Chapter 13

Concept Learning Rates and Transfer Performance of Several Multivariate Neural Network Models

Patrick Suppes
Lin Liang
Stanford University

The purpose of this chapter is to present results for several related learning models, which may be useful both in machine learning and in the analysis of human learning. One motivation has been our desire to develop efficient and powerful methods of concept learning to use in our work on machine learning of natural language (Suppes, Böttner, & Liang, 1995; Suppes, Liang, & Böttner, 1992). In this earlier work, concepts were assumed known and the learning concentrated on language. It is obviously important to combine both concept and language learning for many kinds of situations. In unpublished research we have already used the learning models proposed here in the early stages of robotic concept and language learning.

The models proposed are compared in two major ways: their comparative rates of concept learning, and their comparative transfer performance when the learning of one concept is followed by another. The several models are compared on six sets of data.

1. The well-known Edgar Anderson data on three species of iris. Four measurements are provided on each sample. The data were first made a prominent subject of statistical analysis by Fisher (1936).
2. Measurement data on genuine and forged Swiss bank notes (Flury & Riedwyl, 1988).
3. Measurement data on three species of beetles (Lubischew, 1962).
4. Randomly generated data on an artificial, but interesting, problem of classification. It is that of distinguishing when an observation,

consisting of two "features," the x- and y-coordinates of a point in the plane, lies in a circle of radius r, or in the region bound by this circle and a circle of radius r', with $r' > r$.

5. The classical problem of learning the binary sentential connective exclusive or, usually written XOR.
6. Transfer data from experiments reported in Suppes (1965) on young children learning identity and equipollence of sets.

The first section that follows presents the six models and related theory, beginning with a multivariate normal learning model. The second section introduces a model for comparatively evaluating the performance of the several learning models. The third presents and analyzes the learning results for the several models of data sets 1–5 just described. The fourth section is devoted to problems of transfer, and in particular to data set 6. The final section compares the models and results to a variety of other models discussed in the literature.

SIX ALTERNATIVE LEARNING MODELS

Why so many models? The large number is a reflection of our conviction that we are still a considerable distance from understanding in any very deep way which learning models are most appropriate for which situation. We are especially skeptical of there being one universally best choice. On the other hand, all six of the models are, broadly speaking, neural network models for learning concepts, or, in an equivalent statistical language, learning classifications based on input vectors of feature data.

Thus each model has feature inputs $x_1, \ldots, x_h$, where each feature x_i may be numerical or only binary. The number g of outputs or classification responses is always taken to be finite. Moreover, the learning is always supervised, with the unique correct response being used to update each model after each trial.

Model I: Multivariate Normal Model

The basic assumption of this model is that the features of instances of a concept have a multivariate normal distribution. Surprisingly, this widely used statistical model has been little studied as a learning model, and its learning rate has not been compared to that of alternatives.

First, we make the Bayesian assumption of a prior distribution $\pi = \pi_1, \ldots, \pi_g$ on the classification responses assigning an object with the features $\mathbf{x} = (x_1, \ldots, x_h)$ to one of the classes or groups.

To obtain the posterior probability of each class 1, . . . , g, we use a uniform prior as a special case of a Dirichilet prior and then compute a Dirichilet-like posterior probability $p_{j,n}$ with parameters α and β for class j on trial n as

$$p_{j,n} = (n_j + \alpha\beta^n)/(\sum_{j=1}^{g} n_j + g\alpha\beta^n) \tag{1}$$

where $\alpha > 1$ and $0 < \beta < 1$, n_j is the number of members of class j observed through trial n, and so in our setup $\Sigma\, n_j = n$. For a general prior π, Equation 1 becomes

$$p_{j,n} = (n_j + \pi_j\alpha\beta^n)/(\sum_{j=1}^{g} n_j + \alpha\beta^n) \tag{2}$$

Second, we assume that on trial n the group-conditional distribution $f_{j,n}$ of any object, that is, feature vector $\mathbf{x}$, is given by

$$f_{j,n}(\mathbf{x};\mathbf{m}_{j,n}, \Sigma_{j,n}) = (2\pi)^{-h/2}|\Sigma_{j,n}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mathbf{m}_{j,n})'\Sigma_{j,n}^{-1}(\mathbf{x} - \mathbf{m}_{j,n})\right\} \tag{3}$$

where $\mathbf{m}_{j,n}$ is the vector of conditional feature means for category response j at the beginning of trial n, $\Sigma_{j,n}$ is the corresponding feature covariance matrix for category j and trial n, $\mathbf{x}'$ is the transpose of vector $\mathbf{x}$, and h is the number of features (McLachlan, 1992, p. 53). The recursive learning rules for updating on each trial the vector of sample means $\mathbf{m}_j$ and the sample covariance matrix $\mathbf{s}_j$ are the following:

Incremental Computation of Mean. The number m_{ij} is the sample mean of feature x_i for category j. It is unchanged when a category is not the correct response. Incremental computation of m_{ij} [a connection from the feature i to the category j at the $(n + 1)$th step] is as follows. According to the definition we have

$$\begin{aligned} m_{ij\,n+1} &= \frac{\sum_{k=1}^{n+1} \delta_{j\,k} x_{i\,k}}{\sum_{k=1}^{n+1} \delta_{j,k}} \\ &= \frac{\delta_{j,n+1}x_{i,n+1} + \sum_{k=1}^{n} \delta_{j,k}x_{i,k}}{\sum_{k=1}^{n+1} \delta_{j,k}} \end{aligned} \tag{4}$$

Using the same definition, we also have

$$m_{ij,n} = \frac{\sum_{k=1}^{n} \delta_{j,k} x_{i,k}}{\sum_{k=1}^{n} \delta_{j,k}} \tag{5}$$

Substituting Equation 5 into Equation 4, we find

$$m_{i,j,n+1} = \frac{\delta_{j,n+1} x_{i,n+1} + \sum_{k=1}^{n} \delta_{j,k} m_{ij,n}}{\sum_{k=1}^{n+1} \delta_{j,k}} \tag{6}$$

where $\delta_{j,n}$ denotes that at the nth trial, if the correct response is category j, $\delta_{j,n}$ is 1, and if the correct response is not category j, $\delta_{j,n}$ is zero. (Note that the correct response does not depend on the actual response, but is just the correct classification.)

Incremental Computation of Variance. The variance computations on trial $n + 1$ are also conditioned to the correct response and the incremental computation of variance from a feature i to a category j.

$$\begin{aligned} V_{ij\,n+1} &= \frac{\sum_{k=1}^{n+1} \delta_{j,k} (x_{i,k} - m_{ij,n+1})^2}{\sum_{k=1}^{n+1} \delta_{j,k} - 1} \\ &= \frac{\sum_{k=1}^{n+1} (\delta_{j,k} x_{i,k}^2 - 2\delta_{j,k} x_{i,k} m_{ij,n+1} + \delta_{j,k} m_{ij,n+1}^2)}{\sum_{k=1}^{n+1} \delta_{j,k} - 1} \\ &= \frac{\sum_{k=1}^{n+1} \delta_{j,k} x_{i,k}^2 - 2m_{ij,n+1} \sum_{k=1}^{n+1} \delta_{j,k} x_{i,k} + \sum_{k=1}^{n+1} \delta_{j,k} m_{ij,n+1}^2}{\sum_{k=1}^{n+1} \delta_{j,k} - 1} \end{aligned}$$

$$= \frac{\sum_{k=1}^{n+1} \delta_{j,k} x_{i,k}^2 - \sum_{k=1}^{n+1} \delta_{j,k} m_{ij,n+1}^2}{\sum_{k=1}^{n+1} \delta_{j,k} - 1} \tag{7}$$

Using the same procedure, we can have

$$V_{ij,n} = \frac{\sum_{k=1}^{n} \delta_{j,k} x_{i,k}^2 - \sum_{k=1}^{n} \delta_{j,k} m_{ij,n}^2}{\sum_{k=1}^{n} \delta_{j,k} - 1} \tag{8}$$

Substituting Equation 8 into Equation 7, we have

$$V_{ij,n+1} = \frac{\delta_{j,n+1} x_{i,n+1}^2 + (\sum_{k=1}^{n} \delta_{j,k} - 1) V_{ij,n} + \sum_{k=1}^{n} \delta_{j,k} m_{ij,n}^2 - \sum_{k=1}^{n+1} \delta_{j,k} m_{ij,n+1}^2}{\sum_{k=1}^{n+1} \delta_{j,k} - 1} \tag{9}$$

In order to do the incremental computation of variance, we need current values and mean values, previous mean values, and previous variances. In Model I, separate computation of the variance is made redundant by the covariance computation, but not in other models, which is why we have shown the recursion here.

