Comparing Algorithms and Clustering Data: Components of the Data Mining Process

A thesis submitted to the Department
of Computer Science and Information Systems at
Grand Valley State University
in partial fulfillment of the requirements
for the degree of
Master of Science

By
Glenn A. Growe
December, 1999

Glenn A. Growe
Department of Computer Science and Information Systems
Grand Valley State University
Mackinac Hall
Allendale, Michigan 49401

groweg@gvsu.edu

Acknowledgements

Several people contributed significantly to the quality of this project. Professor Paul Jorgensen taught me to appreciate the role of creative thought in Computer Science. Professor Nenad Jukic broadened my knowledge of database and data analysis. Professor Soon Hong taught me efficient methods for conducting statistical analyses. And all of the above members of my committee have been patient, supportive, and encouraging.

David Wishart of Clustan Ltd. provided helpful comments on this thesis in its proposal form. Tjen-Sien Lim of the Department of Statistics at the University of Wisconsin did likewise. Benjamin Herman, formerly of the University of Chicago, now at Brown University, provided helpful input on the topic of ROC analysis. The limitations of this study are, of course, my own responsibility.

Thanks to my wonderful wife Betty. She took care of many of the practical details of family life, allowing me to focus on this project. Thanks and a hug to our children for putting up with the long hours dad spent (and still spends) at his "puter."

The soil is refreshed when sown with successive changes of seed, and so are our minds if cultivated by different subjects.

The Letters
Pliny the Younger

…the wisdom of mortals consists…not only in remembering the past and apprehending the present, but in being able, through a knowledge of each, to anticipate the future, which grave men regard as the acme of human intelligence.

The Decameron
Giovanni Boccaccio

Abstract

Thirteen classifiers, including neural networks, statistical methods, decision trees, and an instance-based method were employed to perform binary classifications on twelve real-world datasets. Predictive classification performance on test sets was compared using ROC analysis and error percentage. The four best algorithms were neural networks. The hypothesis of no difference between the error rates of the algorithms was rejected by statistical test. The amount of difference in the quality of performance of the classifiers seems to be a characteristic of the dataset. On certain datasets almost all algorithms worked about equally well. For other datasets there are marked differences in algorithm effectiveness.

An attempt to improve classification accuracy by pre-clustering did not succeed. However, error rates within clusters from training sets were strongly correlated with error rates within the same clusters on the test sets. This phenomenon could perhaps be used to develop confidence levels for predictions.

Contents

Chapter 1. Introduction

Progress in storage technology is allowing vast amounts of raw data to accumulate in both private and public databases. It has been estimated that the amount of data in the world doubles every twenty months (Frawley, Piatetsky-Shapiro, & Matheus, 1992). Insurers, banks, hotel chains, airlines, retailers, telecommunications and other enterprises are rapidly accumulating information from day to day transactions with their customers. Wal-Mart every day uploads twenty million point-of-sale transactions into a centralized database (Cios, Pedrycz, & Swiniarski, 1998). As John (1997) writes:

> Knowledge may be power, but all that data is unwieldy.

Statistician David Wishart (1999a) comments:

> Computers promised a fountain of wisdom, but delivered a flood of data.

Unless the accumulated data can be adequately analyzed it becomes useless. To help put this flood of data into a format that can be used, more and more data is being moved into data warehouses whose purpose is decision support. Data warehouses help by having a common format and consistent definitions for fields.

The process of turning some of this stored data into knowledge is the domain of knowledge discovery in databases. Knowledge discovery in databases has been defined as follows:

> Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data (Fayyad, Piatetsky-Shapiro, and Smyth (1996).

Data mining is a component of the knowledge discovery in databases process concerned with the algorithmic means by which patterns are extracted and enumerated from data. (Fayyad, Piatetsky-Shapiro, and Smyth, 1996).

Data mining helps businesses and scientists discover previously unrecognized patterns in their databases. These patterns may help a consumer products company optimize inventory levels and detect fraudulent credit-card transactions. They can help a telecommunications company identify who is likeliest to move to another long-distance phone company. They may help a doctor predict which patients are good candidates for a surgical procedure or are at risk for developing a particular disease.

This knowledge discovery process has several steps. The first step is to define the problem. Often working with a domain expert, the data mining analyst needs to define

specific problems or questions to be answered. The second step is to extract the data, often from several different tables in a database and place it into one table against which data mining algorithms can be run. The third step is to "clean" and explore the data for such things as mislabeled fields and special semantics. Special semantics refers to the practice of assigning numerical values such as zero or 99 to attributes whose actual value is unknown (George, 1997). Next the data is "engineered." The data may be transformed to insure that the data all has the same scale. There may be a decision to drop certain records if they are incorrect or represent cases that could not be used to infer general patterns. Then the analyst selects an algorithm to analyze the data from among the many available or develops an algorithm. Finally, he runs the algorithm(s) on a subset of the data, holding back a portion on which to validate the discovered patterns. Our contributions in this thesis will be to the last two steps of the knowledge discovery process. We will look at choosing algorithms and at clustering data to improve accuracy.

We will examine several statistical and artificial intelligence (AI) methods used to perform various classification tasks. Wilson (1997) defines this problem as follows:

> The problem of classification…is to learn the mapping from an input vector to an output class in order to generalize to new examples it has not necessarily seen before.

Classification rather than continuous function approximation will be the focus because it is the most common question to be answered in data mining situations. Binary classification is the frequently encountered situation where there are two categories. A set of cases or instances is partitioned into two subsets based on whether each has or does not have a particular property. Binary classification is also our focus because there are clear criteria for judging binary classification efforts - percentage correctly classified and receiver operating characteristic (ROC) curves.

Increased knowledge of the accuracy of various classification methods will allow data mining analysts to select from those that are most effective. Knowledge of which classifiers perform best may suggest directions for those seeking to construct new algorithms or to improve upon existing ones.

There is controversy over the relative merits of doing classification using AI tools such as neural networks versus employing statistical methods. Thus, David Banks, (1996) a statistician, writes about neural networks in this tone:

> Computer science has recently developed a new drug, called neural nets.…I come to bury, not to praise.(pp.2,3)

He goes on to cite experimental work which found no marked superiority for neural nets over newer statistical techniques in classification tasks.

Those in the neural network camp almost universally boast of the superiority of neural networks over statistical methods. For example, NeuralWare, Inc. (1991) states in a book it published on neural computing:

> Neural computing systems are adept at many pattern recognition tasks, more so than both traditional statistical and expert systems… The ability to select combinations of features pertinent to the problem gives them an edge over statistically based systems. (p.10)

Advocates of neural networks often claim that statistical models have difficulty dealing with the contradictory and messy data often found in real-world datasets. They feel statistical methods only work with data that is "clean" and which contain consistent correlations. They note that neural networks can fit complex non-linear models to the data while some statistical methods can accommodate only linear relationships.

A third view is that both approaches are evenly matched and which approach is best will depend upon the problem domain. Couvreur and Couvreur (1997) write:

> For us, statistics and neural networks are complementary tools, with considerable overlap not only in their fields of application but also in their theoretical foundations… When compared fairly, neural and modern statistical approaches perform similarly, both in terms of quality of results and in terms of computational cost. In some applications NN's will outperform their statistical counterparts, in others they will not (pp. 2, 5).

We propose to compare the accuracy of various AI and statistical methods on several classification tasks. There may be generalizations that can be drawn about the types of data sets for which certain methods are most appropriate. Statistical methods used in the comparison will include decision trees (CART, CHAID, and QUEST), discriminant analysis, and logistic regression. AI approaches will include various multi-layer perceptron neural networks, learning vector quantization (LVQ) neural networks and other related supervised learning methods.

The need for research comparing classification algorithms is great. Salzberg (1997) writes:

> Classification research, which is a component of data mining as well as a subfield of machine learning, has always had a need for very specific, focused studies that compare algorithms carefully. The evidence to date is that good evaluations are not done nearly enough…(Salzberg, 1997)

Prechelt (1996) surveyed nearly 200 papers on neural network algorithms. Twenty-nine percent were not evaluated on any real-world data. Only 8% compared the algorithm to more than one alternative classifier on real data.

There are several sites on the web where interesting real-world data sets for doing classification can be found. Most of the data that we will use in this thesis comes from such sources. Real world datasets are used with the idea that good performance on them will generalize to similar performance on other real-world tasks.

A second focus of this thesis arises out of the proposition that large datasets may not yield certain significant patterns until they are divided into more homogeneous subgroups through cluster analysis techniques. This is seen as improving the performance of more directed or "supervised" learning methods that are then applied to the subgroups created instead of to the entire dataset.

Chapter 2. Comparing the Algorithms

Classification as we shall use it in this chapter refers to establishing rules so that we can classify new observations into one of a set of existing classes. Observations have attributes. The task of the classifier is to assign an observation to a class given its set of attributes. The rules may be explicit or comprehensible, as in the case of decision trees. Or, as with neural networks, rules may not be capable of explicit formulation.

We assume that we have a number of sample observations from each class. The classifier is presented with a substantial set of the data from which it can associate known classes with attributes of the observations. This is known as training. When such guidance is given the process is known as supervised learning. The rules developed in the training process are tested on the remaining portion of the data and compared with the known classifications. This is known as the testing process. Here the response of the procedure to new observations is a prediction of the class to which the new observations belong. The proportion correct in the test set is an unbiased estimate of the accuracy of the rules implemented by the classifier.

Much of the knowledge gained in the data mining process is in the form of predicted classifications. Customers may be classified as likely or not likely to respond to a bank's solicitations to take out a home loan. Medical patients may be classed at high or low risk for heart disease based upon risk factors.

