PERGAMON

Expert Systems with Applications 17 (1999) 223–232

Expert Systems with Applications

www.elsevier.com/locate/eswa

Association, statistical, mathematical and neural approaches for mining breast cancer patterns

P.C. Pendharkar[a,*], J.A. Rodger[b], G.J. Yaverbaum[a], N. Herman[a], M. Benner[c]

[a]*Information Systems, Penn State University at Harrisburg, 777 W. Harrisburg Pike, Middletown, PA 17057, USA*
[b]*Management Information Systems, Department of Business, University of Pittsburgh, Johnstown, PA 15904, USA*
[c]*AMP Incorporated, Harrisburg, PA, USA*

Abstract

Using several association and classification approaches to study breast cancer patterns, this study illustrates how these approaches can be used to predict and diagnose the occurrence of breast cancer. The results of the study, based on data obtained from a large medical facility in western Pennsylvania, show that data mining can be a viable tool for breast cancer diagnosis. © 1999 Elsevier Science Ltd. All rights reserved.

Keywords: Data mining; Data envelopment analysis; Artificial neural networks

1. Introduction

Data mining (DM), sometimes referred to as knowledge discovery, is a systematic approach to finding hidden patterns, trends and relationships in data. Due to its applicability to information management, decision support, fraud detection, marketing strategy, financial forecasting, process control and many other applications, DM has attracted tremendous attention from both researchers and practitioners. Several approaches to data mining exist and these approaches can be broadly classified into two categories: methodologies and technologies. According to Pass (1997), methodologies used to study the effects of data mining consist of cluster analysis, linkage analysis, visualization, and categorization analysis. The technologies consist of connectionist models/neural networks, decision trees, genetic algorithms, fuzzy logic, statistical approaches and time series approaches.

A classification problem, a sub-problem under the broad area of cluster analysis, one in which observations are assigned to one of several disjoint groups. Classification problems play an important role in medical decision making. Binary classification problems, a sub-set of classification problems, are ones in which data are restricted to one of two groups. These problems (also termed two-group discriminant analysis problems) have a wide applicability to problems ranging from credit scoring, default prediction and direct marketing to applications in finance and medical domains. There are a number of statistical, mathematical, and artificial intelligence approaches used to solve a binary classification problem. A few examples of these approaches, the use of which is based on the type of solution algorithm used, are Fisher's linear discriminant analysis, Logit, Probit, ID3, C 4.5 and neural networks.

The solution to a binary classification problem is a model that is expressed in terms of vector of weights w_i $i \in [1,\ldots,n]$ together with scalars c_1 and c_2 such that given a observation defined by a vector x_i $i \in [1,\ldots,n]$ of attribute values is classified into group 1 if the polynomial $\sum_{i=1}^{n} w_i x_i^p \le c_1$ and into group 2 when $\sum_{i=1}^{n} w_i x_i^p > c_2$ (c_1 and c_2 are generally considered equal). The term p is the order of the polynomial. When p is equal to 1, the resulting model is a linear model.

Breast cancer, a common cancer in women, affects one in every seven women (Wingo, Tong & Bolden, 1995). The traditional approach to detecting breast cancer is mammography. Research suggests that radiologists show considerable variability in how they interpret a mammogram. One study shows that 90% of radiologists recognized fewer than 3% of cancers and 10% recognized about 25% of the cases (Elmore, Wells, Carol, Lee, Howard & Feinstein, 1994). A few researchers have used several statistical and artificial intelligence approaches for predicting breast cancer. The results of these studies indicate that AI approaches can be successfully applied to the prediction of breast cancer (Kovalerchuck, Triantaphyllou, Ruiz & Clayton, 1997).