Incremental Computation of Covariance. Covariance computations for trial $n + 1$ are also conditioned to the correct response and the incremental computation of covariance from a feature i to a category j.

$$S_{il,j,n+1} = \frac{\sum_{k=1}^{n+1} \delta_{j,k} (x_{i,k} - m_{ij,n+1})(x_{l,k} - m_{lj,n+1})}{\sum_{k=1}^{n+1} \delta_{j,k} - 1}$$

$$= \frac{\sum_{k=1}^{n+1} \delta_{j,k}(x_{i,k}x_{l,k} - x_{i,k}m_{lj,n+1} - x_{l,k}m_{ij,n+1} + m_{ij,n+1}m_{lj,n+1})}{\sum_{k=1}^{n+1} \delta_{j,k} - 1}$$

$$= \frac{\sum_{k=1}^{n+1} \delta_{j,k}x_{i,k}x_{l,k} - \sum_{k=1}^{n+1} \delta_{j,k}m_{ij,n+1}m_{lj,n+1}}{\sum_{k=1}^{n+1} \delta_{j,k} - 1}$$

$$= \frac{\delta_{j,k}x_{i,n+1}x_{l,n+1} + \sum_{k=1}^{n} \delta_{j,k}x_{i,k}x_{l,k} - \sum_{k=1}^{n+1} \delta_{j,k}m_{ij,n+1}m_{lj,n+1}}{\sum_{k=1}^{n+1} \delta_{j,k} - 1} \tag{10}$$

Using the same procedure, we have

$$S_{il,j,n} = \frac{\sum_{k=1}^{n} \delta_{j,k}x_{i,k}x_{l,k} - \sum_{k=1}^{n} \delta_{j,k}m_{ij,n}m_{lj,n}}{\sum_{k=1}^{n} \delta_{j,k} - 1} \tag{11}$$

Substituting Equation 11 into Equation 10, we have

$$S_{il,j,n+1} = \frac{\delta_{j,k}x_{i,n+1}x_{l,n+1} + (\sum_{k=1}^{n} \delta_{j,k} - 1)S_{il,j,n} + \sum_{k=1}^{n} \delta_{j,k}m_{ij,n}m_{lj,n} - \sum_{k=1}^{n+1} \delta_{j,k}m_{ij,n+1}m_{lj,n+1}}{\sum_{k=1}^{n+1} \delta_{j,k} - 1} \tag{12}$$

Of course, when $i = l$, $S_{il,j,n+1} = V_{ij,n+1}$, as given by Equation 9. We show in Fig. 13.1 a simple network for the case of just three features and two responses. The pairs of sample means and variances are the weights of the nodes connecting features x_1, x_2, and x_3 to responses R_1 and R_2. The sample covariances are the weights connecting pairs of features to the responses.

Let p_j denote the prior posterior probability for the group G_j, $j = 1, \ldots, g$, as defined in Equation 1 and $f_{j,n}$ the group-conditional distribution as

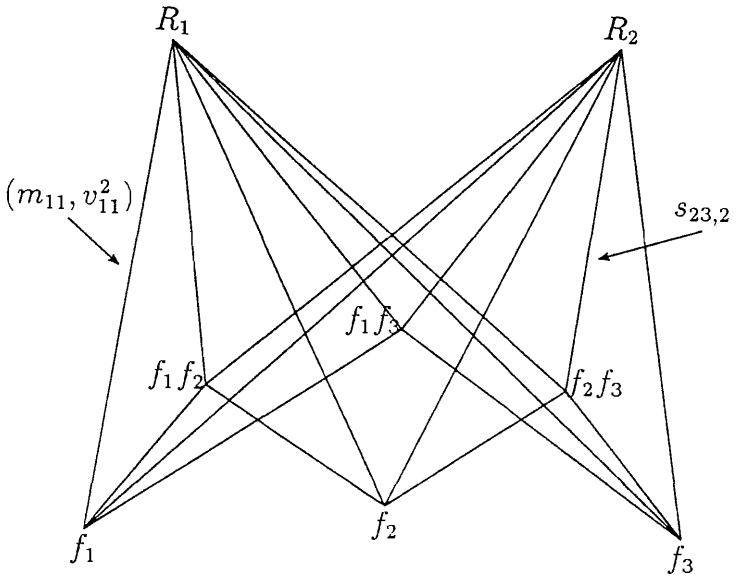

FIG. 13.1. Example of multivariable normal network.

defined in Equation 3. We then use a log ratio to compute the posterior probability of the category given the data. In fact, it is convenient to compute the log.

$$\begin{aligned}\eta_j &= \log(p_j) + \log(|\Sigma_j|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\mathbf{x}-\mathbf{m}_j)'\Sigma_j^{-1}(\mathbf{x}-\mathbf{m}_j)\right\} \\ &= \log(p_j) - \tfrac{1}{2}\log(|\Sigma_j|) - \tfrac{1}{2}(\mathbf{x}-\mathbf{m}_j)'\Sigma_j^{-1}(\mathbf{x}-\mathbf{m}_j)\end{aligned} \tag{13}$$

Bayes' rule assigns an object or stimulus with feature vector $\mathbf{x}$ to G_{j^*} if j^* is the j that maximizes, for $1 \le j \le g$,

$$(\log(p_j) - \tfrac{1}{2}\log(|\Sigma_j|) - \tfrac{1}{2}(\mathbf{x}-\mathbf{m}_j)'\Sigma_j^{-1}(\mathbf{x}-\mathbf{m}_j))$$

That is, Bayes's rule selects the category j^* that is most probable, given the feature vector $\mathbf{x}$.

Model II: Multivariate Features Only Model

This model is a special case of the multivariate normal model. Only the conditional means and variances of the features x_i are used in the model. The covariance matrix is not recursively computed and therefore is not

used in the decision rule. The recursive computations of the means and variances for each response class are just those already given for Model I.

The most probable category choice (minimum probability of error) decision rule for assigning on trial n an input vector $\mathbf{x}$ of features to a response class reduces to the familiar least-square criterion. Pick the response class j for which

$$\sum_{i=1}^{k} \frac{(x_i - m_{ij,n})^2}{V_{ij,n}}$$

is minimum.

Model III: Multinomial Model, Level 1

Like Model II, this model also considers only the features x_i of an input. In Model III, weights w_{ij} that connect feature x_i to response class j are used. They are modified by a linear learning model, rather than conditional variances as in Model II; the conditional means play a similar role in both models. Initially, for $1 \leq i \leq k$ and $1 \leq j \leq g$, $w_{ij} = 1.0$. The linear learning model is defined as follows. First, we introduce a probabilistic aspect with norm $w_{ij,n} = w_{ij,n} / \Sigma_{i=1}^{k} w_{ij,n}$. Second, if for any feature i and incorrect response j on trial n,

$$(\text{norm } w_{ij,n})(x_i - m_{ij,n})^2 > (\text{norm } w_{ij^*,n})(x_i - m_{ij^*,n})^2 \tag{14}$$

where j^* is the correct response, then

$$\begin{aligned} w_{ij,n+1} &= (1 - \theta) w_{ij,n} \\ w_{ij^*,n+1} &= (1 - \theta) w_{ij^*,n} + \theta \end{aligned} \tag{15}$$

Note that learning affects only the weights satisfying the inequality of Equation 14. Third, if the actual response is correct on trial n, no weights are changed:

$$w_{ij,n+1} = w_{ij,n} \quad \text{for} \quad 1 \leq j \leq g,\ 1 \leq i \leq k \tag{16}$$

The parameter θ is a learning parameter, $0 \leq \theta \leq 1$, which is set a priori in machine learning, but estimated from data in the case of human learning. The decision rule on trial n for assigning an input vector $\mathbf{x}$ of features to a response class is a weighted least-square criterion. Choose the response class j for which

$$\sum_{i=1}^{k} (\text{norm } w_{ij,n})(x_i - m_{ij,n})^2$$

is minimum. In general the x_i's are quantitative features, normed by their conditional variances in Model II, and now in Model III normed rather by what is roughly speaking their probabilities of being associated with a correct response, that is, norm $w_{ij,n}$.