There is another important type of classification based upon the concept of clustering. Here neither the number of classes is known in advance nor is the assignment of observations to classes. We shall discuss the relevance of this type of classification to predictive classification in chapter four.

2.1 Previous Comparative Studies

Several previous studies have compared classifiers. The most inclusive study ever done comparing different classifiers was the STATLOG Project carried out in Europe (Michie, Spiegellhalter, and Taylor, 1994). They proceeded from the assumption that the fragmentation among the different disciplines that have worked on classification

problems has hindered communication and progress. They sought to remedy this by bringing together a multidisciplinary team and including classifiers developed by the different disciplines. Included were procedures from classical statistics, modern statistics, decision trees, and neural networks. They considered the results of 22 algorithms from the above areas run on 16 datasets. The datasets were diverse. They included such problems as assessing Australian credit card applicants, recognizing handwritten digits, determining type of ground cover from Landsat satellite images, and predicting recovery level from head injury based upon data collected at the time of injury. The result was that the procedures that worked best varied by dataset. The three individual procedures most often among the best for each of the datasets were one type of neural net (DIPOL92) and two types of statistical procedures (ALLOC 80 and logistic discriminant analysis). Decision trees performed well if the dataset was multimodal. There were other variations. Among the decision tree group of methods almost all performed about the same. Among the neural nets one type was frequently one of the best overall (DIPOL92) and another type was rarely among the best (Kohonen's LVQ).

Shalvik, Mooney, and Towell (1991) compared backpropagation and the ID3 type of decision tree on five real-world data sets. They found backpropagation superior to the decision tree on two datasets with no difference on the other three.

Brown, Corruble, and Pittard (1993) compared backpropagation neural networks and decision trees for multimodal classification problems. Decision trees performed better on datasets which contained irrelevant attributes which they were able to ignore. Neural networks do not have such a capacity for feature selection. Apparently, the neural networks were confused by the presence in the training set of attributes not useful in discriminating the target classes. On two other datasets in which most variables were useful in discriminating the classes neural networks outperformed decision trees. This suggests that to get the best performance from neural networks a procedure to select out the best input variables prior to training is sometimes necessary. Another interesting finding was that neural networks with two hidden layers outperformed those with just one hidden layer.

Ripley (1994) compared discriminant analysis, nearest neighbor, backpropagation neural networks, MARS, and a classification tree on a few classification problems. The measure was percentage correctly classified. The various tools were approximately equally matched. Ripley concluded that:

> Neural networks emerge as one of a class of flexible non-linear regression methods which can be used to classify via regression (p 409).

Curram and Mingers (1994) compared discriminant analysis, decision trees, and neural networks across seven datasets. Four contained real data and three were artificially created. Discriminant analysis performed well when the dataset proved to be linearly separable. On a dataset that was designed to have highly non-linear relatoinships (points classified as either inside or outside a sphere based on their three coordinates) discriminant analysis performed at a chance level. Neural networks performed well on the sphere data and fairly well across all datasets. It did better than discriminant analysis

when there were non-linear relationships between predictors and classes but slightly worse when the data were linearly separable. Decision trees performed worse than the other two methods. It was interesting that on the real world datasets, where its assumptions were likely not strictly adhered to, discriminant analysis proved to be reasonably robust.

Holmstrom, Koistinen, Laaksonen, and Oja (1997) compared several classifiers on handwritten character and phoneme data using percent accurately classified. The two datasets have very different statistical properties. The handwriting data is high dimensional while the phoneme data is low dimensional. The handwriting data has many classes while the phoneme data has just two. The phoneme data is described as having a rich internal structure with a class distribution containing many clusters. Thirteen classifiers were employed. They included variations of classical discriminant analysis, regression-based methods such as MARS, subspace classifiers, nearest-neighbor methods, and two types of neural networks. In the classification of handwritten digits the nearest neighbor and subspace classifier techniques were most effective. A decision tree classifier had the highest error percentage. Combining three classifiers in a "committee" using a majority voting rule for classification provided an improvement over using a single classifier. On the phoneme classification problem kernel classifiers, and nearest neighbor classifiers performed best. Classifiers with relatively simple decision boundaries performed poorly on this dataset. Such results indicate that characteristics of particular datasets are an important determinant of which classification tool will perform best. This also suggests that it will be futile to try to discover one classification tool that will perform best across all datasets.

Lim, Loh, and Shih (1999) compared twenty-two decision tree, nine statistical, and two neural network algorithms in terms of classification accuracy. They assessed classification accuracy by mean error rate and mean rank of error rate. The best methods were a statistical algorithm called POLYCLASS - a spline based "modern version of logistic discrimination analysis." Other top-ranked algorithms were linear discriminant analysis, logistic discriminant analysis, and the decision tree algorithm QUEST with linear splits. The two neural network algorithms (LVQ and radial basis function) were both in the bottom fourth of the methods used. However, more modern and perhaps more powerful neural network algorithms, such as backpropagation, were not used.

Other papers have compared classification approaches on a single dataset. Dietterich, Hild, and Bakiri (1989) compared the performance of a backpropagation neural network and a decision tree algorithm known as ID3. The classification task was the mapping of English text to phonemes and stresses. Backpropagation consistently outperformed the decision tree by several percentage points. The authors comment that there is no universal learning algorithm that can take a sample of training examples for an unknown function and produce a good approximation. Instead, every learning algorithm has its own biases about the nature of the problem to be learned. The difference in performance between backpropagation and ID3 means that they make different assumptions.

Chen (1991) compared three types of neural networks (backpropagation, radial basis functions, and probabilistic neural networks) with the statistical method of nearest neighbor decision rule. The classification target was simulated active sonar waveforms. All three neural networks outperformed nearest neighbor. More advanced statistical techniques were not included.

Sandholm Brodley, Vidovic, and Sandholm (1996) compared six algorithms in predicting morbidity and mortality from equine gastrointestinal colic. The high mortality rate with surgery (40%) and the high cost of the operation (about $10,000) are reasons for only operating on horses that actually have the disease and will likely survive the operation. Linear discriminant, logistic regression, and a neural network did slightly better than a decision tree and a nearest neighbor algorithm. But the results from the neural network were seriously flawed because the test data used in the comparison was also used to choose the best time to stop training the neural net and to set other important aspects of the network's architecture.

Poddig (1995) predicted which of a set of French firms fell into bankruptcy. The predictive attributes were 45 ratios developed from the firm's financial statements 1-3 years before some entered bankruptcy. A backpropagation neural network with multiple hidden layers exceeded the performance of discriminant analysis. Kohonen's LVQ network underperformed the discriminant analysis.

Sen, Oliver, and Sen (1995) compared neural networks and logistic regression in predicting which companies would be merged with other companies. The two techniques performed equally well.

Schwartz Ward, MacWilliam, and Verner (1997) used fourteen variables as potential predictors for improvement after total hip replacement surgery. A neural network was compared with a linear regression model using the same data. Using a receiver operating characteristic (ROC) curve for comparison the neural network was more accurate but the difference did not reach statistical significance.

Pesonen (1977) compared discriminant analysis, logistic regression analysis, and cluster analysis with a backpropagation network in the diagnosis of acute appendicitis. Input variables were 17 clinical signs and age and sex of patients admitted to a hospital suffering from acute abdominal pain. The results of the four classification methods were compared with receiver operating characteristic curve (ROC) analysis as well as by diagnostic accuracy. Discriminant analysis and backpropagation showed slightly better results than the other methods. Interestingly, he found that predicting that a case was acute appendicitis only when all methods agreed on the diagnosis increased accuracy. Pesonen concluded that backpropagation neural networks do not offer any magic but do perform as well as statistical methods.

2.2 Descriptions of Classification Algorithms

The following is a listing of all the supervised learning methods we use:

1. Discriminant Analysis
2. Logistic Regression
3. Classification and Regression Trees (CART)
4. Chi-squared Automatic Interaction Detection (CHAID)
5. QUEST decision tree
6. Model Ware
7. Model Quest
8. Multi-Layer Perceptron neural net (MLP) - Backpropagation Learning
9. MLP Cascade Correlation neural net
10. Learning Vector Quantization (LVQ) neural net
11. MLP Levenberg-Marquardt neural network
12. Resilient Propagation
13. Ward Systems Classifier

Discriminant analysis is the oldest statistical technique for classification. R.A. Fisher first published it in 1936. In it the difference between two classes is maximized by a linear combination of variables. This linear function acts as a hyper-plane that partitions the observation space into classes. Which side of a hyper plane a point falls into determines its classification. Discriminant analysis assumes that the predictor variables are normally distributed. We will use the implementation of discrminant analysis provided in SPSS Version 8.0.

Logistic regression is a version of linear regression used for predicting a classifying variable. Logistic regression builds up a linear model using the logarithm of the odds of occurrence of a class membership. In logistic regression the modeler must select the right variables and account for their possible interactions. There is no normality assumption imposed upon the data. We will use the implementation of logistic regression provided in SPSS Version 8.0.

Decision trees develop a series of rules that classify observations. We will use three types - CART (known as "C&RT" in SPSS's version), CHAID, and QUEST. In all decision trees an observation enters at the root node. A test is applied which is designed to best separate the observations into classes. This is referred to as making the groups "purer." The observation then passes along to the next node. The process of testing the observations to split them into classes continues until the observation reaches a leaf node. Observations reaching a particular leaf node are classified the same way. Many leaves may make the same classification but they do so for different reasons. Decision trees differ from the classical statistical tests in that they do not draw lines through the data space to classify observations. Decision trees may be thought of as drawing boxes around similar observations. Several different paths may be followed for an observation to become part of a particular class. Criticisms of decision trees include that any decision on how to split at a node is made "locally." It does not take into account the effect the

split may have on future splits. And the splits are "hard splits" that often may not reflect reality. Thus an attribute "years of age" may be split at "age > 40." Is someone thirty-nine so different than a forty-one year old? Also, splits are made considering only one attribute at a time (Two Crows Corporation, 1998).

