A CRITICAL OVERVIEW OF NEURAL NETWORK PATTERN CLASSIFIERS*

Richard P. Lippmann
Room B-349, MIT Lincoln Laboratory
Lexington, MA 02173-9108
rpl@sll.mit.edu

ABSTRACT—A taxonomy of neural network pattern classifiers is presented which includes four major groupings. Global discriminant classifiers use sigmoid or polynomial computing elements that have “high” non-zero outputs over most of their input space. Local discriminant classifiers use Gaussian or other localized computing elements that have “high” non-zero outputs over only a small localized region of their input space. Nearest Neighbor classifiers compute the distance to stored exemplar patterns and Rule Forming classifiers use binary threshold-logic computing elements to produce binary outputs. Results of experiments are presented which demonstrate that neural network classifiers provide error rates which are equivalent to and sometimes lower than those of more conventional Gaussian, Gaussian mixture, and binary tree classifiers using the same amount of training data. Many neural network classifiers also provide outputs which estimate Bayesian a posteriori probabilities. Experiments used low-dimensional (2 to 55 input) phoneme classification tasks, a high-dimensional (360 pixel input) handwritten digit classification task, and machine learning data bases. They demonstrate that neural network classifiers provide new alternatives to more conventional approaches. They often provide reduced error rates and always allow other classifier characteristics to be traded off to best match the requirements of a particular problem. Characteristics which often differ dramatically across classifiers include classification time, training and adaptation time, ease of implementation, memory requirements, rejection accuracy, and usefulness of outputs as Bayes probability estimates.

INTRODUCTION
The table in Figure 1 contains a taxonomy of five major types of neural network and conventional pattern classifiers that can be used to classify fixed-length patterns. The first row in this table represents conventional proba-

---
*This work was sponsored by the Defense Advanced Research Projects Agency and the Air Force Office of Scientific Research.
Figure 1: A Taxonomy Including Five Types of Conventional and Neural Network Pattern Classifiers.

Probabilistic classifiers which model likelihood distributions of pattern classes separately using parametric functions. Probabilistic classifiers used for speech recognition include Gaussian linear discriminant and Gaussian mixture classifiers. This most common approach to pattern classification provides good performance when the assumed functional form of class distributions matches real-world data distributions and when there is sufficient training data to estimate parameters. Performance can be poor when class distributions are not modeled well or when training data is limited.

The bottom four rows in Figure 1 include both neural network and conventional classifiers. Global classifiers form output discriminant functions from internal computing elements or nodes that use sigmoid or polynomial functions which have "high" non-zero outputs over a large region of the input space. These classifiers include multi-layer perceptrons trained with backpropagation (back-propagation classifiers), Boltzmann machines, and high-order polynomial networks. Local classifiers form output discriminant functions from internal computing elements that use Gaussian or other radially symmetric functions which have "high" non-zero outputs over only a localized region of the input space. These two types of classifiers make no strong assumptions concerning underlying distributions. They can form complex decision regions with only one or two hidden layers and are typically trained to minimize the mean-squared error between desired and actual network out-

Nearest neighbor classifiers perform classification based on the distance between a new unknown input and previously stored exemplars. Conventional k-nearest neighbor, neural network learning vector quantizer (LVQ) classifiers, and some forms of neural network adaptive resonance theory (ART) classifiers are examples of nearest-neighbor classifiers. These types of classifiers train extremely rapidly but can require large amounts of computation time on a serial processor for classification and also large amounts of memory.

Rule forming classifiers partition the input space into labeled regions using threshold-logic nodes or rules. Inputs are classified by the label of the region the input falls in. These classifiers have binary outputs and include binary decision trees such as the CART binary tree, the hypersphere classifier, perceptrons with hard-limiting nonlinearities trained using the perceptron convergence procedure, and many machine learning approaches that result in a small set of classification rules.

**NETWORKS ESTIMATE BAYES PROBABILITIES**

Bayesian *a posteriori* probabilities, hereafter referred to as Bayes probabilities, are fundamental to statistical approaches to pattern classification. Selecting that pattern class with the highest Bayes probability for each input pattern minimizes the overall classification error rate [1]. Although rule-forming neural network classifiers only provide binary outputs, many other neural network classifiers provide continuous outputs which are estimates of Bayes probabilities. Continuous outputs are useful because they can provide confidence ratings for classification decisions and when input patterns can be rejected for human verification or to request further information. Such a strategy is common in medical diagnosis and when decoding monetary amounts on checks. Bayes probabilities are also required in problems such as speech recognition where outputs from many independent low-level classifiers must be integrated to make higher level decisions. They are also required to compensate for differences between pattern class probabilities in training and test data and to minimize alternative risk functions.