Models IV–VI: Multinomial Models, Levels 2–4

Models IV–VI have the same structure as Model III, with for Model IV the features x_i replaced by the pairwise products $x_i x_{i'}$. For Model V the features x_i are replaced by the triple products $x_i x_{i'} x_{i''}$. For Model VI the features x_i are replaced by the quadruple products $x_i x_{i'} x_{i''} x_{i'''}$. In each of these three models the conditional means $m_{ij,n}$ of Model III are replaced by the appropriate products of conditional means. Weights are for pairs, triples, or quadruples of features, with, of course, all features in a product distinct.

MODEL EVALUATION

We now introduce a learning model that is itself used to evaluate and predict the behavior of the six models introduced in the first section. Let L be a finite nonempty set of learning models whose comparative performance during the course of learning we want to evaluate. Let the evaluation for each model i be the linear learning model, with learning parameter α independent of i. Thus for any model i in L the recursive role is:

$$q_{i,n+1} = \begin{cases} (1-\alpha)q_{i,n} + \alpha & \text{if a correct prediction was made by model } i \text{ on trial } n \\ (1-\alpha)q_{i,n} & \text{otherwise} \end{cases}$$

For the evaluation of a given model, we take the initial probability of a correct response to be the same as that of the model, which will in general be the probability that a random response is correct.

Note that for each model i and trial n, $q_{i,n}$ may be interpreted as the probability of a correct prediction by model i on trial n. This evaluation provides a direct comparison of the different models in L, but in addition it provides a means of smoothing the predictions for a given model i during a single run of a sample of the response categories to be learned. It provides a prediction based on the weighting over recent trials determined by α.

EXPERIMENTAL RESULTS

We begin with the iris data.

Iris Data

In Fig. 13.2 we show the mean learning curve for Model I based on 1,000 statistically independent runs through one cycle of the data of 150 iris specimens. Each run sampled the 150 data vectors randomly without replacement. The mean learning curve is the probability of a correct response in each block of six trials, averaged over the 1,000 runs. The input vector on each trial consists of four measurements in centimeters: sepal length, sepal width, petal length, and petal width. The three-way classification consists of the three species: *Iris setosa*, *Iris versicolor*, and *Iris virginica*.

As can be seen from the figure, the multivariate normal learning model did very well, reaching an asymptote of 0.973 for the probability of a correct response, on average, after less than 90 trials. (The learning curve is based on blocks of six trials, so each data point is the relative frequency of 6,000 responses.) In effect, this model correctly classified all but two specimens after trial 90. Also shown in Fig. 13.2 is the evaluation curve for $\alpha = 0.3$. Of course the evaluation is really useful in smoothing the data on individual runs, and cannot hope to be as accurate as possible, that is, identical with the mean learning curve. The next figure illustrates this point.

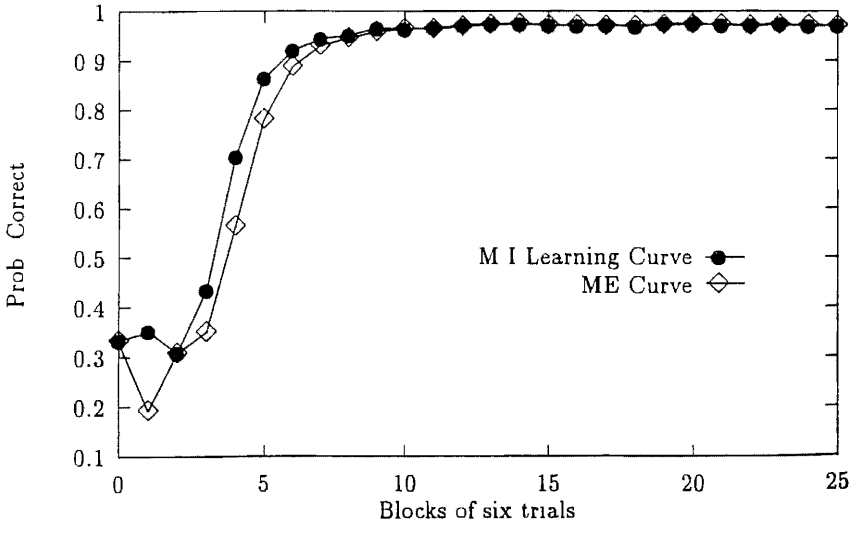

FIG. 13.2. Model I, mean learning curve and mean ME curve for iris data based on 1,000 runs.

In Fig. 13.3, two values of α, 0.1 and 0.3, are used to plot two model evaluation (ME) curves for a single run of Model I. Of course, the larger is α, the better is the fit on average to the individual run, but with $\alpha = 1.0$, we just reproduce by a lag of one trial the individual run and get none of the interesting smoothing effects of the evaluation. On the basis of a great deal of data, but without formal statistical evaluation, we believe $\alpha = 0.3$ is on average near optimal for combining a reasonable degree of smoothing with a reasonable degree of accuracy. How the ME curves smooth an individual learning curve is evident from Fig. 13.3.

In Fig. 13.4 we show, on the basis of 1,000 independent runs—just as in the case of Model I shown in Fig. 13.2—the mean learning curve for Model II, as well as the ME curve based on $\alpha = 0.3$. The results are similar to those for Model I, but not quite as good. The asymptote of the mean learning curve, closely approximated after trial 90, is 0.929.

Similar results are shown for Model III in Fig. 13.5, but the asymptote for the mean learning curve is 0.951, better than Model II but not quite as good as Model I.

In Fig. 13.6 the learning evaluation (ME) curves, with $\alpha = 0.3$, are shown for Models I–III based on the 1,000 independent runs for each model. The most interesting aspect of this figure and others is that it shows that learning is slower for Model I than for either of the other two, even though its asymptotic performance is better. This slowness of learning is not surprising, given the larger number of parameters to be learned from the data. What is surprising is that Model III has a fast rate of learning and also a reasonably good asymptote. Of the three models it is the only one with weights adjusted from trial to trial. In Models I and II all of the parameters are estimated by statistical procedures that are independent of the accuracy of the prediction from trial to trial.

In Fig. 13.7 we show the mean learning curves based on 1,000 independent runs for Models I–VI. With weights based on success or failure

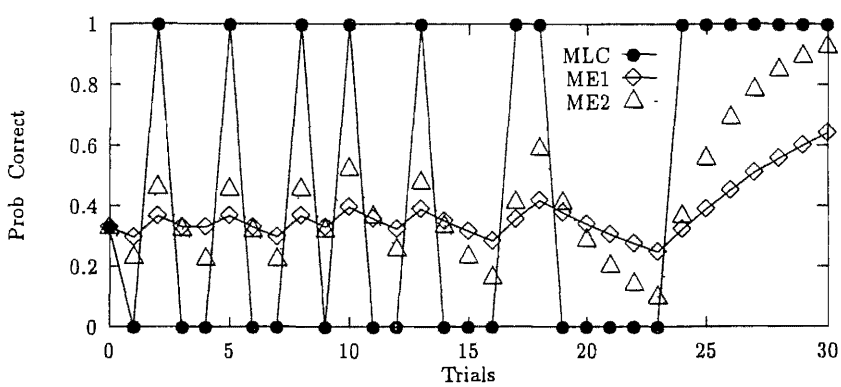

FIG. 13.3. Model I, an individual learning curve and ME curves for two values of α for a single run of iris data (ME1: $\alpha = 0.1$; ME2. $\alpha = 0.3$).