Brieman, Friedman, Olshen, and Stone developed the CART algorithm in 1984. It builds a binary tree. Observations are split at each node by a function on one attribute. The split is selected which divides the observations at a node into subgroups in which a single class most predominates. When no split can be found that increases the class specificity at a node the tree has reached a leaf node. When all observations are in leaf nodes the tree has stopped growing. Each leaf can then be assigned a class and an error rate (not every observation in a leaf node is of the same class). Because the later splits have smaller and less representative samples to work with they may overfit the data. Therefore, the tree may be cut back to a size which allows effective generalization to new data. Branches of the tree that do not enhance predictive classification accuracy are eliminated in a process known as "pruning."

CHAID differs from CART in that it stops growing a tree before overfitting occurs. When no more splits are available that lead to a statistically significant improvement in classification the tree stops growing. Also, any continuously valued attributes must be redone as categorical variables. The implementations of CART and CHAID we will use are from SPSS's Answer Tree Version 2.0.

QUEST is another type of decision tree developed by Loh and Shih (1997). It is unique in that it performs approximately unbiased as to class membership variable selection to split nodes. We will use the implementation of QUEST with linear combination splits available from http://www.stat.wisc.edu/~loh/quest.html.

Model Ware is a modeling tool that can be applied to signal processing, decision/control and classification problems. Model Ware learns from examples via the "Universal Process Algorithm" (UPM). It is in some ways similar to a nearest neighbor algorithm. The UPM requires a set of example data, known as the reference data file. This describes how the system or process behaves under known operating conditions. When it receives an input vector UPM creates a localized model based on a subset of the patterns from the reference library. The selection of exemplars is based on a metric of the similarity of the test vector to each pattern in the reference library. After the exemplars are selected the model computes the response vector. UPM also outputs diagnostic information indicating the quality of each component of the input vector and the overall system health. (Teranet Incorporated, 1992).

The version of Model Ware used in this study is no longer sold. The company that created it markets a product called Model Ware/RT. It is based on UPM's capacity to output diagnostic information about each component of the input vector and about overall system health. This product is marketed exclusively to the semiconductor industry. It is used there in a real-time mode to detect faults in semiconductor manufacturing (O'Sullivan, Martinez, Durham, and Felker, 1995). Model Ware was included in the present study because of evidence that it excels at classification problems (Hess, 1992).

Model Quest (AbTech Corporation, 1996) automatically constructs polynomial networks from a database of input and output values for example situations. The attributes used and their coefficients and the number and types of network elements, network size and structure, and network connectivity are all learned automatically. ModelQuest constructs a network by sequentially hypothesizing many potential network configurations and then rating them according to the predicted square error (PSE) criteria. The PSE test is employed to avoid overly complex networks that perform well on the training data but will perform poorly on future data. Model Quest was originally developed within the U.S. Military for target classification and other purposes. It is currently commercially available and is widely used in data mining applications.

A neural network is a group of highly interconnected processing elements that can learn from information presented to them. Neural networks were inspired by the structure of neuronal connections in the human brain. The neural network's ability to learn and its basis in the biological activities of the human brain classify it as a form of artificial intelligence.

The most widely used neural network is the multi-layer perceptron (MLP) type neural network. MLP networks process information in interconnected processing elements called nodes. Nodes are organized into groups known as layers. An MLP network consists of an input layer, one or more processing layers, and an output layer. The nodes of adjacent layers are connected to transfer the output signals from one layer to the next. Each input connection to a node has a weighting value associated with it. The node produces a single output that is transmitted to many other processing elements. Processing continues through each layer of the network. The network's response emerges at the output layer. During the training process the network's response at the output layer is compared to the known to be correct answers from a training set.

In the most common learning process used by MLP's the difference between the network's output and the correct responses are figured and this error is backpropagated through the network to improve its response. The procedure of processing inputs through the network, figuring errors, and sending the errors back through the network to adjust the weights constitutes the learning process in the backpropagation type of multi-layered perceptrons. Connection weights are adjusted to drive the error to a minimum. Neural networks resemble a directed graph with nodes, connections, and a direction of flow. Vesta Services (1996) produced the MLP using the backpropagation learning method that we will use (QNET).

Cascade correlation is another type of MLP that begins with no nodes initially and then adds them one at a time. Each new node receives inputs from the inputs and the other nodes in the network. Weights for the new nodes are not determined by minimizing mean squared error, as in backpropagation. Rather, the covariance between a new node and the residual error is maximized. Logical Designs Consulting (1994) developed the implementation of cascade correlation we will use.

Another type of MLP uses a particular method to adjust the difference between network outputs and target outputs during training. The Levenberg-Marquardt type of training method has space requirements proportional to the square of the number of weights in the network. This means that networks with a large number of connections between inputs and hidden nodes may be precluded. Hema Chandrasekaran (n.d.) developed the version we will use.

Resilient Propagation is a MLP neural network modified from backpropagation to train more efficiently. The implementation of resilient propagation is from QwikNet v. 2.23 (Jensen, 1999).

NeuroShell Classifier is a neural network using a proprietary algorithm (Ward Systems, 1998). While details of its structure are unavailable it is a tool which might well be selected by those in data mining.

Learning Vector Quantization (LVQ) is a "nearest neighbor" neural net in which each node is designated, via its desired output, to belong to one of a number of classes. The LVQ algorithm involves the use of codebook vectors. These are points within the problem space to approximate various modes of the problem. Several codebook vectors are usually assigned to each class. New patterns are classified based on the class assignment of the codebook vector that is closest to its position. The training process involves iteratively adjusting the positions of the codebook vectors in order to create a distribution that will minimize overall classification error. Logical Designs Consulting (1994) created the implementation of LVQ that we will use.

2.3 Assessing Classification Tool Performance

While we seek to determine the fitness of each algorithm the results obtained when a technique is applied to data may depend upon other factors. These include the implementation of the technique as a computer program and the skill of the user in getting the best out of the technique.

We will use several metrics to assess the performance of classification tools. The first is the traditional one of percentage of cases in the test set incorrectly classified (mean error rate). We will average this number across all datasets to give us a measure of a classifier's overall effectiveness. We will also examine the ranks of the classifiers within datasets. The classifiers with the lowest error rate will be assigned a rank of one, the one with the second lowest error rate will be assigned a rank of two, etc. The average ranks will be assigned in the case of ties.

It has been shown that there are problems with using accuracy of classification estimation as a method of comparing algorithms (Provost, Fawcett, and Kohavi, 1998). It assumes that the classes are distributed in a constant and relatively balanced fashion. But class distributions may be skewed. For example, if your classification task is screening for a rare disease, calling all cases "negative" can lead to a spuriously and trivially high

accuracy rate. If only .1 percent of patients has the disease a test that says no one has the disease will be correct 99.9% of the time. Accuracy percentage is affected by prevalence rates and there is no mathematical way to compensate for this.

Accuracy is also of limited usefulness as an index of a classifier's performance because it is insensitive to the types of errors made. Using classification accuracy as a measure assumes equal misclassification costs - a false positive has the same significance as a false negative. This assumption is rarely valid in real-world classification tasks. For example, one medical test may have as its mistakes almost all false negatives (misses). Another might err in the direction of false positives (false alarms). Yet these two tests can yield equal percentages of correctly classified cases. If the disease detected by the test is a deadly one a false negative may be much more serious than a false positive. Similarly, if the task is classifying credit card transactions as fraudulent the cost of misclassifying a transaction as fraudulent (false alarm) may be much less than missing a case of fraud.

The limitations of using classification accuracy can be overcome by an approach known as receiver operating characteristic (ROC) analysis (Metz, 1978; Swets, 1973). This is the second metric we shall use to evaluate classifier performance. We can begin our look at it by defining decision performance in terms of four categories:

$$\frac{\text{True Positive Decisions}}{\text{Actually Positive Cases}} = \text{True Positive Fraction (TPF)}$$

$$\frac{\text{False Positive Decisions}}{\text{Actually Negative Cases}} = \text{False Positive Fraction (FPF)}$$

$$\frac{\text{True Negative Decisions}}{\text{Actually Negative Cases}} = \text{True Negative Fraction (TNF)}$$

$$\frac{\text{False Negative Decisions}}{\text{Actually Positive Cases}} = \text{False Negative Fraction (FNF)}$$

Since all observations are classified as either positive or negative with respect to membership in a class the number of correct decisions plus the number of incorrect decisions equals the number of observations in that class. Thus, the above fractions are related by:

$$TPF + FNF = 1$$

and

$$TNF + FPF = 1$$

FNF can always be computed from knowledge of TPF. TNF can be computed from knowledge of FPF. It is necessary to know only one fraction from each of the above relations to determine all four of the types of decision fractions.

These concepts allow us to sort out the effects of the prevalence of a class. It also allows us to score separately the performance of a classifier with respect to observations that actually are and are not members of a class.

When we use a classification algorithm its output does not necessarily automatically cause an observation to fall into a particular class. If we have a two category classification problem predicted by one output the distribution of results from observations in the "0" class and from those in the "1" class will overlap (since the test is not perfect). A threshold value for allocating predictions to "0" or to "1" must be chosen arbitrarily. A different choice of threshold yields different frequencies for the types of correct and incorrect decisions. If we change the decision threshold we will obtain a different set of decision fractions. Because TPF and FPF determine all of the decision fractions we just keep track of how they change as the decision threshold is varied. The points representing all possible combinations of TPF and FPF lie on a curve that is called the receiver operating curve (ROC) for a classifier. It is called this because the receiver of the classifier information can "operate" on any point on the curve given a particular decision threshold.

