# Neural Networks in Data Mining

> "The human brain contains roughly $10^{11}$ or 100 billion neurons. That number approximates the number of stars in the Milky Way Galaxy, and the number of galaxies in the known universe. As many as $10^4$ synaptic junctions may abut a single neuron. That gives roughly $10^{15}$ or 1 quadrillion synapses in the human brain. The brain represents an asynchronous, nonlinear, massively parallel, feedback dynamical system of cosmological proportions." (Kosko (1992), pp 13)

It is this same human brain which serves as the model for artificial neural networks' topology and dynamics. (Kosko (1992)) Artificial neural networks have developed from generalized neural biological principles. McCulloch and Pitts (1943) proposed the neuron as a binary threshing device in discrete time. (Ripley (1996)) Neurons are now known to be more complicated than this model having a graded, rather than threshold, response, operating in continuous time, and achieving nonlinear functions of inputs. (Ripley (1996)) Compared to computers, however, neurons are rather slow, with a speed of about 100 meters/second. (Ripley (1996)) To compensate for this lack of speed, the human brain is highly distributed and massively parallel. (Ripley (1996)) As parallel processes go the human brain is unsurpassed. The pattern of connections between neurons proves to be an ideal model for the neural network's architecture (Fausett (1994))

Neural networks are essentially comprising three pieces: the architecture or model; the learning algorithm; and the activation functions. (Fausett (1994)) Neural networks are programed or "trained" to " . . . store, recognize, and associatively retrieve patterns or database entries; to solve combinatorial optimization problems; to filter noise from measurement data; to control ill-defined problems; in summary, to estimate sampled functions when we do not know the form of the functions." (Kosko (1992), p.13) It is precisely these two abilities (pattern recognition and function

estimation) which make artificial neural networks (ANN) so prevalent a utility in data mining.

As data sets grow to massive sizes, the need for automated processing becomes clear. With their "model-free" estimators and their dual nature, neural networks serve data mining in a myriad of ways. (Kosko (1992)) Later in this paper, we illustrate real world data mining situations where neural networks have been used quite successfully. But first, it is necessary to examine the architecture, the learning algorithms, and the activation functions in more detail.

Before we proceed it is interesting to note that several neural network models are similar or identical to statistical models. One of the goals of this paper is to point out the similarities and overlap between these two fields. The nomenclature differences exist in the literature because neural networks' jargon differs from statistical jargon. Below we list comparable terms taken from neural network and statistical literature: (Sarle (1994))

| Neural Network | Statistical |
|---|---|
| features | variables |
| inputs | independent variables |
| outputs | predicted values |
| targets or training values | dependent variables |
| errors | residuals |
| training or learning | estimation |
| error function, cost function | estimation criterion |
| patterns or training pairs | observations |
| (synaptic) weights | parameter estimates |
| higher-order neurons | interactions |
| functional links | transformations |
| supervised learning | regression and discriminant analysis |
| unsupervised learning | data reduction |
| competitive learning or adaptive vector quantization | cluster analysis |
| generalization | interpolation or extrapolation |

The development of ANNs from neural biological generalizations has required some basic

assumptions which we list below:

1. "Information processing occurs at many simple elements called neurons [also referred to as units, cells, or nodes].
2. Signals are passed between neurons over connection links.
3. Each connection link has an associated weight, which, in a typical neural net, multiplies the signal transmitted.
4. Each neuron applies an activation function (usually nonlinear) to its net input (sum of weighted input signals) to determine its output signal." (Faussett (1994), p. 3)

From these assumptions it is easy to see the three pieces of a neural network. The first piece is the neural network's pattern of connections between neurons, the architecture or model. (Fausett (1994)) The second part is the method which determines the weights on the connections, the training or learning algorithm. (Fausett (1994)) These weights represent the information being processed by the neural network. (Fausett (1994)) The third compnent is the function which determines each neuron's output signal, the activation function. The activation function typically maps any real input into a bounded range usually 0 to 1 or -1 to 1. (Sarle (1994)) This activation output, only one per neuron, is sent to several other neurons according to the connection scheme found in the architecture. (Fausett, (1994))

Borrowing from Sarle's (1994) paper, we will use his neural network diagrams for illustration of some of the different types of neural networks. Figure 1 depicts a simple neural network which is nothing more than simple linear regression. Circles and boxes are used to represent neurons, while arrows depict connection links. Observed variables are indicated as circles. Function arguments are indicated by boxes, where most boxes have a bias or intercept

element. The symbol inside the box shows the type of function. Connection links, i.e. arrows, have a set of corresponding weights which need to be estimated. Parallel lines signify a fitting procedure such as least squares or maximum likelihood methods. We start with the simplest neural network and work our way up to the more complicated architectures.