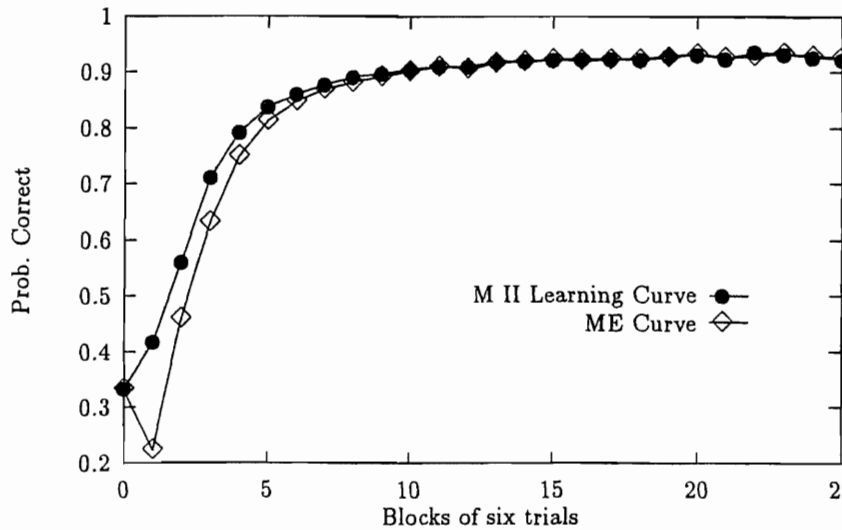

FIG. 13.4. Model II, mean learning curve and mean ME curve for iris data based on 1,000 runs.

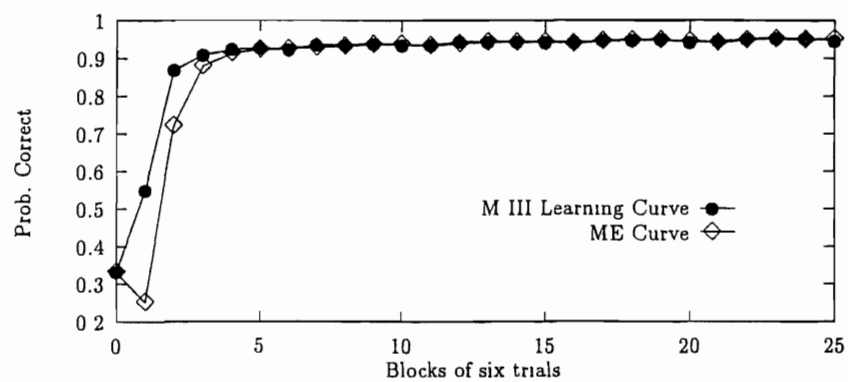

FIG 13.5 Model III, mean learning curve and mean ME curve for iris data based on 1,000 runs

of predictions, Models IV, V, and VI, like Model III, learn faster than Models I and II. These results, and the XOR data, are the only ones we consider for Models IV–VI.

Swiss Bank Note Data

We next consider the Swiss bank note data, consisting of 100 genuine and 100 forged specimens with six measurements on each: length of note, width of margin on the left, width of margin on the right, width of margin

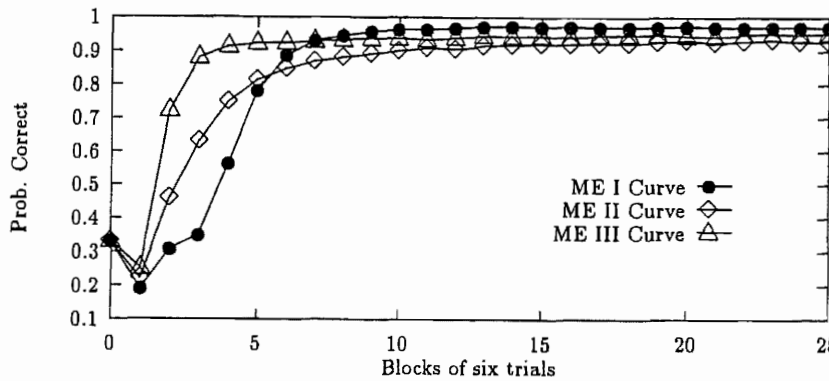

FIG. 13.6. Mean ME curves ($\alpha = 0.3$) of Models I–III for the iris data.

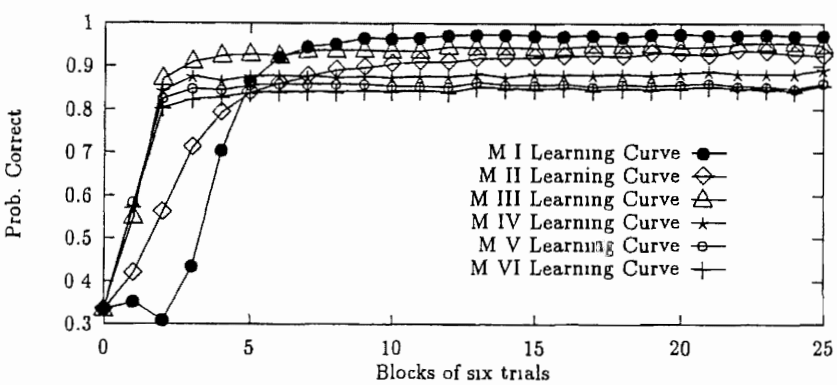

FIG 13.7 Models I–VI, mean learning curves for iris data.

at bottom, width of margin at top, and length of image diagonal. In Fig. 13.8 we show the mean learning curves for Models I–VI based on 1,000 independent runs. Asymptotically—really by the end of the 200 trials—all the models perform well. However, of comparative importance here is the very fast learning rate of Models III–VI.

Beetle Data

We now turn to the beetle data for three species with 21 specimens of *coleoptera concinna*, 31 specimens of *coleoptera heikertingeri*, and 22 specimens of *coleoptera heptapotamica*. Two measurements were made on each specimen. The measurements are the maximal width of the aedeagus in the forepart (in micrometers) and the front angle of the aedeagus (in units of 7.5 degree). In Fig. 13.9 we show the mean learning curves for Models I–III, based on 1,000 runs. Only the multivariate normal learning model

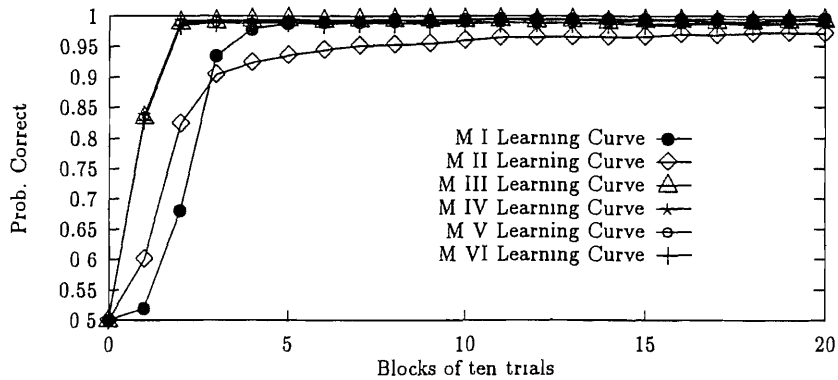

FIG. 13.8. Models I–VI, mean learning curves for bank note data

comes close to an asymptote of 1. By the end of a run of 72 trials it correctly classified 71 out of 72 specimens. On the other hand, Model III had, as previously, a fast learning rate. The ME curves for the same models shown in Fig. 13.10 are similar in shape to the mean curves shown in Fig. 13.9.