In ROC space the TPF is typically plotted on the Y-axis and the FPF is plotted on the X-axis. If the classifier provides valid information the intermediate points on the ROC curve must be above the lower left to upper right diagonal. When this is so a decision to place an observation in a class when it actually is a member of that class is more probable. A ROC curve illustrates the tradeoffs that can be made between TPF and FPF (and hence all four of the decision fractions).

ROC analysis gives us another perspective on the performance of classifiers. An ROC curve shows the performance of a classifier across a range of possible threshold values. The area under the ROC curve is an important metric for evaluating classifiers because it is the average sensitivity across all possible specificities. One point in ROC space is better if it is to the upper left in the ROC chart. This means TPF is higher, FPF is lower, or both. A ROC graph permits an informal visual comparison of classifiers. If a classifier's ROC curve is shifted to the upper left across all decision thresholds it will perform better under all decision cutoffs. However, if the ROC curves cross then no classifier is best under all scenarios. There would then exist scenarios for which the model giving the highest percentage correctly classified does not have the minimum cost. The computer program we will use for figuring ROC curves was developed by Charles Metz, Ph.D. of the Department of Radiology at the University of Chicago (Metz, 1998).

Bradley (1997) investigated the use of the area under the ROC curve (AUC) as a measure of a classification algorithm's performance. He compared six learning algorithms on six real-world medical datasets using AUC and conventional overall accuracy. AUC showed increased sensitivity (a larger F value) in analysis of variance tests. It was also invariant to a priori class probabilities. Bradley recommended that

AUC should be used in preference to overall accuracy as a single number evaluation of classification algorithms.

A major limitation of ROC analysis is that it can only analyze classifier output that is continuously distributed. Many classification algorithms, notably decision trees, can only have discrete outputs (i.e., "1" or "0"). Hence, ROC analysis can be used with most, but not all of the classification algorithms used in this study.

2.4 The Data

We have included twelve datasets in our study. They are described briefly below. Any modifications we need to make to them for our study are also noted. We will remove those observations or cases containing missing data from all datasets.

Breast Cancer Survival This sample relates age at time of operation and number of positive axillary nodes to five-year survival after surgery for breast cancer. There are 306 cases in this dataset from the University of California at Irvine's (UCI) Machine Learning Repository (Blake, Keogh, and Merz, 1998).

Cleveland Clinic Heart Disease Here we are classifying patients as having or not having heart disease based upon 12 cardiac functioning variables. Disease is defined as having a greater than 50% narrowing of arteries on angiographic examination. There is complete data for about 287 subjects in this dataset also obtained from UCI.

Contraceptive Method Choice This data obtained from the UCI database was originally collected by the National Indonesia Contraceptive Prevalence Study in 1987. The data consists of nine demographic attributes for 1,473 married women. The data is modified slightly from the original dataset to include two classifications - does or does not use contraception.

Doctor Visits This dataset contains data on a sample of elderly individuals drawn from the National Medical Expenditure Survey done in 1987. There are 4406 observations and 22 variables. The data was used in a paper from the Journal of Applied Econometrics (Deb and Trivedi, 1997). This journal maintains a site where data from its articles is deposited and can be accessed (http://qed.econ.queensu.ca/jae/).

Earnings This dataset is from Polachek and Yoon (1996) who studied income using data from the Michigan Panel Study of Income Dynamics. Predictors are education (years), job experience (years) and tenure at current job (months). The dependent variable is whether wage level is above or below average. The number of observations is 13,408.

Indian Rice Farm This dataset comes from a forthcoming paper in the Journal of Applied Econometrics by Horrace and Schmidt (in press). The target variable is whether a farm in a village in India is classified as above or below average in efficiency of rice production. Efficiency is defined as the total rough rice in kilograms produced after deducting for harvest costs (which are paid in terms of rough rice) divided by the total area the farmer cultivated in rice. There is data from one thousand and twenty-six Indian farms who average 1.07 acres of rice under cultivation. Predictor variables include the village where the farm is located, the total area cultivated with rice, whether traditional or high-yielding varieties of rice are planted, fertilizer use levels, labor hours expended, and labor pay rate. There is data from 1026 farms.

Italian Household Income The target variable in this dataset is the classification of an Italian household's net disposable income as above or below the median. Predictors are such variables as husband and wife's hours of work, number of children between certain ages, work experience, education, and whether or not they resided in northern Italy. This data is from a forthcoming paper in the journal of Applied Econometrics by Aaberge, Colombino, and Steiner (in press).

Own Home This data is derived from the 1987 wave of the Michigan Panel Study of Income Dynamics as used in the study of Lee (1995). Variables such as husband/wife educational and vocational variables as well as number and age of children are related to whether or not the home they live in is owned by the household or otherwise. The number of observations is 3382.

Pima Indians Diabetes This UCI dataset provides 8 medical attributes for 768 women of Pima Indian heritage. Predictors include such attributes as 2-hour serun insulin level, body mass index, diabetes pedigree function, age, skin fold thickness, and diastolic blood pressure. The cases are classified according to whether or not they carry a diagnosis of diabetes.

Working Wives Various demographic variables and type of husband's insurance coverage is related to hours worked per week by the wife (Olson, 1998). To turn this into a classification problem cases are categorized as wife working more or less than 32 hours per week. This dataset also comes from the Journal of Applied Econometrics database. There are over 22,000 cases.

Wage Differences This dataset from the Journal of Applied Econometrics is taken from the second Malysian Family Life Survey done in 1989. Educational, ethnicity, and family asset attributes were related to income (Marcia and Schafgans,1998). To make this a classification problem income level is classified as above or below the average. Because many of the women did not work outside the home only males are included in our study. There are more than 4,000 such cases.

Yeast Proteins Here we predict the cellular localization sites of proteins in yeast cells. There are 8 predictors. We will limit our study to the two most prevalent classes. This gives us 889 instances in this dataset from UCI.

2.5 Experimental Procedure

Eighty percent of each dataset will be used for training the algorithms and twenty percent will be held back as a test set. For the backpropagation, cascade correlation, and Levenberg-Marquardt neural networks ten percent of the training data (8 percent of the total) will be put into a file used to prevent overtraining (Masters, 1993). Assignment of data to training, overtraing prevention, and tresting files will be randomized. Three hidden layers was used with the backpropagation neural network and two with the Levenberg-Marquardt network to insure the ability to model complex relationships. Training of these neural networks stops when the error level on overtraining prevention file passed through the neural net model reaches its minimum and no improvement occurs for 10,000 iterations for backpropagation networks. Tuning discriminant analysis using the stepwise technique to remove non-contributory variables was not done because this might have given an advantage over the other methods. Performance on the test sets using percentage accurately classified and ROC analysis forms the basis for comparing the algorithms.

Chapter 3. Pre-clustering

Another approach to classification is cluster analysis. Cluster analysis is an exploratory data analysis tool where there are no pre-set classes, although the number of classes may be set. Because in cluster analysis classes must be constructed without guidance it is known as an unsupervised learning technique. This is akin to how people or animals learn about their environment when they are not told or directed what to learn.

Clusters are formed when attributes of observations tend to vary together. Cluster analysis constructs "good" clusters when the members of a cluster have a high degree of similarity to each other (internal homogeneity) and are not like members of other clusters (external homogeneity). However, there is no agreement over how many clusters a dataset should be partitioned into. There are no guidelines on the number of clusters that would be optimal to aid supervised learning efforts.

Statisticians have developed clustering procedures which group observations by taking into account various metrics to optimize similarity. The major type of cluster analysis, which will be used in this study, is hierarchical clustering.

Hierarchical clustering begins with putting each observation into a separate cluster. Clusters are then combined successively based upon their resemblance to other clusters. The number of clusters is reduced until only one cluster remains. A tree or dendrogram can represent hierarchical clustering. Each fork in the tree represents a step

in the clustering process. The tree can be sectioned at any level to yield a partition of the set of observations. At its early stages the dendrogram is very broad. There are many clusters that contain very similar observations. As the tree structure narrows the clusters comprise coarser, more inclusive groupings.

Berry and Linoff (1997) have proposed cluster analysis as a precursor to analyzing data with supervised learning techniques. Especially in a large dataset elements may form subgroups or clusters. Members of a cluster may have much in common with other members of their cluster and differ in important ways from members of other clusters. Each cluster may have its own "rules" that relate the attributes of its members to classifying variables. Thus, to enhance accuracy it may be advantageous to first group elements of a dataset by cluster and then apply the classification algorithms successively to each cluster. In this way they will learn each cluster's unique "rules" for relating attributes to classes and thereby more accurately classify the members of each cluster. Berry and Linoff (1997) write about this approach as follows:

> It is possible to find rules and patterns within strong clusters once the noise from the rest of the database has been eliminated… Automatic cluster detection can be used to uncover hidden structure that can be used to improve the performance of more directed techniques… Once automatic cluster detection has discovered regions of the data that contain similar records, other data mining tools have a better chance of discovering rules and patterns within them. (pp.212,214,215)

The process of grouping data into subgroup classifications has been described as "pre-clustering."