"A function which computes a linear combination of the variables and returns the sign of the result is known as a *Perceptron* after the work of F. Rosenblatt." (Ripley (1996), p.116) The first step is for the perceptron to compute the net input as a linear combination of the inputs possibly with an intercept or bias term. (Sarle (1994)) Next an activation function, usually nonlinear, is applied to the net input producing an output. (Sarle (1994)) Normally each output uses the same activation function, although there are cases where different activation functions are used. (Sarle (1994))

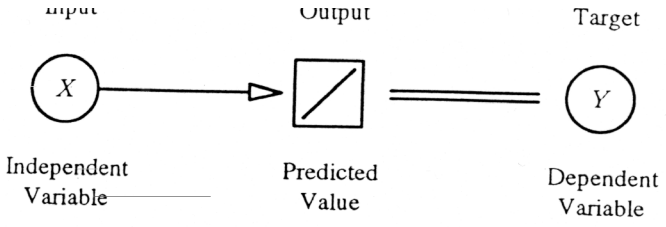


**Figure 1**

Below we list the most commonly used activation functions

- linear or identity: $f(x)=x$

- hyperbolic tangent: $f(x)=\tanh(x)$

- logistic: $f(x)=\frac{1}{(1+\exp(-x))}=\frac{(\tanh(\frac{x}{2})+1)}{2}$

- threshold: $f(x)=0 \quad if\ x<0, \quad 1\ otherwise$

- Gaussian: $f(x)=\exp(\frac{-x^2}{2})$

"The activation function in a perceptron is analogous to the inverse of the link function in a generalized linear model (GLIM)." (Sarle (1994), p.4) The major difference between activation functions and inverse link functions are that activation functions are usually bounded, but inverse link functions like the identity, reciprocal, and exponential are not bounded. (Sarle (1994))

Following the activation function comes the fitting of the data. Rumelhart and McClelland (1986) proposed to fit the perceptron using least squares' method. Given $(\mathbf{x}^p, \mathbf{t}^p)$, let $\mathbf{y} = f(\mathbf{x};\mathbf{w})$ be the network's output. A parameter vector $\mathbf{w}$ is then chosen so as to minimize

$$E(w)=\sum_p \|t^p - f(x^p;w)\|^2. \qquad (1)$$

(Ripley (1996)) The summand here is over all outputs and over the training set. (Sarle (1994))

We borrow the symbols for neurons from Sarle (1994) and show them below in Figure 2. A perceptron with a linear activation function is really just a linear regression model. (Sarle (1994)) In Figure 3 we show a multivariate (multiple) linear perceptron. With a logistic activation function

the perceptron becomes a logistic regression model, as shown in Figure 4. (Sarle (1994)) A linear discriminant model is obtained with the threshold activation function see Figure 5. (Sarle (1994)) Figure 5, because there is only one output, is often referred to as an *adaline*. (McClellend and Rumelhart (1986)) Note that only with the threshold activation function can we get a true multi layer extension of the perceptron. (Ripley (1996))