Concentric Circle Data

The concentric circle problem is shown in Fig. 13.11. Response 1 is the correct response for points (x, y) lying in the interior of the inner circle, and response 2 is correct for those in the outer circle but not in the inner circle. The two features are just the x and y coordinates of a point. We

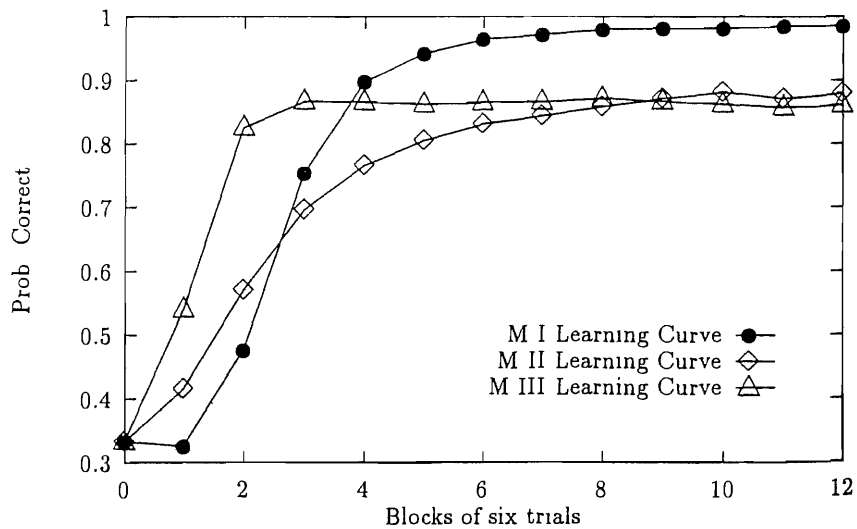

FIG 13 9 Models I–III, mean learning curves for beetle data.

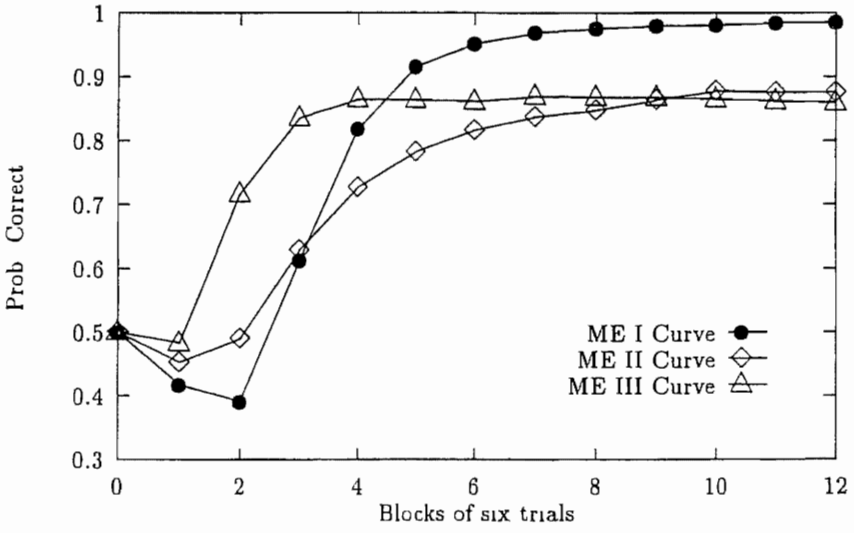

FIG. 13.10 Models I–III, mean ME curves for beetle data.

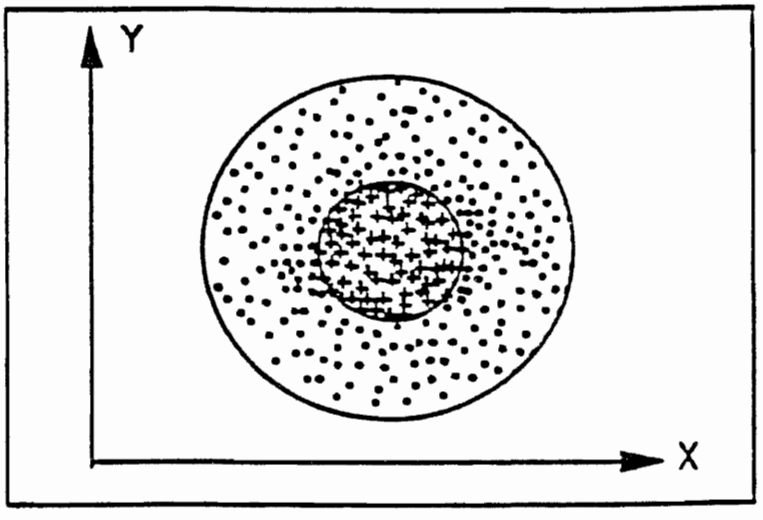

FIG. 13.11. Data points for the problem of two concentric circles.

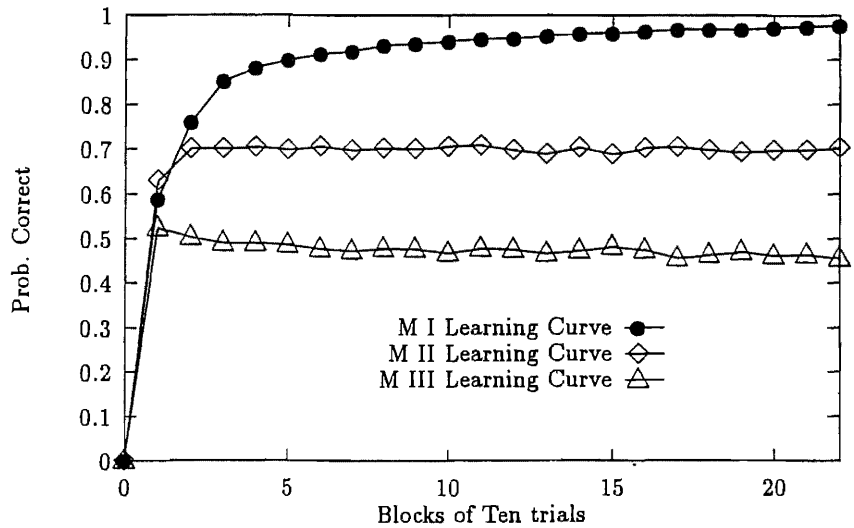

FIG. 13.12. Models I–III, mean learning curve for the concentric circle problem.

ran Models I–III on 1,000 runs of one cycle consisting of 110 feature pairs from the inner circle and 110 from the outer circle lying outside the inner circle. The mean learning curves are shown in Fig. 13.12. Only Model I, the multivariate normal model, did really well. Models II and III performed rather poorly, mainly because the asymptote mean values of the features are the same for both response-conditional distributions. Model III is worse than Model II because no account is taken of the conditional variances, which are different for the two regions.

XOR Data

A classical problem for neural nets is to learn the exclusive or (XOR) sentential connective, which is formally identical to addition mod 2. More explicitly, by the XOR problem, we mean the problem defined by the truth table for XOR. The stimulus inputs are the component truth value pairs, and the correct output is the truth value of the pair for XOR. We take, as in other cases, 1 and −1 for the input values. Ignoring sampling estimates for Model I, then if the probability of each stimulus is 0.5, the variance of the input is 1, and, as is easily computed, the two conditional covariance matrices have the form

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

The first is for truth value 1 and the second for truth value −1. Both matrices are singular. The multivariate normal model fails on this task because the covariance matrix is singular and consequently the posterior probability cannot be computed, because the covariance matrices must be inverted. As is clear from a priori analysis of Models I–III, they cannot solve the problem, and Models V and VI are inapplicable because they require more than two features as input. Model IV, on the other hand, learns in one cycle of four trials, so we do not show a figure for this model. In the language of mathematical learning theory, this is a case of one-trial learning for each pair of feature inputs.