Dr. David Wishart, creator of the Clustan cluster analysis software, responded in the following way to the question of his opinion of this use for cluster analysis:

> The essence of clustering is to break down a heterogeneous dataset into homogeneous subsets. …cluster your data into homogeneous subsets which you can describe, and then work individually on the subsets.
>
> In the context of, say, supermarket shoppers, there are different types - the bargain hunter, the quality foods seeker, the organics cook, the anti - GM (genetically-modified) lobbyist, and so on. Each of these types needs a different marketing strategy to achieve good sales response. So they have to be identified and analyzed separately.
>
> In my banking study the same thing happened. The bank was surprised to find it had different types of account holders, some of which were not profitable. They were then able to focus on the profitable ones, and either disengage or convert the non-profitable ones. In essence, developing different sets of rules for cluster subgroups.
>
> I think this works for almost any types of dataset. …In data mining contexts, it probably works best with large datasets, because there's always the hope that you might get a surprise hidden in a lot of data (e.g. the profitability of

> account types) or discover a nugget of hidden data (e.g. in the context of health insurance claims, a tiny group of fraudsters operating a scam) (Wishart, 1999).

A further example Wishart mentions comes from the field of astronomy. In the Hertzsprung-Russel diagram stars are plotted by temperature and luminosity. "Dwarf" and "giant" stars are in separate clusters. Within each cluster there is a different relationship between temperature and luminosity. The correlation is negative for the dwarfs and positive for the giants. If just one correlation were figured for the dataset of all stars the corrleations within the two clusters would wash each other out. This would erroneously indicate no relationship between temperature and luminosity. Yet within the clusters for the two types of stars there are clear "rules" governing the relationship between temperature and luminosity.

Despite the plausibility of this use for cluster analysis there does not appear to be any empirical studies supporting this approach in the data mining or statistical literature. Dr. Jon Kettenring, of Bellcore, is a Fellow and past president (1997) of the American Statistical Association. He gave a presentation entitled "Massive data sets, data mining, and cluster analysis" before the Institute for Mathematics and Its Applications. He was asked if he was aware of any empirical studies which demonstrate that cluster analysis improves the performance of supervised learning done within the clusters. His reply was:

> No, I am not aware of any such studies. There may be some, but in fact these are points of view that are much easier to state than substantiate (Kettenring, 1999).

3.1 Experimental Procedure

Hierarchical clustering with Ward's method as a linkage rule are applied to the training sets derived from several of our datasets using the ClustanGraphics software program (Wishart,1999c). Neural networks are believed to need relatively large training sets (Masters, 1993). Since the training sets are partitioned by cluster analysis we restricted the clustering to datasets containing more than 1,000 cases. The datasets used in this analysis include the Doctor Visits, Italian Household Income, Earnings, Own Home, Wage Differences, and Working Wives. We partition a training set into four clusters based only on the values of the predictive attributes. The statistical and AI classifiers then create a predictive model for the cases that were put into each cluster. Test set data is then be put into its cluster-of-best-fit. The predictive models created for each cluster classify the test set cases assigned to that cluster. For each dataset the accuracy of the models created for the clusters are compared with those created for the entire training set. While the number of clusters created is arbitrarily set at four, this should give us at least some hints as to whether breaking training sets into clusters routinely aids the supervised learning process. The design also allows us to evaluate how clustering and classifying algorithms interact in their effect on accuracy.

The study also looks to see if the error rate for a model applied to a cluster within the training set from which it is derived predicts the error rate for members of the same

cluster within the test set. This could yield confidence levels for predictive classifications of new data. First, the training set itself is passed through the classification model developed from the training set. Almost all the algorithms used in our study work by constructing an abstract description for mapping vector inputs onto classes. Even some members of the training set will not be classified correctly by this concept description. The testing set is also classified. Next, the training and testing set are clustered at the four-cluster level based on a cluster model constructed from the training set. The error rate is figured for each member of both the training and the testing set grouped by cluster membership. The error rate for a given cluster in the training set is compared with the error rate for the same cluster in the test set. If there is a positive relationship, the error level for clusters in the training set could be taken as indicating a confidence level for predictions of that cluster among new cases presented for classification.

Chapter 4. Results

The error rates for the algorithms in each dataset are presented in Tables 2 through 13. ROC data (when applicable) is also included. ROC curves for each dataset are presented in Figures 1 through 12 in Appendix A. The mean error of the classifiers across datasets in ascending order is presented in Table 14. The mean rank of the error rate of the classifiers is shown in Table 15. The mean rank of the classifiers by ROC area under the curve measurement (A(z)) is shown in Table 16. To compute ranks an algorithm was given a score of "1" if it had the lowest error rate, "2" if it had the second lowest error, etc. If two algorithms had an equal error rate, the average rank was assigned. From inspection of the tables and figures we can draw several conclusions:

1. QNET and Model Quest are consistently good classifiers.
2. The four best classifiers are all neural networks.
3. The worst algorithm (LVQ) is also a neural network.
4. The first four classifiers are the same whether ranked according to error rate or ROC area under the curve criteria.
5. For certain datasets there are almost no differences between the quality of classification performance by the various algorithms. In other datasets there are wide differences between the quality of the classifier's decision performances.

The statistical significance of the differences of the mean ranks of algorithms for error rates within datasets was analyzed using Friedman's Test. This test gave a significance probability of <.001 (Chi Square = 65.745, df = 12). This indicates that the null hypothesis that the algorithms have equal error rates on average is rejected.

The author contacted the developers of the most accurate method (QNET) and asked them if they would comment on the reasons for its outstanding performance. QNET implements a standard backpropagation multi-layer perceptron neural network. William Riba, a developer of QNET wrote:

There are a couple of things we paid close attention to in our development of QNET. We spent a lot of time on accuracy optimization. There are computational shortcuts - which we tested and were tempted to take for speed improvements, but were ultimately rejected because they compromised accuracy. With our attention to accuracy you'd think we'd have developed a real slow trainer. Luckily we gained back speed through loop optimizations and the use of an optimizing Intel compiler for the computational sections (buggy as heck - but worth it). It claims to make better use of the CPU's floating point unit - resulting in increased accuracy and speed. Last, we paid close attention to QNET's default settings - again for the purpose of developing more accurate models, not for fastest training. So I would credit an attention to detail more than algorithmic differences (Riba, 2000).

Table 1. Breast Cancer Survival (2 predictors, 306 cases). The Loh and Shih freeware version of QUEST did not run on this dataset. The version of QUEST from SPSS, Inc. was substituted.

			ROC			
Methods	Error rate	Rank	a	b	Az	Rank
Cascade Correlation	24.59%	2.5	.67	.63	.71	3
C&RT	29.51	7				
CHAID	29.51	7				
Discriminant Analysis	29.51	7	.81	.90	.73	2
Levenberg-Marquardt	32.79	10.5	.51	.94	.64	7
Logistic Regression	32.79	10.5	.83	.91	.73	1
LVQ	34.43	12				
Model Quest	24.59	2.5	.53	.48	.68	6
Model Ware	31.15	9	.28	.57	.59	9
QNET	26.23	4	.64	.55	.71	4
QUEST	37.71	13				
Resilient Propagation	22.95	1	.56	.47	.69	5
Ward Classifier	27.87	5	.41	.56	.64	8

Table 2. Cleveland Clinic Heart Disease (11 predictors, 287 cases)

Methods	Error rate	Rank	ROC a	b	Az	Rank
Cascade Correlation	28.07%	10	.96	.80	.77	9
C&RT	31.58	11				
CHAID	28.07	8				
Discriminant Analysis	19.30	2	1.80	1.30	.86	4
Levenberg-Marquardt	28.07	8	1.06	.52	.82	6
Logistic Regression	21.06	3.5	1.80	1.21	.87	2
LVQ	56.14	13				
Model Quest	26.32	6	1.34	.99	.83	5
Model Ware	24.43	5	1.59	1.03	.87	3
QNET	15.79	1	1.82	1.18	.88	1
QUEST	21.05	3.5				
Resilient Propagation	33.33	12	1.33	1.18	.81	7
Ward Classifier	28.07	8	1.20	1.02	.80	8

Table 3. Contraceptive Method Choice (9 predictors, 1473 cases)

Methods	Error rate	Rank	ROC a	b	Az	Rank
Cascade Correlation	30.51	4	1.14	1.14	.75	4
C&RT	28.47	1				
CHAID	30.85	5				
Discriminant Analysis	37.97	9	1.35	1.14	.81	3
Levenberg-Marquardt	34.24	7	.95	1.24	.72	6.5
Logistic Regression	34.76	8	.88	1.24	.71	8
LVQ	41.02	12				
Model Quest	28.81	2	3.15	2.26	.90	2
Model Ware	40.34	10	.79	1.04	.71	9
QNET	29.83	3	4.39	2.72	.94	1
QUEST	65.08	13				
Resilient Propagation	32.46	6	.92	1.07	.73	5
Ward Classifier	41.02	11	.95	1.24	.72	6.5

Table 4. Doctor Visits (11 predictors, 5190 cases)

Methods	Error rate	Rank	ROC a	b	Az	Rank
Cascade Correlation	18.69%	7	.90	.91	.75	8
C&RT	18.88	8.5				
CHAID	19.29	10				
Discriminant Analysis	23.80	11	.97	.98	.79	2
Levenberg-Marquardt	18.50	5	.92	.93	.75	7
Logistic Regression	18.50	5	.96	.96	76	6
LVQ	39.13	13				
Model Quest	18.40	3	1.02	.98	.78	3
Model Ware	23.99	12	.57	.91	.66	9
QNET	18.30	2	.99	.94	.76	4
QUEST	18.88	8.5				
Resilient Propagation	18.50	5	1.01	.95	.85	1
Ward Classifier	17.92	1	.96	.91	.76	5

Table 5. Earnings (3 predictors, 13,408 cases)