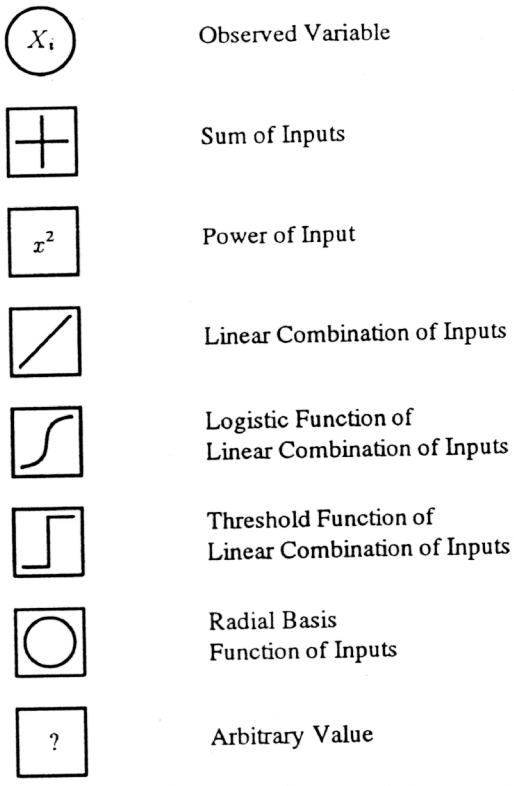


**Figure 2**

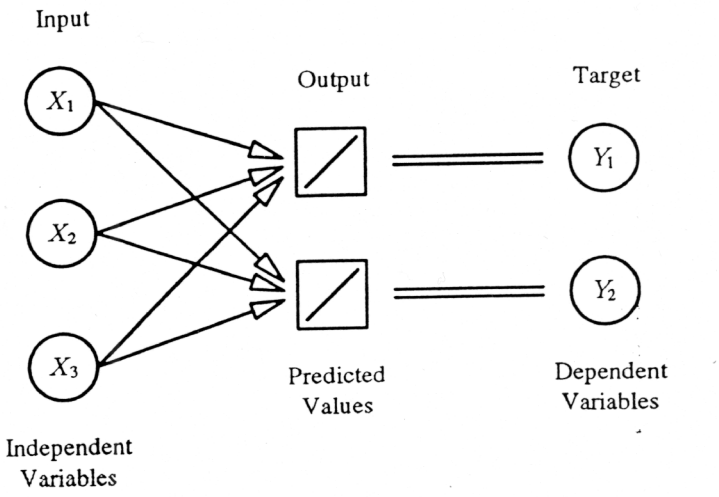


**Figure 3**

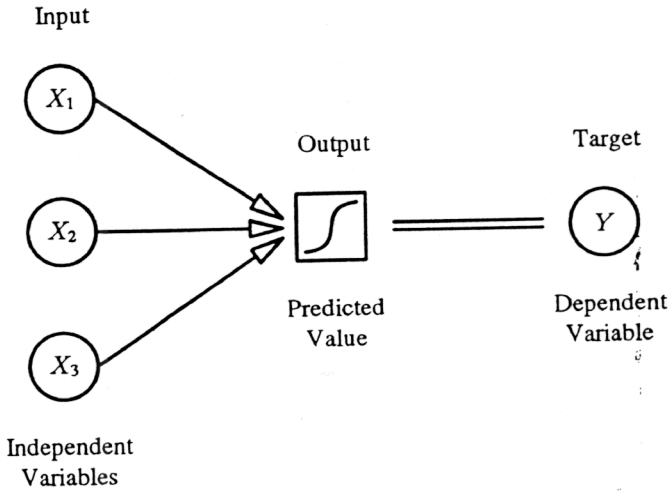


**Figure 4**

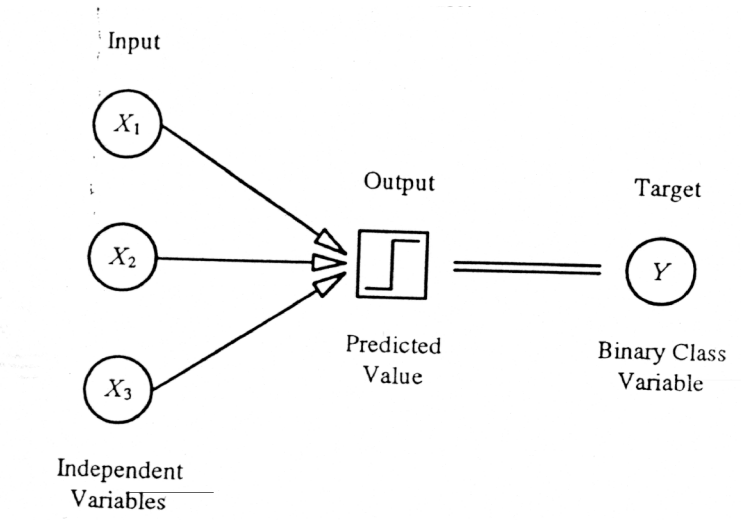


**Figure 5**

The Multi Layer Perceptron (MLP), also called a feed-forward network, involves estimated weights between the inputs and a hidden layer where the hidden layer has a nonlinear activation function. (Sarle (1994)). Figure 6 depicts a Multi Layer Perceptron where the hidden layer uses a logistic activation function. (Sarle (1994)) often there are multiple inputs and outputs in an MLP as shown in Figure 7. (Sarle (1994)) It is also possible to have the number of hidden units less than the number of inputs or outputs, see Figure 8. The networks listed above may be represented by the following function

$$y_k = f_k\left(\alpha_k + \sum_{j \to k} w_{jk} f_j\left(\alpha_j + \sum_{i \to j} w_{ij} x_i\right)\right). \qquad (2)$$

where $y_k$ is the output, $\alpha_k$ is the bias, $w_{ij}$ are the link weights, $f_j$ is the activation function, and the inputs are $x_j$. (Ripley (1996)) Equation (2) represents a general class of functions rather than just a parametric input-output association. (Ripley (1996))

The neural network literature finds that neural networks " . . . with linear output units and a