Of interest is a more general model for the two features, f_1 and f_2, namely Model VII, which has in addition to f_1 and f_2 the constructed nodes f_1^2, f_2^2, and f_1f_2. The mean learning curve for 100 runs of two cycles as shown in Fig. 13.13 has one striking characteristic. On the fourth—that is, the last—trial of the first cycle the probability of correct response is zero. The conceptual reason for this, however, has a straightforward explanation. Suppose trial 4 has the presentation of features (−1, 1). Then on the first three trials (1, 1) and (−1, −1) had to be presented. The consequence of this fact is that the conditional mean values of the features for response 1 at the beginning of trial 4 are $f_1 = f_2 = 0$ and $f_1^2 = f_2^2 = f_1f_2 = 1$. In contrast, for the other conditional distribution, only (1, −1) has been presented and so at the beginning of trial 4 the conditional mean values of the features for response 2 are $f_1 = 1$, $f_2 = f_1f_3 = -1$, and $f_1^2 = f_2^2 = 1$. Thus, when the least square computation is made for the two possible responses on trial 4, response 2 has a minimum. In particular, it is 6 for

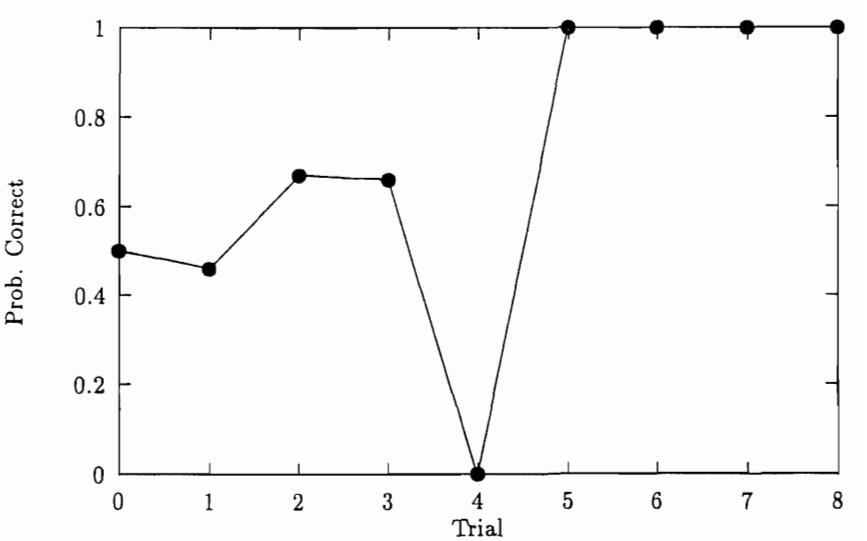

FIG. 13.13. Model IV, mean learning curve for the XOR.

response 1 in contrast to 8 for response 2, when we have ignored the weights in this small number of trials. The analysis is completely similar for presentation of any of the other three feature pairs on trial 4.

ANALYSIS OF TRANSFER

In a great variety of contexts, positive or negative transfer in learning one skill or task after another plays a crucial role in the rate of learning. Problems of transfer are also a way of moving conceptually from a static to a dynamic environment. In spite of a number of claims to the contrary, it is our impression that dynamic environment changes, particularly those involving classical problems of transfer, have as yet been little studied in the literature on neural networks. This is also a problem for standard statistics including Bayesian principles of rationality.

As a setting in which to study transfer, we used the data and analysis of transfer on 24 first-grade children learning for 56 trials equipollence of sets, that is, the concept of sets having the same number of elements, and then learning for 56 trials the concept of identity of sets, as given in Suppes (1965). The sets depicted by the simple stimulus displays had one, two, or three elements. The sets were shown in pairs, and the children were instructed to press one button when the pairs displayed were "the same" and the other button when they were not. In these experiments, no given pair of stimulus displays was repeated. This was done to prevent learning by pure stimulus association.

What we have not yet explained is what features did Model III—the only model studied here—use as input. Of course, we would have liked to model exactly the perceptual features attended to by the children, but we had no possibility of determining with any precision what those features were. Consequently, we devised a set of 12 relatively high-level features that we took as input. Given two sets in the form of $\{f_1,f_2,f_3\}$ and $\{g_1,g_2,g_3\}$, with f_2, f_3, g_2, and g_3 possibly null displays, and $f_i \neq f_j$, $g_i \neq g_j$ for $i \neq j$, our first nine features were: f_1g_1, f_2g_1, f_3g_1, f_1g_2, f_2g_2, f_3g_2, f_1g_3, f_2g_3, and f_3g_3, where

$$f_ig_j = \begin{cases} 1 & \text{if } f_i = g_j \\ -1 & \text{if } f_i \neq g_j \\ 0 & \text{if } f_i \text{ or } g_j \text{ is null} \end{cases}$$

The three other features were at a higher level still, namely, O, $I\bar{O}$, and $E\bar{I}$, where

$$O = \begin{cases} 1 & \text{if the two sets displayed are identical in the sense of ordered sets} \\ -1 & \text{otherwise} \end{cases}$$

$$I\bar{O} = \begin{cases} 1 & \text{if the two sets displayed are identical but not identical in the sense of ordered sets} \\ -1 & \text{otherwise} \end{cases}$$

$$E\bar{I} = \begin{cases} 1 & \text{if the two sets displayed are equipollent but not identical} \\ -1 & \text{otherwise} \end{cases}$$

In what follows, we also use the notation O, I, $\bar{O}$, $\bar{I}$, $I\bar{O}$, E, $E\bar{I}$, and $\bar{E}$ for subsequences of trials on which the pair of stimulus displays exemplified one or more of these features.

Models with β-Weighted Means. We found in transfer experiments that the usual Bayesian posterior or maximum likelihood estimates of means, which we computed recursively from trial to trial, is too insensitive to change after a substantial number of trials. (This is a general problem with a static Bayes approach to learning in a changing environment.) The weighting rule is again in terms of a linear model. If response j is the correct response on trial n,

$$m_{ij,n+1} = (1 - \beta)m_{ij,n} + \beta x_{i,n}$$

When this weighted estimate is used, we designate the model with a β. Throughout this section when we refer to Model III, it is really to Model β-III, with $\beta = 0.1$, except in one case noted with $\beta = 0$. In the data analysis that follows, the comparison of the first graders with Model III stops at 56 trials, for that was the number of trials in the experiment with the children. We continued the learning of Model III for another cycle of 56 trials, as can be seen in Figs. 13.14–13.18. The mean curves for Model III shown in these figures are based on 100 runs. As remarked earlier, the first graders data are based on 24 children.

In Fig. 13.14 we show for the first graders and Model III the comparison of learning identity first and learning identity after equipollence. Two aspects of the results are salient. In these mean curves for the first 56 trials, that is, seven blocks of eight trials, the four curves are very similar. The first graders and Model III have almost identical mean performance.

Second, in the overall mean performance there is little evidence of positive or negative transfer for either the first graders or Model III. However, as we examine subconcepts generating subsequences of learning trials, the evidence of positive or negative transfer, depending on the subconcept, becomes quite salient.

There are two natural ways to analyze in further detail the data on transfer. The first and simplest is to divide the pairs of stimulus displays

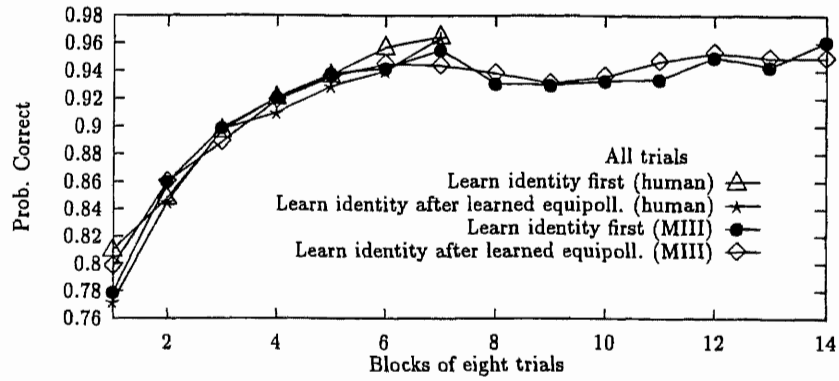

FIG. 13.14. Comparison of transfer curves using all trials.

into two groups: those exemplifying the concept of identity (I) and those not exemplifying it ($\bar{I}$). This analysis is shown in Figs. 13.15–13.18. Included with it are the mean learning curves for 48 other first graders who learned identity of sets first. In Fig. 13.15 we show four curves. Two are for the human, that is, first-grade, students. One curve is for first learning to respond correctly to the pairs of displays that are $\bar{I}$, that is, not identical. The second is the transfer curve, that is, the learning curve of $\bar{I}$ after 56 trials on equipollence. Model β-III, the multinomial learning model with β-weighted means, showed almost the same negative transfer as the first graders, with $\beta = 0.1$. Model III, without β-weighted means (i.e., $\beta = 0$), is hopelessly slow in learning identity of sets after equipollence, so the results with $\beta = 0$ are not shown.