Many proofs have been presented, including those in [7], which demonstrate that outputs of local and global neural network classifiers minimize commonly used squared error and cross-entropy cost functions when they are Bayes probabilities. Figure 2 presents an example which illustrates that outputs of Radial Basis Function (RBF), Multi-Layer Perceptron (MLP) and high-order polynomial (GMDH) classifiers all estimate Bayes probabilities accurately. Estimation accuracy is best in the central region for this problem which has two input classes that have Gaussian mixture likelihood distributions. All networks in this example had one input, two outputs, and were trained with four thousand training samples per class. The sum of network outputs across all classes should be 1.0 if outputs accurately estimate Bayes probabilities. Figure 3 demonstrates that the sum of network outputs for the
Above RBF, MLP, and GMDH classifiers is near 1.0 in the central region with sufficient training data. This occurred even though outputs were never explicitly constrained during training to sum to 1.0. Further simulations and a more thorough discussion of the relation between network outputs and Bayes probabilities are available in [7].

CLASSIFIER COMPARISONS
A listing of many different neural network and conventional classifiers which groups classifiers using the above taxonomy but focuses on differences in network outputs is presented in Figure 4. Descriptions and references for many of these classifiers are available in [1, 4, 5, 6]. As noted above, probabilistic, global, and local classifiers can produce smoothly varying outputs which estimate either the likelihood of the input data or Bayes probabilities. Nearest neighbor and rule forming classifiers produce binary outputs that attempt to
minimize the number of misclassifications. In addition to these differences, classifiers often differ dramatically in (1) Memory and computation requirements, (2) Training and classification times, (3) Complexity, (4) Rejection accuracy, (5) Ability to identify outliers, (6) Ease of adaptation, (6) Performance with noisy features and missing data, and (7) Performance with binary or continuous inputs.

**Figure 3:** Sum of Two Outputs for Each of the Three Networks Whose Outputs are Plotted in Figure 2.

Four studies described below compared many of the neural network and conventional classifiers in Figure 4 using real-world data from different application areas. In all studies, the complexity or size of classifiers was carefully...
One series of experiments summarized in [4, 6] compared eight neural network and six conventional classifiers using a talker-dependent recognition task (7 digits, 22 cepstral inputs per digit, 10 talkers, 70 training and 112 testing patterns per talker), a talker-independent vowel recognition task (10 vowels, 2 formant frequency inputs, 67 talkers, 338 training and 333 testing patterns), and two artificial tasks (Bullseye and Disjoint) with two input dimensions that require either a single convex or two disjoint decision regions. Error rates for Gaussian, Back-Propagation, Radial Basis Function, Hypersphere, k-nearest neighbor, learning vector quantizer, feature-map, and CART binary tree classifiers on these data bases are presented in Figure 5. The shaded area represents one binomial standard deviation above and below average classifier performance. Except for a few exceptions, classification error rates are statistically indistinguishable. Practical classifier characteristics on these problems including training time, classification time, and memory requirements, however, differed by more than an order of magnitude.
A second study used a more complex talker-dependent phoneme classification task with 4,600 training and 7,600 test frames extracted from the initial consonant segment of the letters “B”, “D”, and “G” from the TI-46 word database [2]. Inputs to classifiers were adjoined cepstral vectors from one to five 10 msec speech frames (11 to 55 cepstral inputs). Classifiers were evaluated using the 10 training and 16 testing tokens available for each word and talker. Experiments comparing back-propagation, radial basis function, k-nearest neighbor, Gaussian, and Gaussian mixture classifiers demonstrated that radial basis function and Gaussian mixture classifiers provided error rates as low as the other classifiers. Error rates for these two classifiers obtained using different numbers of adjoined speech frames and different numbers of centers (Gaussian mixture components and radial basis function hidden nodes) are presented in Figure 6. As can be seen, neural network radial basis function classifiers provide error rates which are significantly lower than those of tied Gaussian mixture classifiers with similar numbers of parameters. This is presumably because the radial basis function classifier minimizes the overall error rate directly instead of approximating class distributions separately.

Figure 6: Classifier Error Rates for Gaussian Mixture and Radial Basis Function Classifiers on a Phoneme Classification Task Using 1 to 5 Adjoined Speech Frames as Inputs.