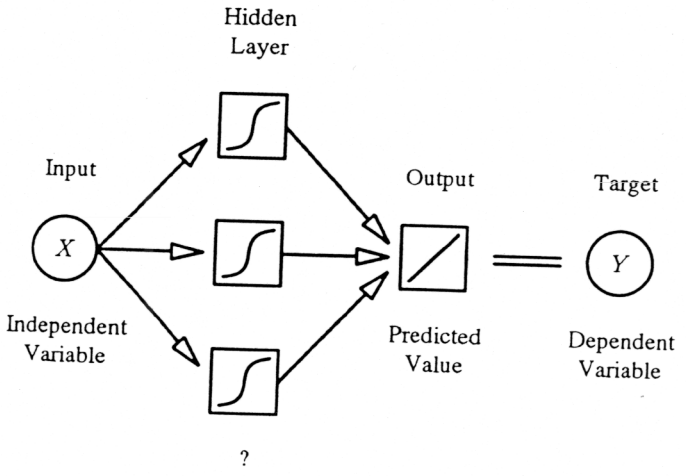


**Figure 6**

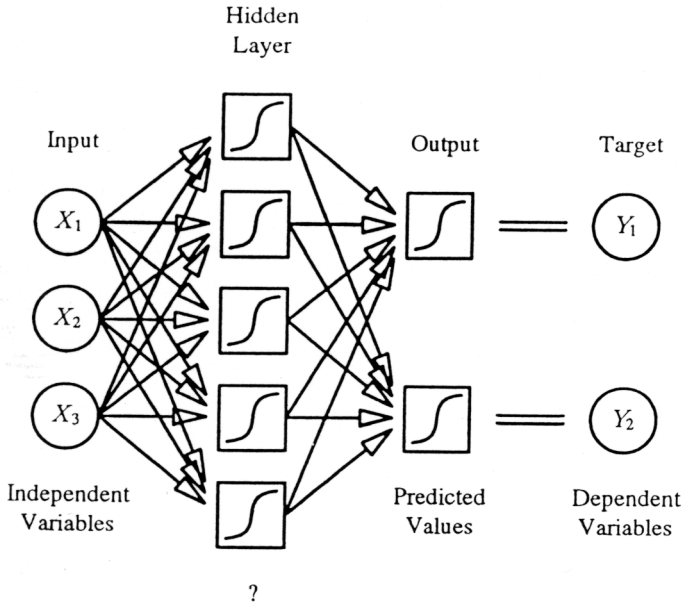


**Figure 7**

single hidden layer can approximate any continuous function *f* uniformly on compacta, by increasing the size of the hidden layer, and this implies many other types of approximation." (Ripley (1996), p. 147) See Ripley (1996) for detailed proofs of these results. By selecting the number of hidden layers and the number of hidden nodes in each layer it is possible to vary the complexity of the neural network model. (Sarle (1994)) Few hidden nodes are similar to polynomial regression. (Sarle (1994)) A process with a moderate number of hidden units begins to resemble projection pursuit regression, except that the neural network uses a rigid activation function, while projection pursuit utilizes a flexible nonlinear smoother. (Ripley (1996) and Sarle (1994)) If the number of hidden neurons increases with the sample size, the feed-forward network acts as a " . . . nonparametric sieve that provides a useful alternative to methods such as kernel regression and smoothing splines." (Sarle (1994), p. 5)

In short, "MLPs are general-purpose, flexible, nonlinear models that, given enough hidden neurons and enough data, can approximate virtually any function to any desired degree of accuracy. In other words, MLPs are universal approximators. MLPs can be used when you have little knowledge about the form of the relationship between the independent and dependent variables." (Sarle (1994), p. 5)

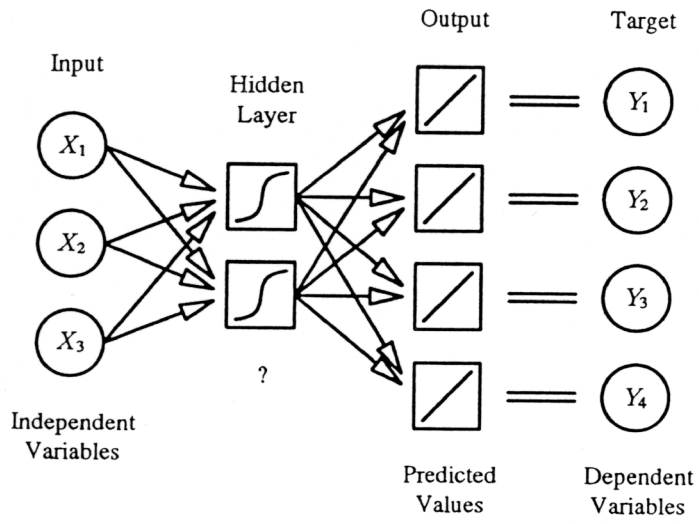


**Figure 8**

Polynomial regression, although fast, has difficulty approximating too many "wiggles" in a curve and tends to infinity when extrapolated. (Sarle (1994)) With smoothing splines it is necessary to determine the placement of the knots. (Sarle (1994)) Feed-forward neural networks take more computer time, but are more stable than higher order polynomials, and there are no knot locations to consider. (Sarle (1994)) Also MLPs are easily extended to multiple inputs and outputs without exponential growth in the number of parameters. (Sarle (1994))