Methods	Error rate	Rank	ROC a	b	Az	Rank
Cascade Correlation	35.27%	10.5	.77	.99	.71	6
C&RT	34.34	7				
CHAID	35.05	8				
Discriminant Analysis	35.15	9	.75	1.00	.70	8
Levenberg-Marquardt	34.30	6	.81	.97	.72	2.5
Logistic Regression	35.27	10.5	.74	.99	.70	7
LVQ	41.95	13				
Model Quest	33.82	1	.83	.98	.72	1
Model Ware	40.23	12	.48	1.03	.63	9
QNET	34.12	4	.80	.98	.71	4
QUEST	33.89	2				
Resilient Propagation	34.30	5	.81	.97	.72	2.5
Ward Classifier	34.04	3	.78	.93	.72	5

Table 6. Italian Household Income (9 predictors, 2,953 cases)

Methods	Error rate	Rank	ROC a	b	Az	Rank
Cascade Correlation	20.64%	5	1.51	.87	.87	8
C&RT	21.66	9.5				
CHAID	23.18	11				
Discriminant Analysis	20.81	6.5	1.64	.94	.88	6
Levenberg-Marquardt	20.98	8	1.61	.89	.89	5
Logistic Regression	20.81	6.5	1.65	.96	.88	7
LVQ	45.82					
Model Quest	18.78	1	1.62	.87	.89	4
Model Ware	24.53	12	1.09	1.07	.77	9
QNET	20.47	4	1.69	.92	.89	1
QUEST	21.66	9.5				
Resilient Propagation	20.30	2.5	1.76	1.03	.89	2
Ward Classifier	20.30	2.5	1.70	.96	.89	3

Table 7. Wage Differences (9 predictors, 8,748 cases)

Methods	Error rate	Rank	ROC a	b	Az	Rank
Cascade Correlation	36.46%	10	.56	1.05	.65	7
C&RT	30.11	4				
CHAID	31.89	6				
Discriminant Analysis	38.11	12	.46	.98	.63	9
Levenberg-Marquardt	30.74	5	.99	1.08	.75	2
Logistic Regression	37.31	11	.47	.99	.63	8
LVQ	55.14	13				
Model Quest	29.83	3	.99	1.07	.75	1
Model Ware	33.89	9	.71	1.01	.69	6
QNET	29.60	2	.96	1.05	.75	3
QUEST	32.91	8				
Resilient Propagation	29.09	1	.92	1.02	.74	4
Ward Classifier	32.00	7	.84	1.02	.72	5

Table 8. Own Home (12 predictors, 3,382 cases)

Methods	Error rate	Rank	ROC a	b	Az	Rank
Cascade Correlation	33.58%	10	.87	1.09	.72	8
C&RT	30.03	6				
CHAID	30.92	8				
Discriminant Analysis	36.83	12	.94	1.03	.74	6
Levenberg-Marquardt	30.47	7	1.09	1.08	.77	4
Logistic Regression	28.99	3	1.11	1.02	.78	2
LVQ	37.72	13				
Model Quest	28.85	2	1,06	.97	.78	3
Model Ware	36.54	11	.64	.94	.68	9
QNET	27.07	1	1.13	.96	.79	1
QUEST	29.29	4				
Resilient Propagation	31.51	9	.84	1.03	.72	7
Ward Classifier	30.03	5	1.06	1.04	.77	5

Table 9. Pima Indians Diabetes (6 predictors, 724 cases)

Methods	Error rate	Rank	ROC a	b	Az	Rank
Cascade Correlation	19.31%	1.5	1.86	1.42	.86	3.5
C&RT	22.07	8.5				
CHAID	28.28	12				
Discriminant Analysis	19.31	1.5	1.86	1.42	.86	3.5
Levenberg-Marquardt	23.45	11	1.04	.89	.78	9
Logistic Regression	21.38	7	1.68	1.17	.86	2
LVQ	33.11	13				
Model Quest	20.69	5.5	1.32	.87	.84	6
Model Ware	22.07	8.5	1.57	1.28	.83	8
QNET	20.00	3.5	1.51	1.05	.85	5
QUEST	22.76	10				
Resilient Propagation	20.69	5.5	1.69	1.39	.84	7
Ward Classifier	20.00	3.5	1.73	1.18	.87	1

Table 10. Indian Rice Farms (22 predictors, 1,026 cases)

Methods	Error rate	Rank	ROC a	b	Az	Rank
Cascade Correlation	37.56%	11	.10	1.39	.52	6
C&RT	35.13	9				
CHAID	37.08	10				
Discriminant Analysis	31.71	4	.04	1.24	.51	7
Levenberg-Marquardt	38.05	12	-.03	1.14	.49	9
Logistic Regression	32.20	5.5	.04	1.24	.51	8
LVQ	47.80	13				
Model Quest	32.68	7.5	.66	.78	.70	4
Model Ware	30.24	3	.11	1.12	.53	5
QNET	29.27	2	.70	.56	.73	2
QUEST	32.68	7.5				
Resilient Propagation	26.34	1	.70	.77	.71	3
Ward Classifier	32.20	5.5	.84	.82	.74	1

Table 11. Working Wives (20 predictors, 22,272 cases). The Loh and Shih freeware version of QUEST did not run on this dataset. The version of QUEST from SPSS, Inc. was substituted.

Methods	Error rate	Rank	ROC a	b	Az	Rank
Cascade Correlation	24.48%	7.5	1.30	.94	.83	8
C&RT	24.81	9				
CHAID	24.30	5				
Discriminant Analysis	25.20	11	1.31	.93	.83	6
Levenberg-Marquardt	23.82	2.5	1.36	.97	.84	4
Logistic Regression	25.20	10	1.30	.92	.83	7
LVQ	54.92	13				
Model Quest	16.35	1	1.36	.94	.84	1
Model Ware	31.21	12	.83	1.03	.72	9
QNET	24.14	4	1.35	.93	.84	2
QUEST	24.48	7.5				
Resilient Propagation	23.82	2.5	1.40	1.05	.83	5
Ward Classifier	24.36	6	1.39	1.01	.84	3

Table 12. Yeast Proteins (7 predictors, 892 cases)

Methods	Error rate	Rank	ROC a	b	Az	Rank
Cascade Correlation	34.83%	2	.71	.98	.69	2
C&RT	34.83	2				
CHAID	37.64	9				
Discriminant Analysis	37.08	6.5	.68	1.01	.68	3
Levenberg-Marquardt	38.77	11	.48	.97	.64	9
Logistic Regression	37.36	8	.68	1.03	.68	4
LVQ	44.38	13				
Model Quest	37.08	6.5	.55	.85	.66	7
Model Ware	40.45	12	.59	1.11	.65	8
QNET	34.83	2	.78	1.07	.70	1
QUEST	35.96	4.5				
Resilient Propagation	38.20	10	.60	.93	.67	6
Ward Classifier	35.96	4.5	.61	.85	.68	5

Table 13. Mean Error of Methods Across Datasets in Ascending Order

Methods	Mean Error rate
QNET	25.80
Model Quest	26.35
Resilient Propagation	27.72
Ward Systems Classifier	28.05
C&RT	28.45
Cascade Correlation	28.61
Logistic Regression	28.76
QUEST	28.85
Levenberg-Marquardt	29.52
Discriminant Analysis	29.57
CHAID	29.60
ModelWare	31.35
LVQ	44.24

Table 14. Mean Rank of Error Rate in Ascending Order

Methods	Mean Rank of Error Rate
QNET	2.71
Model Quest	3.42
Resilient Propagation	5.04
Ward Systems Classifier	5.17
Cascade Correlation	6.75
C&RT	6.88
Logistic Regression	7.38
QUEST	7.58
Discriminant Analysis	7.63
Levenberg-Marquardt	7.75
CHAID	8.25
Model Ware	9.63
LVQ	12.83

Table 15. Mean Rank of ROC A(z) in Ascending Order

Method	Mean Rank on A(z)
QNET	2.42
ModelQuest	3.58
Resilient Propagation	4.54
Ward Systems Classifier	4.62
Discriminant Analysis	4.96
Logistic Regression	5.17
Levenberg-Marquardt	5.92
Cascade Correlation	6.04
ModelWare	7.75

Each training dataset was partitioned by cluster analysis into four clusters (pre-clustering) and models were created for each cluster using three classification algorithms - discriminant analysis, QNET, and C&RT. These models were used to classify test cases that belonged to the same cluster. The errors made within each cluster of the test set were summed and compared with the error level when the classification model was developed from the entire dataset. The results are presented in Table 16. The mean error

level is slightly greater using pre-clustering, obviating the possibility that this method has an overall positive effect upon classification accuracy in the datasets used.

Table 16. Comparison of Error Rates for Three Algorithms Run on Entire Training/Testing Datasets and on Training/Testing Datasets Partitioned Into Four Clusters Using Ward's Method

Dataset	Methods	Error Rate Without Pre-Clustering	Error Rate With Pre-Clustering
Doctor Visits			
	Discriminant Analysis	23.80%	23.99%
	QNET	18.30	19.27
	C&RT	18.88	18.12
Earnings			
	Discriminant Analysis	35.16%	39.19%
	QNET	34.12	33.63
	C&RT	34.34	34.08
Italian Household Income			
	Discriminant Analysis	20.81%	22.17%
	QNET	20.47	20.47
	C&RT	21.66	23.86
Wage Differences			
	Discriminant Analysis	38.11%	39.20%
	QNET	29.60	29.49
	C&RT	30.11	31.89
Own Home			
	Discriminant Analysis	36.83%	33.73%
	QNET	27.07	29.73
	C&RT	30.03	34.07

Working Wives		
Discriminant Analysis	25.20%	25.21
QNET	24.14	24.00
C&RT	24.81	21.85
Mean	27.41%	27.99%

Error rate within a cluster from the training set was correlated with error rate within those same clusters in the test set. Using both the decision tree algorithm C&RT and the neural network QNET this relationship was found to be extremely robust (Tables 17 and 18). Relative error rates computed by cluster in the training set can thus give us some indication of the confidence that can be placed upon predictions of new or test cases belonging to the corresponding clusters.

