1985 [5]; Sejnowski and Rosenburg, 1986 [6]) with limited computational resources. We note that the efficiency of network approaches has yet to be compared with that of conventional nonlinear regression approaches and we have not yet carried out such a comparison ourselves. Nevertheless, we are already optimistic about the potential performance of the incremental methods because of their computational advantages. To a large extent, more complex relationships can be handled simply by adding more interaction terms, such as pairwise products, to the input vector. This increases the effective n for the various techniques. Since the incremental connectionist methods are of order n more computationally efficient, they should be able to consider far more such interaction terms then the conventional methods.

3 The Manufacturing Line

Our fluorescent bulb manufacturing line is a cascade of dozens of processes and is highly nonlinear. Figure 2 shows the line being monitored by a learning system. Each stage shown is itself composed of multiple complex processes.

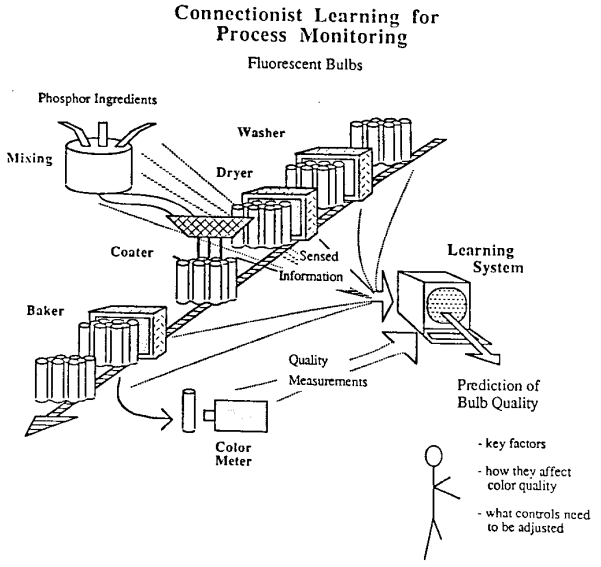

Figure 2: Adaptive Networks for Process Monitoring

While our goal is to compare these algorithms on an actual fluorescent bulb line, we have constructed a simulator to obtain the results presented in this paper. The stages included in the simulation are the bulb washer and dryer, the phosphor mixer, the coater, the four-stage dryer, and the baker. For much of the manufacturing process, small bundles of several dozen bulbs are processed as a unit, and our simulation reflects this level of detail throughout. Bundles of bulbs are "passed through" each stage and the quality of coating is observed at the end of the final stage. Rules in each stage affect coating quality of each bundle of bulbs while it is within that stage. These rules coarsely emulate the observations of the plant experts with whom we have consulted. Sensors are also simulated in every stage. The values of the variables that the sensors measure are determined by rules specifying their interrelations and also by random effects. Some variables are following random walks; some are tied to the time of the day or to other variables; most are a combination. For example, wash water temperature affects coating quality and depends on the time of day and the environmental air temperature, both of which are "sensed" independent variables.

Figure 2 shows the gathering of sensed information from the various stages along with quality measurements taken at the final stage of the process. These data are processed by the learning system. The system predicts bulb quality and compares this prediction to the quality measurement in order to improve its predictive capability. A human observing the output of the learning process may use the learned correlations between sensor values and predicted bulb quality to adjust controls on the process line, as in conventional process control methods. However, our experiments involve only the discovery of such correlations. Also, we are not presently predicting color quality (as is shown in the figure), but percentage of defects in the phosphor coating of the bulbs.

Results have been obtained from the simulation of the fluorescent line using an ADALINE and also a new, faster procedure called the NADALINE (Sutton, 1988 [9]). We compare these methods with conventional linear regression and show the results of this comparison.

4 The Algorithms

4.1 ADALINE

The ADALINE (ADAptive LInear Neuron) consists of a time-indexed vector of real-valued parameters or weights,

$$W(k) = [w_0(k), w_1(k), \cdots w_n(k)]^T \tag{1}$$

that multiplies an input vector,

$$X(k) = [x_0(k), x_1(k), \cdots x_n(k)]^T \tag{2}$$

to form a weighted sum:

$$y(k) = W^T(k)X(k) \tag{3}$$

$$= \sum_{i=0}^{n} w_i(k)x_i(k) \tag{4}$$

where k is the discrete time index. The $x_i(k)$ are the n sensor measurements, except that $x_0(k) \equiv 1$; the corresponding weight, w_0, plays the role of the "threshold" or "bias" in some other connectionist models. We may have, for example, $x_5(k) = 150$ degrees, the temperature of the bulb wash water at time step k. The resulting output signal, $y(k)$ is the ADALINE's estimate of the correct or desired output. The desired output, $d(k)$ is supplied externally and is used to form an error $e(k) = d(k) - y(k)$. This error is used to adjust the values of the weights in the following way:

$$w_i(k+1) = w_i(k) - \alpha e(k)x_i(k) \tag{5}$$

for every $i = 0, 1, 2, \cdots, n$, and α a positive learning rate parameter. A thorough analysis of the theory of this algorithm may be found in [11].