[*] Corresponding author.
E-mail addresses: pxp19@psu.edu (P.C. Pendharkar); gjy1@psu.edu (G.J. Yaverbaum)

0957-4174/99/$ - see front matter © 1999 Elsevier Science Ltd. All rights reserved.
PII: S0957-4174(99)00036-6

The binary classification approaches described above, such as neural networks, discriminant analysis, and C 4.5, use the data to identify the discriminant function/rules that discriminate between the two classes (Cancer/No-Cancer). In effect, these approaches “learn” from the data and thus apply to the concept of machine learning in AI. Among several factors that impact the effective learning of the discriminant function is the ratio of the number of examples belonging to classes 1 and 2. If the mining data set contains several examples from class 1 and very few examples from class 2, there is bias in the discriminant function that the technique identifies and it follows that the bias results in lower reliability of the technique. All techniques need data from both classes before the discriminant function is learned by the system. For this reason, if the mining data set contains examples from only one class, it is nearly impossible and inappropriate to use neural networks, discriminant analysis, or C 4.5.

Recently, Troutt, Rai and Zang (1995) proposed data envelopment analysis (DEA) for binary classification problems. This technique is useful because it can learn the discriminant function when data about only one class is available. For example, if hospital data contains information about patients with breast cancer, then DEA is the only viable technique for learning the discriminant function. Pendharkar and Kumar (1998) show that DEA can provide the best information when data about both classes is available. The research, however, on using DEA for binary classification has so far been theoretical and experimental studies have not yet been conducted to test this technique on real data.

In this paper, we compare the performance of DEA and artificial neural networks (ANN) using discriminant analysis for mining breast cancer patterns. We use association rules to study associations of different female hormones with the occurrence of breast cancer and also bench mark the performance of ANN and DEA against the standard parametric Fisher’s linear discriminant analysis (FLDA) technique. The contributions of our study are two-fold: first we benchmark a new and relatively unexplored approach, DEA for discriminant analysis, against the established parametric (FLDA) and non-parametric technique (ANN); and second we illustrate the utility of data mining to learn breast cancer patterns.

The remainder of the paper is organized as follows: Section 2 summarizes recent literature in data mining; Section 3 provides a brief description of DEA and NN for discriminant analysis; Section 4 provides a description of the data used for this study and the results of our experiments; and Section 5 concludes the paper with a summary of our findings and directions for future research.

2. Review of literature on data mining

Data Mining is a broad area involving the discovery of different patterns within historical databases. In general, data mining involves the identification of patterns that are not, otherwise, easily obtained using traditional descriptive statistical techniques such as mean, median, mode and standard deviation. Much of the research in data mining relates to machine learning of rules and the observation of patterns in very large customer databases (Cheung, Ng, Fu & Fu, 1996) and knowledge acquisition for various knowledge base systems (Bhattacharyya & Pendharkar, 1998). We review the current literature on data mining in both of these areas.

Past research into identifying rules and patterns in very large databases has focused on learning association rules in transactional databases, clustering the abstract objects and finding similarities within different sequences in a database. However, the majority of research in mining and learning association rules (Agrawal, Faloutsos & Swami, 1993; Agrawal and Srikant, 1995; Han & Fu, 1995; Mannila, Toivonen & Verkamo, 1994; Park, Chen & Yu, 1995; Savasere, Omiecinski & Navathe, 1995; Srikant & Agrawal, 1995) has focused on learning association rules of the form

$$P_1 \wedge P_2 \ldots \wedge P_n \Rightarrow Q_1 \wedge Q_2 \ldots \wedge Q_m$$

where P_i, and Q_j for $i \in [1, \ldots, n], j \in [1, \ldots, m]$ are a set of attribute-values from data sets. In general, mining of association rules problem can be represented in the following form:

> Let $I = \{i_1, i_2, \ldots, i_n\}$ represent the set of items in a database D. Further, let T represent a transaction that includes a set of items. We can then write $T \subseteq I$. If X represents a set of items then we can say that T *contains* X if and only if $X \subseteq T$. The association rule is an implication of the form $X \Rightarrow Y$ where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \phi$. The confidence factor, CF, of the association rule in database D, is then obtained by finding the percentage of transactions in database D that contain X and also contain Y. The evidence E for the association rule in database D is the percentage of transactions in database D that contain $X \cup Y$. If η_{minCF} is the minimum confidence threshold and, η_{minE} is the minimum evidence threshold, then the problem of mining association rules is to find all the association rules whose confidence and evidence is greater than the respective thresholds.