As is clear from Fig. 13.15, the learning on $\bar{I}$ trials of the first graders and Model III were very close in rate in the first part of the experiment, the learning of identity. But Model III was slower than the first graders

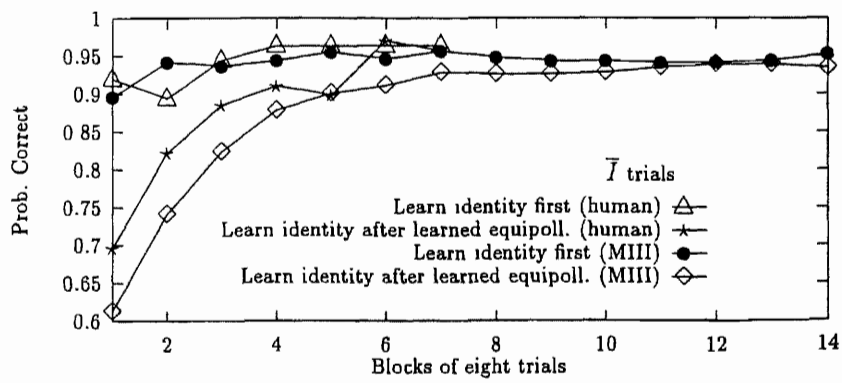

FIG. 13.15. Comparison of transfer curves for concept $\bar{I}$.

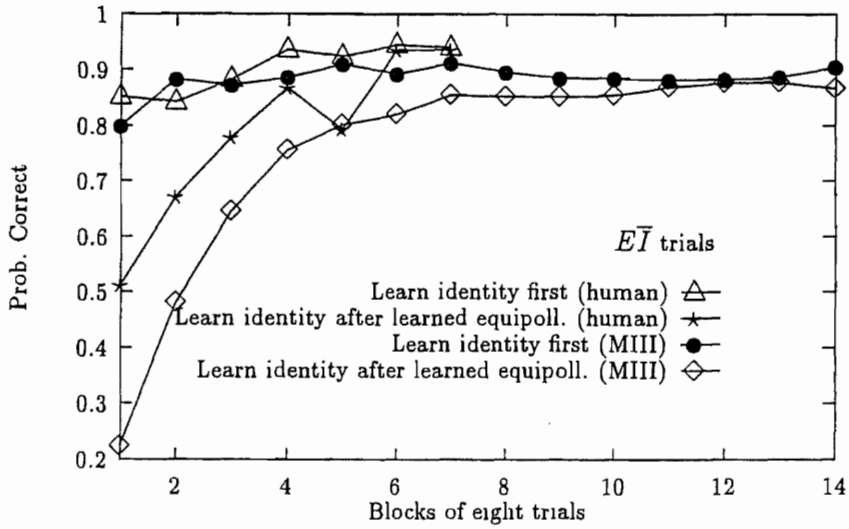

FIG. 13.16. Comparison of transfer curves for concept $E\bar{I}$.

and therefore showed more negative transfer on the $\bar{I}$ trials after first learning equipollence.

This greater negative transfer also holds, as shown in Fig. 13.16, for the $E\bar{I}$ trials, the critical ones to change when learning identity after equipollence.

In Fig. 13.17 we show the I trials before and after transfer, which is positive. We notice at once, from inspection of the graphs, that the rates for Model III and the first graders learning identity first are nearly the same.

In Fig. 13.18 we show the learning of identity on $I\bar{O}$ trials before and after transfer. Now the learning before transfer, that is, the learning of I on $I\bar{O}$ trials initially, is not at approximately the same rate for Model III and the first graders. In this case Model III is faster than the first graders, contrary to the approximately same rate for I trials before transfer, as shown in Fig. 13.17. The explanation is that on O trials, not shown in either figure, I was learned faster by the first graders than by Model III. This is not surprising, because the O trials present the natural concept of identity for children, and Model III does not have their prior experience.

On the other hand, the positive transfer for Model III is greater than for the first graders. This is true for both Figs. 13.17 and 13.18. We emphasize that on I or $I\bar{O}$ trials we expect for both first graders and any reasonable model positive, not negative, transfer. What is interesting is comparison of the amount of positive transfer. The greater positive transfer for Model III arises mainly from complete retention of the earlier learning of the correct responses to $I\bar{O}$ trials when equipollence was the concept being learned. The first graders exhibit less such retention, as is easily seen by comparing the proportion correct on the first block of trials of learning

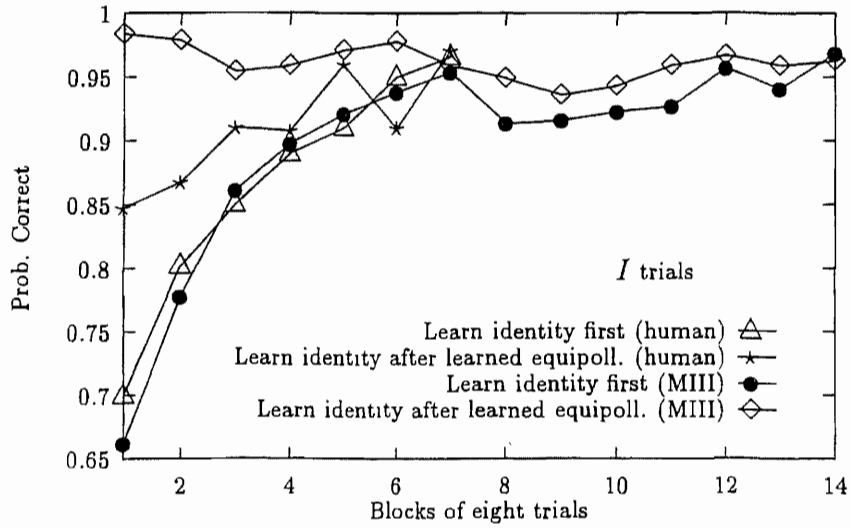

FIG. 13.17. Comparison of transfer curves for concept *I*

identity of sets after equipollence. This proportion is about 0.97 for Model III and about 0.74 for the first graders.

COMPARISONS OF MODELS

We consider in this final section several other models and the problems they address, and summarize our own analyses.

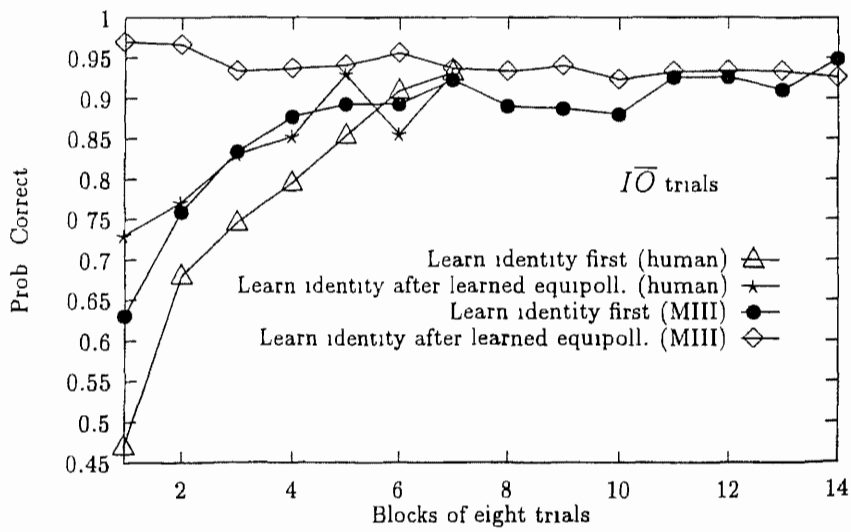

FIG. 13.18. Comparison of transfer curves for concept $I\bar{O}$

Back-Propagation Models

The kind of probabilistic models we have concentrated on are conceptually related to, but in detail different from, neural net models using a back-propagation algorithm. The general consensus in the literature is pretty much that for the kinds of learning tasks considered here, back-propagation methods are comparatively slow. Using the back-propagation algorithm described in Wasserman (1989, 1993) with two hidden units, we obtained for a single run of $150 \times 20 \times 60 = 180{,}000$ trials the learning curve for the iris data shown in Fig. 13.19. The comparison with Fig. 13.2 for the multivariate normal model (MI) with one cycle of 150 trials is striking, but in fact the comparison is even more dramatic, for a compatible probability correct was reached by Model I after only 30 trials, and only 20 trials by Model III. Typically, as other investigators have found, and as is the case for the iris data, classification performance of the back-propagation network required a large number of learning trials and the resulting asymptotic performance was no better than Models I and III. This indicates that situations where linear learning methods are superior in performance to back-propagation can occur, and may indeed be common.