MLPs are most commonly trained by the generalized delta rule of the Rumelhart-McClelland group. (Ripley (1996)) Their procedure was to use a steepest descent to minimize the total squared error of the output (1) with an update rule of

$$w_{ij} \leftarrow w_{ij} - \eta \frac{\delta E}{\delta w_{ij}}$$

and since the partial derivative can be expressed as

$$\frac{\delta E}{\delta w_{ij}} = \sum_{p} y_i^p \delta_j^p$$

This has become known as the generalized delta rule. (Ripley (1996)) "Further, as the $\delta$'s can be computed from output to input across the network both the process of calculating the derivatives and the descent algorithms are known as back-propagation." (Ripley (1996), p. 149)

According to Fausett (1994) back propagation is defined to be, "A learning algorithm for multi layer neural nets based on minimizing the mean, or total, squared error." (Fausett (1994), p. 423) To train a neural network by back propagation is a three-step process: "...the feed-forward of the input training pattern, the calculation and back propagation of the associated error, and the adjustment of the weights." (Fausett (1994), p. 290) The diagram below, taken from Caudill and Butler (1993), depicts the back propagation process.

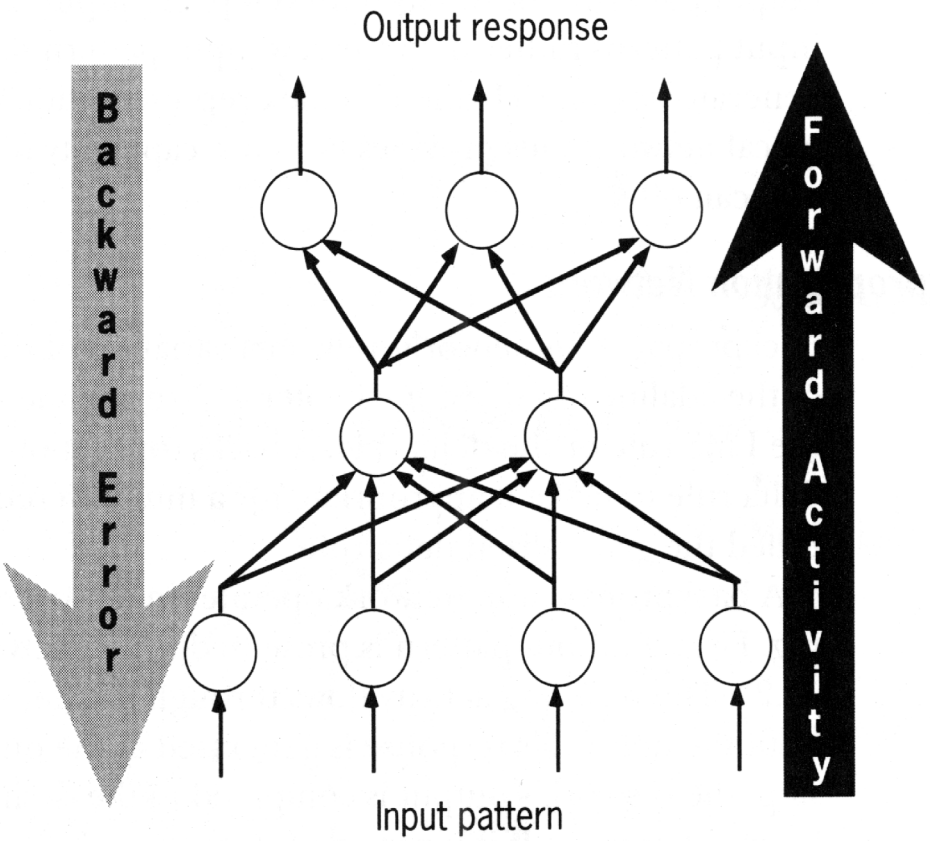


**Figure 9**

This learning method is just one from a family of supervised learning neural networks. The purpose of supervised learning is to predict one or more target variables from one or more input variables. (Sarle (1994)) "Supervised learning is usually some form of regression or discriminant analysis." (Sarle (1994), p.6) MLPs are the most common networks in the supervised learning family. (Sarle (1994)) Another example of this type of learning might is polynomial regression. A functional link network where the polynomial terms are given constant weight is shown in Figure 10. (Sarle (1994)) In general

functional links can be transformations of any type that do not require extra parameters, where the activation function is the identity. (Sarle (1994)) Other forms of supervised learning also include learning vector quantization and counter propagation. (Fausett (1994))