4.2 The NADALINE

The NADALINE, or Normalized ADALINE, is the same algorithm as the ADALINE with two differences. The most important difference is that the sampled inputs, x_i, are *normalized*:

$$x'_i(k) = \frac{x_i(k) - \mu_i(k)}{\sigma_i(k)} \tag{6}$$

where $\mu_i(k)$ is the mean of all of the values of $x_i(k)$ up to the present time k, and $\sigma_i^2(k)$ is their variance. Equations 4 and 5 describe the NADALINE with the $x_i(k)$ replaced by the normalized $x'_i(k)$. The second difference is that equation 5 does not apply for $w_0(k)$, the weight corresponding to the input that is always 1. Instead, $w_0(k)$ is set to the mean of all the values of d up to but not including the k^{th} value.

In experiments on other problems, these two minor changes have been found to result in significant reductions in time to learn, typi-

cally by an order of magnitude or more [9]. The process monitoring application is of the sort in which normalization would be expected to help substantially, because it involves a wide variety of sensed variables with widely varying means and variances. The theory of the NADALINE is discussed further by Sutton [9].

4.3 Linear Regression

Conventional linear regression is a technique for computing the weight vector W^* that minimizes the mean square error:

$$\frac{1}{k}\sum_{j=1}^{k}(d(j) - W^{*T}X(j))^2 \tag{7}$$

In a sense this represents the optimal solution. However, that perspective is based on assumptions such as stationarity, statistical independence, and noise models that rarely strictly hold in real applications. Thus, it is possible for other techniques to perform better that linear regression, as we in fact found for the NADALINE in the results discussed below.

The linear regression algorithm we used was as follows. Let D be the k-vector of all desired responses $d(k)$ seen up to time k. Let Z be the $n \times k$ matrix whose columns are all the X vectors seen up to the current time k. The minimum mean square weight vector is then given by:

$$W^* = (Z^T Z)^{-1} Z^T D \tag{8}$$

The complexity of this algorithm is $O(n^2)$ in space and $O(kn^2 + n^{2.81})$ in computation. Other implementations can reduce the computation to $O(kn^2)$, that is, one factor of n more complex than the ADALINE or NADALINE. Once regression is performed, W^* is used to find the predicted value, $y(k)$, of the independent variable, $d(k)$ (say, bulb coating quality), given the vector of sampled values of the independent variables $X(k)$ via $y(k) = W^{*T}X(k)$ (i.e., just as in the ADALINE, excepting W^*). A reference providing the details is Draper and Smith [3].

5 Experimental Results

To compare the algorithms, we ran the simulation for eight days of simulated time, taking data samples every fifteen minutes. The first seven days of data was used as a training set, and the eighth day's as a testing set. All algorithms used exactly the same training and testing set. For linear regression, the first seven days of data was stored and then used to compute W^*. For the ADALINE and NADALINE, the weight vector was updated incrementally for the first seven days and then held constant during the eighth day. A comparison of the actual and predicted coating quality over the eighth day for each of the three algorithms is shown in Figures 3, 4, and 5. The average prediction errors over the eighth day of the algorithms were linear regression, 0.244, ADALINE, .737, and NADALINE, .259 (in percentage of coating defects per bundle of bulbs). We see that linear regression and the NADALINE performed well both in predicting short term changes and in following the general trend. The ADALINE could only predict the average level of coating defects given seven days worth of data. We hypothesize that its inability to learn was caused by the widely different variances in the values of the sensor readings. The learning rate for the NADALINE was $\alpha = .01$ and that of the ADALINE, $\alpha = .00001$. Without normalization, a small learning rate is required in order to be stable. In fact, for a value of $\alpha = .0001$ the ADALINE was unstable. For this reason, in order for the ADALINE's performance to converge to that obtained by linear regression, more training data would need to be presented.

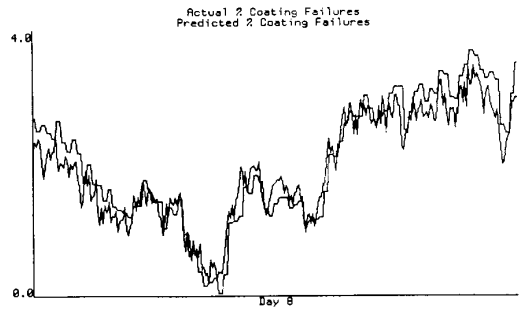

Figure 3: Linear regression: Predicted and actual quality versus time for one day after a training period of one week.

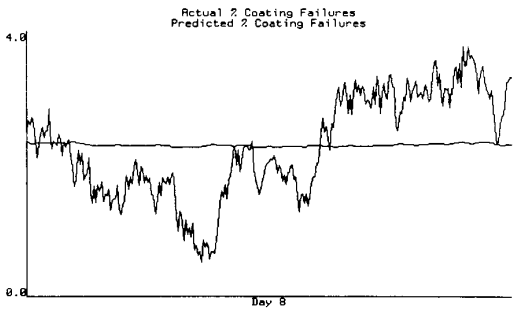

Figure 4: ADALINE: Predicted and actual quality versus time for one day after a training period of one week.