After some conversation about the data sets and models in this paper, David Rumelhart agreed to run one of his more powerful back-propagation models on the iris data. Using 10 hidden units, he trained the model on the first 100 iris samples and tested on the last 50. With 1,000 runs of the training data, no mistakes were made on the test data.

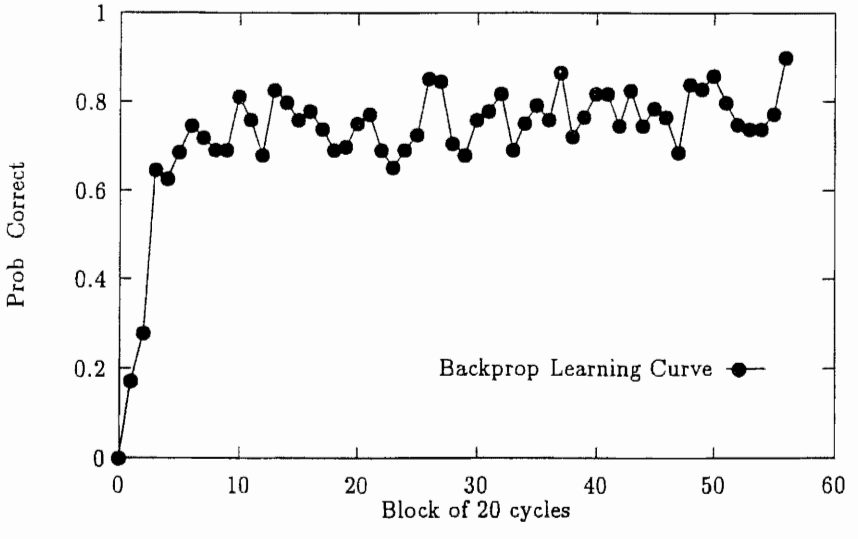

FIG. 13.19. Single run of back-propagation model on iris data for 180,000 trials.

Probability Matching

A principle feature of stochastic learning models developed three to four decades ago is *probability matching*. What is meant by this may be explained by examining some details of the linear model used in Model III for change of weights and the ME model of evaluation. In simplest form, let there be only one input feature, a red light signaling that it is time to make either an R_1 or R_2 response. Because there is only one constant input feature assumed at the beginning, we may ignore it and formulate the model just in terms of responses and reinforcement, E_1 and E_2, that is, information about the correct response. (We limit ourselves here to just two responses.) The linear theory is formulated for the probability of a response on trial $n + 1$, given the entire preceding sequence of responses and reinforcements. For this preceding sequence we use the notation x_n. Thus, x_n is a sequence of length $2n$ with 0s and 1s in the odd positions indicating responses R_1 and R_2, and 1s and 2s in the even positions indicating reinforcing events E_1 and E_2. The axioms of the linear theory are as follows:

Axiom L1. If $E_n = 1$ and $P(x_n) > 0$, then

$$P(R_{n+1} = 1|x_n) = (1 - \theta)P(R_n = 1|x_{n-1}) + \theta$$

Axiom L2. If $E_n = 2$ and $P(x_n) > 0$, then

$$P(R_{n+1} = 1|x_n) = (1 - \theta)P(R_n = 1|x_{n-1})$$

Here, as usual, θ is to be thought of as the learning parameter. We consider what is from a theoretical standpoint one of the simplest cases: noncontingent reinforcement. This case is defined by the condition that the probability of E_1 on any trial is constant and independent of the subject's responses. It is customary in the literature to call this probability π. Thus, $P(E_n = 1) = \pi$, $P(E_n = 2) = 1 - \pi$.

For the noncontingent case, we can derive from L1 and L2 the same asymptotic mean result,

$$\lim_{n \to \infty} P(R_n = 1) = \pi \tag{17}$$

Because π is the asymptotic probability of response R_1, we have probability matching of responses and reenforcements on average. It is clear that Equation 17 is not the prediction of the multivariate normal model (MI), for with the Bayes posterior for the two response classes, asymptotically R_1 would be chosen with probability 1 when $\pi > \frac{1}{2}$. Similar remarks apply to essentially all the models previously discussed, with one exception. In the

models with β-weighted means, probability matching will occur with $\beta = 1$. More detailed results can be obtained using the methods developed in Estes and Suppes (1959) for analyzing the detailed properties of the linear model.

Summary of Analysis

The model evaluation methods introduced in the second section are seen to be most useful when applied to single runs of test data. This is particularly evident in the resulting smoothing of the learning to be seen in Fig. 13.3. More work will be required to determine which values of the parameter α are best for what purposes. The value $\alpha = 0.3$ worked rather well in our preliminary investigations.

We considered six data sets and seven models. In Table 13.1 we summarize the results in two ways. We show, for each data set, which model exhibited the fastest learning and which model had the best asymptotic result. In several cases there were ties, so several models are entered in the table.

The most striking fact about the table is that no one of the seven models comes close to being uniformly best. Models I and III were best most often, but their virtues are clearly different. Model I is most often best asymptotically—three out of the six cases. Model III had most often the fastest learning rate, again three out of six cases.

We also note that multinomial models V and VI were not best in any of the 12 categories of Table 13.1.

Actually, the situation is a little more complicated than we have described, because considering the β-weights explicitly increases dramatically the number of models. Moreover, the need for the β-weights is quite clear both conceptually and computationally in the case of the transfer data (Section 4). Because the β-weights fly in the face of much standard statistical practice in estimating parameters, but not in time-series analysis, their

TABLE 13.1
Comparison of Model Performance

Data Set	*Fast Learning*	*Best Asymptotic*
Iris	III	I
Bank notes	III, IV	I, III, IV
Beetles	III	I
Circles	I, II	I
XOR	IV, VII	IV, VII
Transfer	β-III	β-III

relevance for parameter estimation in transfer experiments warrants further study.

REFERENCES

Estes, W., & Suppes, P. (1959). Foundations of linear models. In R. Bush & W. Estes (Eds.), *Studies in mathematical learning theory* (pp. 137–179). Stanford: Stanford University Press.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179–188.

Flury, B., & Riedwyl, H. (1988). *Multivariate statistics, a practical approach.* London: Chapman and Hall.

Lubischew, A. A. (1962). On the use of discriminant functions in taxonomy *Biometrics, 18*, 455–477.

McLachlan, G. (1992). *Discriminant analysis and statistical pattern recognition* New York: Wiley.

Suppes, P. (1965) On the behavioral foundations of mathematical concepts. *Child Development Monograph*, Serial 99, Vol. 30, No. 1.

Suppes, P., Bottner, M., & Liang, L. (1995). Comprehension grammars generated from machine learning of natural languages. *Machine Learning, 19*, 133–152.

Suppes, P., Liang, L., & Böttner, M. (1992). Complexity issues in robotic machine learning of natural language. In L. Lam & V. Naroditsky (Eds.), *Modeling complex phenomena* (pp. 102–127). New York: Springer-Verlag.

Wasserman, P. (1989). *Neural computing, theory and practice.* New York: Van Nostrand Reinhold.

Wasserman, P. (1993) *Advanced methods in neural computing.* New York: Van Nostrand Reinhold.