Many algorithms have been proposed for mining association rules. Popular algorithms are A PRIORI (Agrawal & Srikant, 1995), DHP (Chen, Han & Yu, 1996b), PARTITION (Savasere et al., 1995) and DMA (Cheung et al., 1996), with the A PRIORI algorithm being one of the most popular for mining association rules in a centralized database. In an A PRIORI algorithm, the entire database is scanned once and large item sets are identified. The item sets are then arranged in ascending order of size. Let C_1 be the set of one item large item sets L_1 generated by the first scan. A set C_2 of two item sets is created using $L_1 * L_1$ where

$*$ is an operation for concatenation given by:

$$L_k * L_k = \{X \cup Y | X, Y \in L_k, |X \cap Y| = k - 1\}$$

where k is the scan iteration.

$$C_2 = \binom{|L_1|}{2}$$

Agrawal and Srikant (1995) describe the A PRIORI algorithm with an easy to follow example. DHP and PARTITION are extensions to the A PRIORI algorithm that make it computationally efficient. In the DHP algorithm, hash tables are used to prune candidates during various scan iterations (Cheung et al., 1996). A PARTITION algorithm divides a database into partitions such that each can be processed effectively (Savasere et al., 1995). While A PRIORI, DHP and PARTITION algorithms were designed for centralized databases, real life transactional data is often contained in distributed databases. To solve this problem, Cheung et al. (1996) proposed the distributed mining of association rules (DMA) algorithm for mining association rules in a distributed database environment. DMA extends the original A PRIORI algorithm in a distributed database environment for circumstances where a small number of candidate sets are obtained and messages are exchanged between various sites in the distributed database for mining association rules.

Other research on machine learning of rules and patterns in very large databases has focused on using cluster analysis and pattern-based similarity search approaches. Cluster analysis is the grouping of abstract objects into groups of similar objects. In a large database context, cluster analysis provides an approach to divide large databases into smaller similar components. Cluster analysis is studied in statistics (Cheeseman & Stutz, 1996; Jain & Dubes, 1988), machine learning (Fisher, 1987; Fisher & McKusick, 1989) and data mining literature (Ester, Kriegel & Xu, 1995; Ng & Han, 1994; Weiss & Kapouleas, 1989). In statistics, Bayesian classification approaches have been used for clustering (Cheeseman & Stutz, 1996). Various *unsupervised* learning approaches were used for clustering in machine learning. The machine learning approaches differ from statistical approaches in that distance-based measures (used in statistical approaches) are replaced by measures that check for similarity between objects (Chen, Park & Yu, 1996a). Other approaches use conceptual clustering based methods with probability analysis. The probability analysis approaches, however, make the assumption that probability distributions of the attributes are independent of each other (Fisher, 1987, 1995). This assumption is challenged by a few researchers who believe that correlation between the attributes often exists (Chen et al., 1996b).

Pattern-based similarity approaches have been used for matching sequences in temporal databases (Agrawal, Faloutsos & Swami, 1993; Agrawal, Lin, Sawhney & Shim, 1995; Faloutsos & Lin, 1995; Faloutsos, Ranganathan & Manolopoulos, 1994; Li, Yu & Castelli, 1996; Mannila et al., 1994). There are two types of similarity search queries that support various data mining operations. First is called the *object similarity query* whereby a user searches for the collection of objects that are within the user defined distance from the queried object. The second type of query is called the all-pair similarity query where the objective is to find all the pairs of elements that are separated by user specified distances (Chen et al., 1996a,b. The similarity measures that are used fall into two categories: Euclidean distance similarity measures (Faloutsos & Lin, 1995; Faloutsos et al., 1994) and correlation based similarity measures (Li et al., 1996).