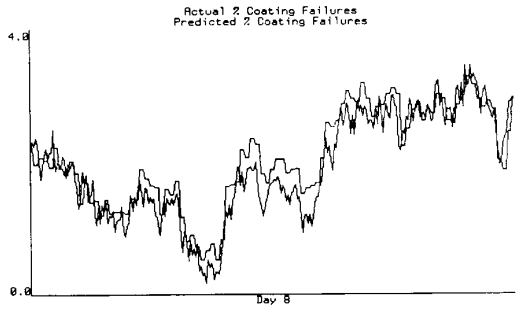

Figure 5: NADALINE: Predicted and actual quality versus time for one day after a training period of one week.

6 Conclusions

We have shown in simulation that connectionist learning networks can monitor manufacturing processes to determine causal relationships with an accuracy competitive with that of conventional statistical techniques. Moreover, the network operates on-line, in real-time, and with substantial savings in computational complexity as compared with conventional CIM techniques. Our comparisons have been between single-layer. linear networks and linear regression. Similar comparisons could be made between multi-layer nonlinear networks and various forms of nonlinear regression. While one of the most attractive features of connectionist networks is their ability to handle nonlinearity, our results suggest that even in the simpler case they can offer significant advantages.

7 Future Work

Multi-layer learning networks, such as back-propagation networks (Rumelhart, Hinton, and Williams, 1985), could similarly be compared with nonlinear incremental regression methods. More complex connectionist networks that are able to learn nonlinear mappings will produce more accurate correlations than linear methods. The ability to learn nonlinear mappings is necessary in modelling complex manufacturing process. This is an especially important point in light of the application that is our focus.

Another important area of study is that of the temporal aspects of causal relationships. The effect of a variable in an early stage of the manufacturing process may not be evident until the quality is checked at the very end of the process. This means that there is a temporal delay in observed cause and effect, and the length of this delay is probably not known. All the methods discussed here, both connectionist and conventional, assume that all data samples are selected independently, and that all causal influences have their effect within a single sampling interval; the methods all ignore the temporal relationships between samples. However, since these relationships do provide significant additional information about the causal structure of the manufacturing process, this information can in principle be used to form better estimates. Temporal-difference learning methods (Sutton,1988 [9]) are one simple way of extending the connectionist techniques discussed here to take advantage of the information contained in the temporal sequencing of observations. These methods can also be used to account for delays of varying length between cause and effect.

Finally, we would like to consider the use of the correlations that are drawn in the monitoring stage for closed loop control of the manufacturing line. In a sense this is done now by humans in a heuristic manner that has evolved with the expertise of the plant operators. A preliminary step will be to supply the operators with the correlations we obtain from applying our techniqeus to process monitoring. They may be able to use this information on-line to adjust controls in the manufacturing line and then report the ways that they found these correlations to be useful in improving the quality of the line.

8 Acknowledgements

Many thanks go to GTE employees John Doleac, Bob DaSilva, Oliver Selfrige, Paul Feltri, Dominic Checca, Niru Patel.

REFERENCES

[1] Aho, A.V., Hopcroft, J.E., Ullman, J.D. *The Design and Analysis of Computer Algorithms*, Reading, Massachusetts: Addison-Wesley, 1974

[2] Barto, A.G. & Anderson, C.W., "Structural Learning in Connectionist Systems," in Proceedings of the Seventh Annual Conference of the Cognitive Science Society, 43-53, Irvine, Ca., 1985

[3] Draper, N.R. & Smith, H., *Applied Regression Analysis*, 1966, New York, New York: John Wiley and Sons, Inc.

[4] Hinton, G.E., "Connectionist Learning Procedures," Technical Report CMU-CS-87-115, 1987, Computer Science Department, Carnegie-Mellon University, Pittsburgh, Pa., to appear in *Artificial Intelligence*

[5] Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1985), *Learning internal representations by error propagation*, (Institute for Cognitive Science Technical Report 8506), 1985, La Jolla, Ca: University of California, San Diego. Also in D.E. Rumelhart, & J.L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations*, Cambridge, MA: MIT Press.

[6] Sejnowski, T.E. & Rosenburg, C.R., "Parallel Networks that Learn to Pronounce English Text," *Complex Systems*, 1987, *1*, 145-168

[7] Strassen, V., "Gaussian elimination is not optimal," *Numerische Mathematik 13*, 1969, 354–356.

[8] Sutton, R.S., "Learning to Predict by the Methods of Temporal Differences," 1988, *Machine Learning*, Vol. 3, 9-44

[9] Sutton, R.S., "NADALINE: A Normalized Adaptive Linear Element that Learns Efficiently," Technical Report TR88-509.4, 1988 GTE Laboratories Incorporated, Waltham, Ma., submitted to the Second IEEE Conference on Neural Information Processing Systems - Natural and Synthetic

[10] Widrow, B., & Hoff, M.E., "Adaptive switching circuits," 1960, *1960 WESCON Convention Record Part IV*, 96–104.

[11] Widrow, B., & Stearns, S.D., *Adaptive Signal Processing*, 1985, Englewood Cliffs, NJ: Prentice-Hall.