In contrast with supervised learning is unsupervised learning or training. In unsupervised training, the self-organizing neural networks, without the aid of supervised training, group similar inputs together and determine what an average element of each group resembles. (Fausett (1994)) No target values are explicitly provided. The net updates the weights so that the most similar input observations are grouped together in clusters of outputs. (Fausett (1994)) Then the neural network produces a representative input from each cluster. (Fausett (1994)) In this way the input values serve both as inputs and as targets. (Sarle (1994))

Unsupervised competitive learning produces binary outputs. This type of network is shown below in Figure 11. Note that only one output neuron has a value of 1 while all the rest are 0. This type of neuron is called the winner-take-all neuron, the architecture being called Kohonen networks. The winning neuron is the one with values most similar to the inputs as measured by an inner-product similarity measure.

A classification of neural networks can be made by determining whether or not they learn with supervision and whether or not they contain closed synaptic loops (feedback). A feedback neural network is one which simultaneously learns and recalls patterns. (Kosko (1992)) Both neurons and the weights change when these nets learn and when they recall. (Kosko 1992)) Figure 12 (Kosko (1992)) displays a general taxonomy

of several common neural network models.

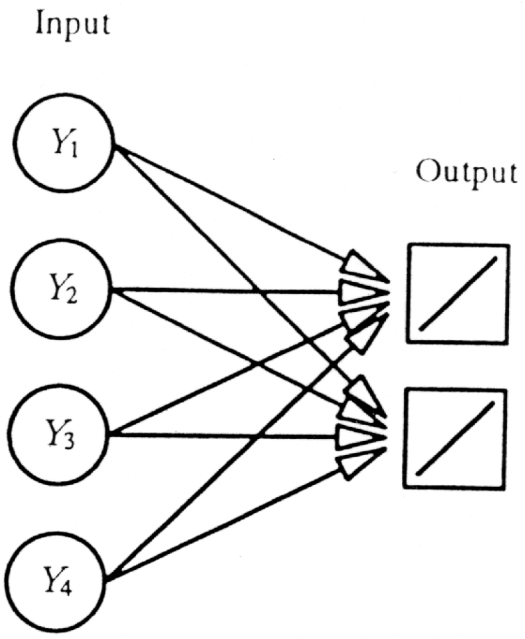


**Figure 10**

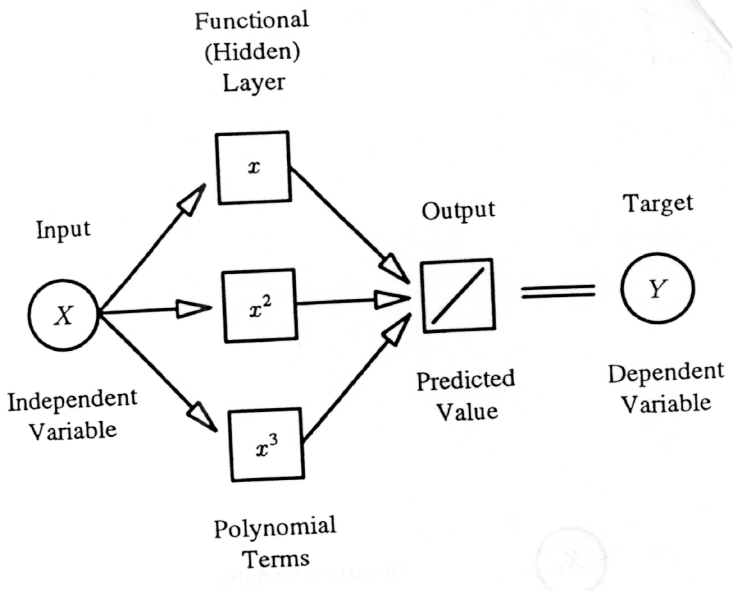


**Figure 11**

NEURAL NETWORK TAXONOMY

==========

DECODING

| | Feed-forward | Feedback |
|---|---|---|
| Supervised | Gradient Descent<br>LMS<br>Back propagation<br>Reinforcement Learning | Recurrent Back propagation |
| Unsupervised | Vector Quantization<br>Self-organizing maps<br>Competitive learning<br>Counter-propagation | RABAM<br>Brownian Annealing<br>Boltzmann Learning<br>ABAM<br>ART-2<br>BAM<br>Cohen-Grossberg<br>Hopfield Circuit<br>Brain-State-In-A-Box<br>ADAPTIVE RESONANCE<br>ART-1<br>ART-2 |