Table 17. Error Rate by Clusters Within Training and Testing Set Using C&RT Algorithm

	Training Set	Testing Set
Doctor Visits		
Cluster 1	13.01%	15.70%
Cluster 2	10.34	10.42
Cluster 3	27.48	28.00
Cluster 4	26.81	26.06
Earnings		
Cluster 1	26.80%	29.16%
Cluster 2	36.05	37.15
Cluster 3	34.81	36.11
Cluster 4	37.29	33.95

Italian Household Income

Cluster 1	21.77%	27.95%
Cluster 2	32.26	23.33
Cluster 3	13.73	17.10
Cluster 4	14.37	14.91

Wage Differences

Cluster 1	26.79%	26.41%
Cluster 2	27.84	29.45
Cluster 3	33.15	32.25
Cluster 4	36.32	32.27

Own Home

Cluster 1	20.17%	30.68%
Cluster 2	27.70	32.14
Cluster 3	22.97	36.36
Cluster 4	14.27	15.71

Working Wives

Cluster 1	26.44%	28.20%
Cluster 2	22.78	23.54
Cluster 3	25.68	26.57
Cluster 4	18.21	20.33

Correlation between percentages in training and testing set: $r = .853$, $df=23$, $p<<.01$

Table 18. Error Rate by Clusters Within Training and Testing Set Using QNET Algorithm.

	Training Set	Testing Set
Doctor Visits		
Cluster 1	13.44%	15.70%
Cluster 2	11.49	8.33
Cluster 3	31.73	26.00
Cluster 4	27.56	24.85

Earnings

Cluster 1	27.28%	28.77%
Cluster 2	36.42	37.15
Cluster 3	35.57	36.98
Cluster 4	38.04	30.24

Italian Household Income

Cluster 1	21.66%	25.59%
Cluster 2	26.88	30.00
Cluster 3	13.73	16.06
Cluster 4	15.50	14.04

Wage Differences

Cluster 1	26.20%	26.67%
Cluster 2	27.55	27.34
Cluster 3	33.26	32.02
Cluster 4	36.50	32.76

Own Home

Cluster 1	20.24%	27.27%
Cluster 2	27.88	29.29
Cluster 3	14.87	27.28
Cluster 4	14.39	14.29

Working Wives

Cluster 1	25.46%	26.44%
Cluster 2	21.13	21.64
Cluster 3	25.32	27.09
Cluster 4	18.89	19.01

Correlation between percentages in training and testing set: $r = .870$, $df=23$, $p<<.01$

Chapter 5. Discussion

1. Algorithm superiority is somewhere between selective and generalized.

One of the findings of this study is that the algorithms used do systematically differ in their general accuracy. Brodley (1993) has asserted that any superiority of a learning algorithm is only "selective" and limited to a given task or dataset:

> The results of empirical comparisons of existing learning algorithms illustrate that each algorithm has a *selective superiority;* (author's italics) it is best for some but not all tasks. Given a data set, it is often not clear beforehand which algorithm will yield the best performance…In every case, the algorithm can boast one or more superior learning performances over others, but none is always better. (Brodley, 1993)

This may not be the most accurate description of the comparative performance of the classification algorithms used in our study. The superiority of the best algorithms in our study is not as selective as in Brodley's conception. A statistical test did reject the notion that there was no difference in classifier performance across all datasets. On the other hand, any superiority is not totally general either. Superiority somewhere between general and selective is perhaps the best characterization of our results. For example, the best overall performer, QNET, is the absolute best for only two datasets and ties with two other methods for first place in another, but it never ranks below fourth of the thirteen classifiers. And among the weaker performers, the CHAID decision tree has a mean rank of 8.25. It is never better than fifth or worse than twelfth. It seems that, at least among our datasets, it would be difficult to claim any "selective superiority" for it. These findings are presented in the table below.

Table 19. Algorithm Superiority Across Datasets for Two Algorithms

	Times Best	Average Rank	Range
QNET	3(one tie)	2.71	1 - 4
CHAID	0	8.25	5 - 12

2. Newer methods for classification are coming into their own.

The good performance of several newer algorithms suggests that they have earned their place in data classification endeavors. There have been concerns that expectations for neural networks in particular, following a long historical pattern in the artificial intelligence field, have been inflated. But, as Banks (1996) notes:

> The ultimate arbiter among these many competing methods must be performance (Banks, 1996).

The present results suggest that neural nets have a contribution to make to classification efforts. The most recent, extensive, and methodologically elegant published comparison

of classification algorithms is the work of Lim, Loh, and Shih (1999). They compared twenty-two decision trees, and nine statistical but only two neural network algorithms. And one of the two neural network algorithms was LVQ. Michie, Spiegelhalter, and Taylor, (1994) Poddig, (1995) and the present study found this early neural network algorithm (circa 1988) to be among the least accurate classifiers (due either to the implementation used or to the algorithm itself). The results of the current study suggest that future work in this field will benefit from including a range of the more modern neural network algorithms. They also suggest that neural networks not be left out for consideration when the concern is classifying real-world data for some practical purpose.

3. The amount of divergence between the classifiers on accuracy measures varies as a function of the dataset.

Examining the error rates and ROC graphs by dataset reveals that for some datasets it would seem not to matter which algorithm you selected to do classification - they almost all work about the same. For other datasets there are very definite "winners" and "losers" among the classification algorithms. For instance, looking at the ROC curves for the "Earnings" dataset (Figure 5) most overlie each other so closely that they cannot be discriminated. Similarly, the range of error rates for 11 of the classifiers for this dataset falls narrowly from 33.82% to 35.27%. For the "Wage Differences" dataset inspection of the ROC curves (Figure 7) shows that algorithm performance varies substantially across the range of possible cutoff points for classification. And the error rate measures similarly show a broad range of results from 29.09% to 37.31%. The reason for this divergence in the amount of variability of algorithm performance between datasets is uncertain.

4. Accuracy optimization techniques should be a priority in the computer programming of classification algorithms.

Comments from one of the developers of our best method, QNET, were cited above. They indicate that with dependence upon computers to implement algorithms close attention needs to be paid to programming techniques aimed at accuracy optimization. Technical choices about programming issues will greatly affect the accuracy of iterative, computationally intensive classification algorithms. To achieve highly accurate classifier performance it is necessary to consider details of the algorithm's implementation on a computer system.

5. Cluster analysis can be explored as a method to indicate confidence levels for classifiers.

The attempt to increase classification accuracy by first clustering the training and testing data, and then developing and testing the classification model within the clusters failed. It was no more accurate than just developing one model by training the algorithm on all the data. Possibly clustering methods other than Ward's method could be tried. And it may be that this approach will work on datasets other than those included in the present study.

The error rates within clusters in training sets are highly predictive of error rates for those clusters in testing sets. The relative rankings of the accuracy of the clusters

within the training data can be used to indicate a confidence level for predictions within those clusters from new data or a testing set. Thus, if cases are classified at the four cluster level predictions on new or test set cases could be ranked from 1 (most confidence) to 4 (least confidence). This would be based upon their membership in clusters that in the training set the classification model had greater or lesser success classifying correctly. This is a new use for cluster analysis that can be explored further.

Chapter 6. Improvements and Future Directions

As with any large project at its completion, the author can see ways in which it could have been improved as well as directions he would like to pursue in the future. Here are some of those ideas.

1. Several interesting new classification algorithms became available as this project neared its conclusion. Tjen-Sien Lim developed an advanced decision tree methodology, known as PLUS (Lim, 1999). Nauck (1999) presented her implementation of NEFCLASS - a combined neural network-fuzzy logic approach to classification. Another new type of classifier is support vector machines (SVM). SVM's combine linear modeling and instance-based learning. Software has been offered which implements this new technique. (Witten and Frank, 2000).
2. It is known that with smaller datasets single train and test partitions may provide an inaccurate estimation of the true error rate of a classification algorithm (Weiss and Kulikowski, 1991). A random sub-sampling procedure known as 10-fold cross validation has been developed to minimize any estimation bias. Typically, as in Lim, (1999) this is implemented by randomly dividing a dataset into ten disjoint subsets, each containing the same number of records. A classification model is constructed from nine of the subsets and tested on the one withheld subset. This process is repeated ten times, each with a different subset withheld. Accuracy across the ten subsets is averaged to provide an estimate for the classifier. This procedure would have been interesting to employ with some of the smaller datasets used in the present study.
3. New ways have been discovered to make classification algorithms more accurate. For example, it has been found that if a classifier is even weakly accurate more accurate results can be obtained by running the algorithm several times on different samples of the training set and combining the resulting models. This is known as "bagging." A procedure called "boosting" is another way of combining several models into a single predictive model (Schapire, Freund, Bartlett and Lee, 1998). Such unique ways of employing the training and testing data, along with improving classification algorithms, offer enhanced opportunities to solve the complex data classification problems our technological world presents to us.

In conclusion, we have seen that to understand and enhance the data mining process we have relied upon tools traditionally belonging to both statistics and computer science. As statistician William Shannon (1999) wrote:

I think there is a challenge for statisticians to start learning machine learning and computer science, and machine learners to start learning statistics. These two fields rightly fall under the broad umbrella of "data analysis."