The aforementioned research related to data mining of large databases primarily focused on the development of algorithms and improvement of the performance of existing algorithms. The focus was on mining association rules, aggregating data in clusters and searching for relevant concepts in large databases using similarity search techniques.

A new stream of researchers have used statistical methods, case based reasoning and machine learning concepts for mining decision rules in order to acquire knowledge for expert systems. Unlike, the data mining research in large databases, these researchers used relatively smaller sets of databases that contained expert decisions. The problem that was addressed in the majority of this research was that of mining and developing models for classification/discrimination and forecasting.

Statistical techniques used for classification problems have included Fisher's linear discriminant analysis (LDA) (Fisher, 1936), quadratic and logistic discriminant analysis. Each of the techniques differs with respect to assumptions about group distributions and functional forms of the discriminant function. Linear models, given the ease of result of interpretation and reliability, were generally preferred for decision making (Hand, 1981); non-linear models, though more accurate on the training data, tended to show sharp declines in performance on unseen test samples (Altman, Eisenbeis & Sinkey, 1981). The major drawback with statistical methods is the fact that real-world data often did not satisfy the parametric distribution assumption. Non-parametric methods relax the parametric assumption and are less restrictive. Among the popular non-parametric methods used were the k-nearest neighbor and linear programming methods (Freed & Glover, 1981; Hand, 1981).

Machine learning techniques that are used for classification fall into two categories: (1) connectionist; and (2) inductive learning models where the discriminant function is expressed in a symbolic form using IF–THEN rules or decision tree. The Back-propagation neural network (Rumelhart, Hinton & Williams, 1986) was the most commonly used algorithm for connectionist schemes. Several induction algorithms were suggested for classification. Among the popular induction algorithms are CART (Breiman, Friedman, Olshen & Stone, 1984), ID3 (Quinlan, 1986), and CN2 (Clark & Niblett, 1987). A third set of techniques that has recently received attention for classification problems is based on the evolutionary computation paradigm which includes genetic algorithms and genetic programming (Bhattacharyya & Pendharkar, 1998; Koehler, 1991).

This wide variety of approaches made the selection of a particular technique that "best" matches a given problem a difficult task. The literature includes a number of studies comparing the performance of machine learning with statistical approaches (Atlas et al., 1990; Chung & Silver, 1992; Fisher & McKusick, 1989; Shavlik et al., 1991; Wolpert & Macready, 1995). Reviewing a number of comparative studies on symbolic and neural network methods, Quinlan (1986) emphasized that no single method uniformly demonstrates superior performance.

A study called "No Free Lunch" (NFL) theorems on search (Wolpert & Macready, 1995), pointed out that "positive performance in some learning situations must be balanced by negative performance in others" (Wolpert & Macready, 1995), i.e. search algorithms perform the same when performance is averaged over a sufficiently large group of problems. The NFL results emphasized that conclusions regarding a technique's performance can be made only with respect to the specific problem type being examined. Thus, the technique selected should be chosen with an understanding of its respective strengths and limitations. Bhattacharyya and Pendharkar (1998) performed several experiments on simulated data to compare the performance of statistical discriminant analysis, genetic algorithms, C 4.5, genetic programming and neural networks for classification. Their results indicate that the performance of the technique depends upon input data distribution characteristics that include group variance heterogeneity, distribution kurtosis and normality.

3. Description of mathematical and neural techniques used in the study

We used three techniques in this study. The three techniques are DEA (non-linear non-parametric mathematical technique), discriminant analysis (linear parametric statistical technique) and artificial neural networks (non-linear non-parametric technique). This section provides a formal description of DEA and ANN for binary classification problems.