==========

Figure 12

> "Many NN models are similar or identical to popular statistical techniques such as generalized linear models, polynomial regression, nonparametric regression and discriminant analysis, projection pursuit regression, principal components, and cluster analysis, especially where the emphasis is on prediction of complicated phenomena rather than on explanation. ...There are also a few NN models, such as counter propagation, learning vector quantization and self-organizing maps, that have no precise statistical equivalent but may be useful for data analysis." (Sarle, (1994), p. 1)

It is doubtful that neural networks will ever supersede statistical methodologies, but they are not in competition with each other. (Sarle (1994)) As long as "model-free" procedures are needed, neural networks will be around filling this need, and the disciplines which are utilizing neural networks seem to be expanding.

One of the earliest uses for neural networks was as a pattern associator. (McClelland and Rumelhart (1986)) Specifically neural networks have been developed in the automatic recognition of handwritten characters, digits, and letters. (Fausett (1994)) MLPs trained by back propagation have been used for reading handwritten zip codes. (Fausett (1994))

Neural networks have also made inroads in the discipline of speech recognition. In the first stage the neural net distinguishes between vowels and consonants. (Fausett (1994)) The next stage shows the neural network recognizing the boundaries between words. (Fausett (1994)) "After as few as 10 passes through the training data, the text is intelligible. Thus, the response of the net as training progresses is similar to the development of speech in small children." (Fausett (1994), p. 10)

Neural networks as pattern associators are used in the medical community as electrocardiogram interpreters. (Edelstein (1996)) Another application found by Anderson

et al. in the mid-1980s is the idea of "Instant Physician." (Fausett (1994)) This autoassociative memory neural network "Brain-State-In-A-Box" contains a massive number of medical records complete with symptoms, diagnosis, and treatments. (Fausett (1994)) Upon completion of training, the neural network, when presented with symptoms, will respond with the "best" diagnosis and treatment. (Fausett (1994))

In the signal processing discipline, one of the earliest function estimation applications of neural networks was to suppress noise on a telephone line. (Fausett (1994)) An adaline type of neural network, called the adaptive filter, is used to remove echo from the incoming signal. (Fausett (1994)) Function estimation is a far more prevalent neural network application than pattern recognition. In business, one example might be the mortgage origination underwriting neural network. (Fausett (1994)) As function estimators neural networks are applied everywhere from routinely searching databases for credit card risks and loan default profiles to magazine renewal features. (Edelstein (1994)) Using neural networks as a tool, data warehousing firms are harvesting information from datasets in the process known as data mining. The difference between these data warehouses and ordinary databases is that there are actual manipulation and cross-fertilization of the data helping users make more informed decisions. (Aubrey (1996))

Data mining is the business of answering questions that you've not asked yet. (Edelstein (1996)) Data mining reaches deep into databases. (Edelstein (1996)) Typically there are five common types of information: associations, sequences, classifications, clusters, and forecasting. (Aubrey (1996)) "Associations happen when occurrences are linked in a single event." (Edelstein (1996), p. 48) With sequences the events are coupled

over time. The most common action in data mining is classification. "It recognizes patterns that describe the group to which an item belongs. It does this by examining existing items that already have been classified and inferring a set of rules." (Edelstein (1996), p. 48) Similar to classification is clustering. The major difference being that no groups have been predefined. The fifth application is forecasting. This is different from predictions because it estimates the future value of continuous variables based on patterns within the data.

Neural networks, depending on the architecture, provide associations, classifications, clusters, and forecasting to the data mining industry. It is just one of the four primary tools used in data mining: decision trees, rule induction, neural networks and data visualization. Which tool to utilize depends on the size and scope of the data warehouse it is operating in. (Edelstein (1996)) "Buying a $99 neural net program and throwing it against a terabyte data warehouse is not likely to produce any useful results." (Edelstein (1996), p. 48) Because neural networks don't have a clear interpretation of the model, " . . . it is cutting-edge software development best left to the experts." (Aubrey (1996), pp 568) Or at least someone with the statistical understanding and training to comprehend what the model is doing, and whether or not it is suitable to the process at hand. (Sarle (1994))

Instead of continuing to espouse ANNs ubiquity, let us cite two concrete examples of neural networks being used in the data mining context. The first example is from the IRS, which uses neural networks and polynomial regression to " . . . pinpoint potential tax noncompliance cases." (Mena (1994), p. 39) Their tools focused on feature classification and extraction, descriptive statistics, polynomial networks, and clustering algorithms. (Mena (1994)) Through these tools, the database variables were reduced from 150 to eight.