References

Aaberge R., Colombino, U. and Steiner, S. (in press) "Labor supply in Italy. An empirical analysis of joint household decisions, with taxes and quantity constraints" Journal of Applied Econometrics

AbTech Corporation (1996) Model Quest: Users Manual Charlottesville, Virginia.

Banks, David (1996) "Working without a net" Classification Society of North America Newsletter July, #45, pp. 2-11.

Berry, Michael and Linoff, Gordon (1997) Data Mining Techniques New York: John Wiley and Sons.

Blake, C., Keogh, E., and Merz, C.J. (1998) UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

Bradley, A. (1997) "The use of the area under the ROC curve in the evaluation of machine learning algorithms" Pattern Recognition 30, 1145-1159.

Brodley, C. (1993) "Addressing the selective superiority problem: Automatic algorithm/model class selection" Proceedings of the Tenth International Machine Learning Conference Amherst, MA. pp. 17-24.

Brown, D., Corruble, V. and Pittard, L. (1993) "A comparison of decision tree classifiers with backpropagation neural networks for multimodal classification problems" Pattern Recognition 26, 953-961.

Chanrasekaran, H. (n.d.) "Training of MLP using Levenberg-Marquardt Algorithm" Image Processing and Neural Networks Laboratory, University of Texas at Arlington. Available at: http://nexus.uta.edu/eeweb/ip/software./Lm.mlp.txt

Chen, C.H. (1991) "On the relationship between statistical pattern recognition and artificial neural networks" In Neural Networks In Pattern Recognition and Their Applications, New York, World Scientific.

Cios, K. Pedrycz, W. & Swiniarski, R. (1998) Data Mining Methods for Knowledge Discovery. Kluwer: Boston.

Couvreur, K. and Couvreur, P. (1997) "Neural networks and statistics: a naïve comparison" Belgian Journal of Operations Research, Statistics, and Computer Science 36, 217-225.

Curram, S.P., and Mingers, J. (1994) "Neural networks, decision tree induction and discriminant analysis: an empirical comparison" Journal of the Operational Research Society 45, 440-450.

Deb, P. and Trivedi, P.K. (1997) "Demand for medical care by the elderly: A finite mixture approach" Journal of Applied Econometrics 12, 313-336.

Dietterich, T., Hild, H., and Bakiri, G. (1995) "A comparison of ID3 and Backpropagation for English text-to-speech mapping" Machine Learning 18, 51-80.

Dietterrich, T. (1998) "Approximate statistical tests for comparing supervised classification learning algorithms" Neural Computation 10, 1895-1924.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996) "Knowledge discovery and data mining: Towards a unifying framework" Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon.

Frawley, W.J. Pietetsky-Shapiro, G. & Zimmerman, H.G. (1992) "Knowledge discovery in Databases: An overview" AI Magazine 13, 57-70.

Hess, Paul (1992) "Model Ware: Applications of the Universal Process Modeling (UPM) Algorithm" Technical Report prepared for Teranet Incorporated (currently Triant Technologies, Inc. Vancouver, BC, Canada).

Holmstrom, L., Koistinen, P., Laaksonen, J., Oja, E. (1997) "Neural and statistical classifiers - taxonomy and two case studies" IEEE Transactions on neural Networks 8, 5-17.

Horrace, W. and Schmidt, P. (in press) "Multiple comparisons with the best, with economic applications" Journal of Applied Econometrics

Jensen, C. (1999) QwikNet v.2.23 Kirtland, Washington.

John, George (1997) Enhancements to the Data Mining Process Doctoral dissertation, Stanford University.

Kettenring, Jon. Personal communication, February, 8, 1999.

Lee, M. (1995) "Semiparametric estimation of simultaneous equations with limited dependent variables: A case study of female labor supply" Journal of Applied Econometrics 10, 187-200.

Lim, Tjen-Sien (1999) User's guide for PLUS Version 1.0

Lim, T.-S. and Loh, W.-Y. and Shih, Y.-S. (in press) "A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms" Machine Learning.

Logical Designs Consulting, Inc.(1994) THINKS Neural Networks for Windows La Jolla, CA.

Loh, W.Y. and Shih, Y.S. "Split selection methods for classification trees" (1997) Statistica Sinica 7, 815-840.

Marcia, M.A. and Schafgans (1998) "Ethnic wage differences in Malaysia: Parametric and semiparametric estimation of the Chinese-Malay wage gap" Journal of Applied Econometrics 13, 481-504.

Masters, T. (1993) Practical Neural Network Recipes in C++ Boston: Academic Press.

Metz, C. (1978) "Basic principles of ROC analysis" -Seminars in Nuclear Medicine 8, 283-298.

Metz, C. (1998) ROCKIT 0.9B User's Guide Department of Radiology, University of Chicago, Available at http://www-radiology.uchicago.edu/krl/toppage11.htm

Michie, D. Spiegelhalter, D.J. Taylor, C.C.(eds.) (1994) Machine Learning, Neural and Statistical Classification Chichester, Horwood.

Nauck, U. (1999) Design and Implementation of a Neuro-Fuzzy Data Analysis Tool in Java. Technical University of Braunschweig. Software available at http://fuzzy.cs.uni-magdeburg.de/nefclass/nefclass-j/_dld/

NeuralWare, Inc. (1991) Neural Computing Pittsburgh, Privately published manual.

Olson, C. (1998) "A comparison of parametric and semiparametric estimates of the effect of spousal health insurance on weekly hours worked by wives" Journal of Applied Econometrics 13, 543-565.

O'Sullivan, P. J., Martinez, J., Durham, J., and Felker, S. "Using UPM for real-time multivariate modeling of semiconductor manufacturing equipment" Paper

presented at the SEMATECH APC/AEC Workshop VII, November 5-8, 1995, New Orleans, Louisiana.

Pesonen, E. (1997) "Is neural network better than statistical methods in diagnosis of acute appendicitis?" In: Medical Informatics Europe '97 Pappas, C., Maglaveras, N., and Scherrer, J.R. (eds.) IOS Press, Amsterdam, Netherlands.

Polachek, Solomon and Yoon, Bong Joon (1996) "Panel estimates of a two-tiered earnings frontier" Journal of Applied Econometrics 11, 169-178.

Poddig, Thorsen (1995) "Bankruptcy prediction: A comparison with discriminant analysis" In Neural Networks in the Capital Markets Refenes, A.P. (ed.) John Wiley and Sons, New York.

Prechelt, L. (1996) A quantitative study of experimental evaluations of neural network algorithms: current research practice" Neural Networks 9,

Provost, F. Fawcett, T. and Kohavi, R. (1998) "The case against accuracy estimation for comparing induction algorithms" Proceedings of the Fifteenth International Conference on Machine Learning, Madison, WI.

Riba, William Personal Communication, June 7, 2000.

Ripley, B.D. (1994) "Neural networks and related methods for classification" Jounal of the Royal Statistical Society, B 56, 409-456.

Salzberg, S. (1997) "On comparing classifiers: pitfalls to avoid and a recommended approach" Data Mining and Knowledge Discovery 1, 317-327.

Sandholm, T., Brodley, C. Vidovic, A., and Sandholm, M. (1996) "Comparison of regression methods, symbolic induction methods and neural networks in morbidity diagnosis and mortality prediction in equine gastrointestinal colic" AAAI Spring symposium series, Artificial intelligence in medicine: Applications of current technologies, pp. 154-159, Stanford University, CA.

Schapire, R., Freund, Y., Bartlett, P. and Lee, W.S. (1998) "Boosting the margin: A new explanation for the effectiveness of voting methods" Annals of Statistics 26, 1651-1686.

Schwartz, M.H., Ward, R.E. MacWilliams, C. and Verner, J.J. (1997) "Using neural networks to identify patients unlikely to achieve a reduction in body pain after total hip replacement surgery" Medical Care 35, 1020-1030.

Sen, T.K., Oliver, R. and Sen, N. (1995) "Predicting Corporate mergers" In Neural Networks and the Capital Markets Refenes,A.P. (ed.) John Wiley and Sons, New York.

Shannon, W. Comments posted to machine learning discussion group maintained by T.S. Lim

Shavlik, J.W., Mooney, R.J. and Towell, G.G. (1991) "Symbolic and neural learning algorithms: An experimental comparison" Machine Learning 6, 111-143.

Swets, J. "The relative operating characteristic in psychology" (1973) Science 182, 990-1000.

Teranet Incorporated (currently Triant Technologies, Inc.) (1992) Model Ware User's Manual Nanaimo, BC, Canada.

Two Crows Corporation (1998) Introduction to Data Mining and Knowledge Discovery, Second Edition Patomac, MD.

Vesta Systems, Inc.(1996) QNET V.2.1 User's Manual Chicago, IL.

Ward Systems Group, Inc. (1998) NeuroShell Classifier Frederick, MD.

Weiss, S. and Kulikowski, C. Computer Systems That Learn: Classification and Prediction Methods From Statistics, Neural Networks, Machine Learning and Expert Systems. Morgan Kaufman Publishers, San Francisco.

Wilson, R. (1997) Advances in Instance-Based Learning Doctoral Dissertation, Brigham Young University.

Wishart, David. (1999a)Personal communication, June 19, 1999.

Wishart, David. (1999b)Personal communication, February, 5, 1999.

Wishart, David (1999c) Clustan Graphics Primer Edinburgh, Scotland.

Witten, I. and Frank, E. (2000) Data Mining Morgan Kaufman Publishers, San Francisco, California. Software available at http://www.cs.waikato.ac.nz/ml/weka/