A recent study [3] demonstrated that global, local, and nearest neighbor classifiers can provide similar low error rates even with very high dimensional inputs (360 pixels) on a handwritten digit classification task. Classifiers were tested using a database with 30,600 training and 5,060 test patterns which were normalized to form 15x24 gray scale images with 360 pixels each. A multi-layer perceptron classifier was trained with back-propagation and had an architecture which was carefully tailored to this problem. It was compared to k-nearest neighbor and radial basis function classifiers. All classifiers were tuned to provide good performance and compared on this database without rejections. Results are shown in Table 1.
Table 1: Results of Handwritten Digit Recognition Experiments Using Multi-Layer Perceptron, k-Nearest Neighbor, and Radial Basis Function Classifiers.

<table>
<thead>
<tr>
<th></th>
<th>Back-Prop</th>
<th>KNN</th>
<th>RBF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error Rate (No Rejections)</td>
<td>5.15%</td>
<td>5.14%</td>
<td>4.77%</td>
</tr>
<tr>
<td>Free Parameters</td>
<td>5,472</td>
<td>11,016,000</td>
<td>371,000</td>
</tr>
<tr>
<td>Training Time (Hours)</td>
<td>67.7</td>
<td>0.0</td>
<td>16.5</td>
</tr>
<tr>
<td>Classification Time (Secs/Char)</td>
<td>0.14</td>
<td>0.22</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Error rates across the three classifiers are statistically indistinguishable. This is somewhat surprising given the high dimensionality of the problem and the enormous differences (more than three orders of magnitude) in the number of parameters used by the different classifiers. Training time and classification time results in Table 1 clearly illustrate the differences in practical characteristics of the three classifiers. The back-propagation classifier requires days of training on a DECstation 3100 but provides the most rapid classification. The k-nearest neighbor classifier requires no training time but needs megabytes of storage and many seconds to classify each digit. The RBF classifier requires much less training time than the back-propagation classifier at the expense of more storage. Further experiments explored the number of patterns that must be rejected to reduce the error rate to a low 0.3% per digit. Patterns were rejected based on the outputs of back-propagation and RBF classifiers and the makeup and volume of space covered by the k-nearest neighbors for the k-nearest neighbor classifier. Outputs of the RBF classifier were most useful for accurately rejecting patterns that cause errors. A per-digit error rate of 0.3% was achievable with this classifier by rejecting only 19% of the digits. To achieve this same accuracy, the back-propagation classifier had to reject more than 30% of all patterns and the k-nearest neighbor classifier had to reject more than 66% of all patterns.

A number of studies have compared neural network and machine learning classifiers. Figure 7 plots error rates reported in [8] for four machine learning and statistical data bases using nearest neighbor, Gaussian, back-propagation, CART tree, and a machine learning rule-based classifier. The shaded area again represents one binomial standard deviation above and below average classifier performance. These four data bases, which are similar to those used to evaluate many machine learning algorithms, were surprisingly small. The maximum number of classes was 3, numbers of patterns were limited, and inputs were often binary and of low dimensionality. Similar low error rates were provided by back-propagation, CART binary tree, and rule classifiers. Higher error rates were provided by nearest neighbor and Gaussian classifiers for the cancer and hypothyroid problems where best performance was provided by classifiers which could ignore all but a few important input features.
SUMMARY
Many different neural network classifiers have been developed and compared to conventional classifiers over the past few years. Studies using speech, handwritten digit recognition, and machine learning data bases have demonstrated that error rates with these classifiers are as low as and sometimes lower than error rates of more conventional statistical and machine learning classifiers when using the same amount of training data.

Studies have demonstrated that neural network classifiers provide alternative tradeoffs in classifier characteristics including training and classification times, memory requirements, complexity, ease of implementation, and rejection accuracy. Global neural network classifiers include multi-layer perceptrons trained with back-propagation and Boltzmann machines. They generally have low memory requirements and provide rapid classification but have long training times. Local neural network classifiers include radial basis function classifiers and other kernel discriminant classifiers. They generally have higher memory requirements but shorter training times. Nearest neighbor neural network classifiers include learning vector quantizer, feature-map, and some versions of adaptive resonance theory classifiers. They generally have higher memory requirements but can provide rapid training and adaptation. Rule Forming neural network classifiers include binary perceptrons trained...
with the perceptron convergence procedure and the hypersphere classifier. They can provide rapid adaptation but may have high memory requirements. They are most appropriate when only a binary classification decision is required.

Simulations and theoretical proofs also demonstrate that many Global and Local neural network classifiers including back-propagation and radial basis function classifiers estimate Bayesian a posteriori probabilities. Such networks are useful when outputs of multiple networks must be used to make higher level classifications and when classification confidence must be estimated to reject patterns.

References