The accuracy of the polynomial network was confirmed by other neural networks. (Mena 1994)) These other networks included back propagation, a learning vector quantization, and a self-organizing map. (Mena (1994)) "Experimentation with the conversion and scaling of the data, the learning coefficients (momentum), the convergence thresholds, the number of hidden nodes, and other parameters of each network was performed through several training and testing sessions--with no improvement over the polynomial results." (Mena (1994), p. 39)

The second example, Wrangler, (a manufacturer of men's and boy's jeans, shirts and knitwear) is taken from industry. (*Intelligent Systems Report* (*ISR)* (1996)) Wrangler is using AI technology, namely neural networks " . . . in a reengineering effort to improve its total supply chain. . . . The most visible benefits have been increased sales volumes, lower inventory investments and improved inventory turns for retail customers and for Wrangler." (*ISR* (1996), p. 1) A feed-forward neural network is used to generate forecasts from consumer demand data which now drive production planning. The old way had been to let retail buyers' orders drive production. It has worked so well that retailers now share their advertising and promotion plans in an effort to fine tune inventories and replenishments. (*ISR* (1996)) "Better forecasts based on consumer demand, improved manufacturing processes, major reductions in cycle times, reengineered distribution systems, and ongoing retail [continuous replenishment program] CRPs have all contributed to reduced inventory, lower costs, higher sales, and improved profitability at Wrangler." (*ISR* (1996), p.1)

We have examined neural networks in closer detail and have discovered their similarity to statistical methods, their ability to be universal approximators, and their "model-free"

interpretations of the situation. They are flexible--the hidden nodes and hidden layers can vary, as well as the activation functions. As the sample size increases the number of parameters does not grow exponentially. Neural networks can be supervised or unsupervised, feed-forward or feedback. However, by far the most amazing quality is their dual nature to recognize patterns or estimate functions.

In data warehouses, neural networks are just one of the tools used in data mining. ANNs are used to find patterns in the data and to infer rules from them. Neural networks are useful in providing information on associations, classifications, clusters, and forecasting. Both the IRS and Wrangler have used neural networks in a data mining situations with good success. More examples would have been given, but the Internet search of data mining and neural networks revealed only these cases. We anticipate as time passes, and data mining grows more case studies will become available.

We have also seen that for best results with neural networks a working knowledge of statistical models is desired. With all the common material between the two disciplines, neural networks and statistics, better communication between them would be advantageous to both. Computers have a long way to go before they can rival the human brain on the same parallel scale but neural networks are a start in the right direction.

# References

Anderson, J.A. and Rosenfeld, E. (eds) (1988) Neurocomputing: Foundations of Research, The MIT Press, Cambridge, MA, USA.

Aubrey, D. (1996) "Mining for dollars," *Computer Shopper*, Aug, **16-8**, p. 568(3)

Caudill, Maureen and Butler, Charles (1993), Understanding Neural Networks, Volume 1: Basic Networks, The MIT Press, Cambridge, MA, USA.

Caudill, Maureen and Butler, Charles (1993), Understanding Neural Networks, Volume 2: Advanced Networks, The MIT Press, Cambridge, MA, USA.

Edelstein, H. (1996) "Mining data warehouses," *InformationWeek*, Jan 08, **561**, p. 48(4).

Fausett, Laurene (1994), Fundamentals of Neural Networks: Architectures, Algorithms and Applications, Prentice-Hall, New Jersey, USA.

Kosko, Bart (1992), Neural Networks and Fuzzy Systems, Prentice-Hall, New Jersey, USA.

McClelland, J.L. and Rumelhart D.E. (1986) , Explorations in Parallel Distributed Processing: A handbook of Models, Programs, and Exercises, The MIT Press, Cambridge, MA, USA.

McCulloch, W.S. and Pitts, W. (1943) "A logical calculus of ideas immanent in neural activity," *Bulletin of Mathematical Biophysics*, **5**, 115-133. Reprinted in Anderson and Rosenfeld (1988).

Mena, J. (1994), "The adaptive tax collector; AI modeling tools are helping the Internal Revenue Service detect fraud, reduce risk, and improve collections," *AI Expert*, Nov, **9**, p. 39(3).

Sarle, Warren S. (1994), "Neural Networks and Statistical Models," *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, April, pp 1-13.

"Wrangler replenishes with neural nets", *Intelligent Systems Report*, Jan **13-1**, http://lionhrtpub.com/ISR/ISR-1-96/1-96-wrangler.html.