3.1. Data envelopment analysis

DEA was a technique introduced by Charnes, Cooper and Rhodes (1978) to compare efficiencies of decision-making units (DMUs). The basic ratio DEA model seeks to determine a subset of k DMUs that determine the envelopment surface when all k DMUs consist of m inputs and s outputs. The envelopment surface determined by solving k linear programming models (one for each DMU) where all k DMUs appear in the constraints of the linear programming model. Let for DMU $i \in [0, \ldots, k], x_{mk} \geq 0$ denote the mth input value and $y_{sk} \geq 0$ denote the sth output value. The envelopment surface is determined by solving the following set of k linear programs.

$$\text{Max } \xi_0 = \frac{\sum_s O_s y_{sk_0}}{\sum_m I_m x_{mk_0}} \qquad k_0 = 1, \ldots, k \tag{1}$$

Subject to:

$$\frac{\sum_s O_s y_{sk_0}}{\sum_m I_m x_{mk_0}} \leq 1, \qquad \forall k \tag{2}$$

$$O_s, I_m \geq 0 \tag{3}$$

O_s, and I_m are output and input multipliers that are determined by the model. In the event of only one output of unity, model (1)–(3) can be rewritten as follows:

$$\text{Min } \eta_0 = \sum_m I_m x_{mk_0} \qquad k_0 = 1, \ldots, k \tag{4}$$

subject to:

$$\sum_m I_m x_{mk_0} \geq 1 \qquad \forall k \tag{5}$$

$$I_m \geq 0 \tag{6}$$

All the DMUs that have $\eta_0 = 1$ ((4)–(6)) are deemed efficient and lie on the efficient frontier (envelopment). Troutt et al. (1995) proposed that DEA develops an acceptance boundary for use in case-based expert systems for classification. Under the assumptions of conditional monotonicity, convexity of the acceptable set, representative of the sample cases, unrestricted selectivity, and no Type II error in acceptance, the cases that lie on or above the efficient frontier were deemed acceptable cases. The efficient frontier was obtained by solving the following linear programs:

$$\text{Max } \xi_0 = \frac{O_1}{\sum_m I_m x_{mk_0}} \qquad k_0 = 1, \ldots, k \tag{7}$$

subject to:

$$\frac{O_1}{\sum_m I_m x_{mk}} \geq 1, \quad \forall k \tag{8}$$

$$O_1, I_m \geq 0 \tag{9}$$

For a new case x^{new}, a set of linear programs (7)–(9) were solved by including the new case x^{new} in the set of already existing efficient cases E^* (called the reference set) obtained from Eqs. (7)–(9). It was possible that the x^{new} can alter the frontier. The following rule was used to decide the

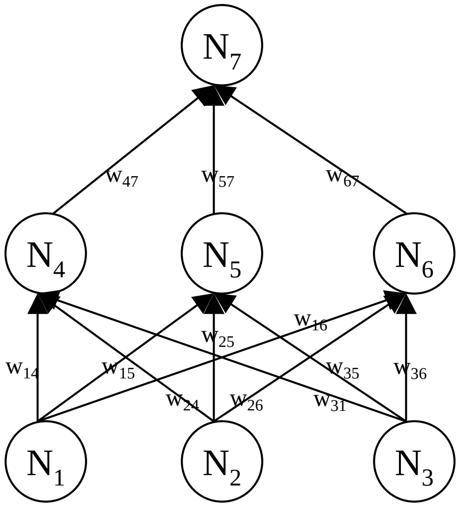

Fig. 1. A three-layer (of nodes) artificial neural network.

acceptability of the new case:

$$\text{If } \frac{O_1^*}{\sum_m I_m^* x_m^{\text{new}}} < 1 \qquad \text{then accept the case} \tag{10}$$

Else if $x^{\text{new}}$is efficient and does not alterE^*

then accept the case. (11)

Else reject the case.

O_1^*, I_m^* are the weights obtained by solving the linear programs similar to Eqs. (7)–(9) by including the new case in E^*. Alternately, under the assumption of convexity, the acceptability of the new case was determined by a simple convexity test. If a new case can be represented as a convex combination of the set of all cases in E^*, then the case is accepted; otherwise the case is rejected. Under the convexity of acceptable cases assumption, a new case was accepted if the following linear program is feasible:

$$\text{Max } \lambda \tag{12}$$

subject to:

$$\sum_j \lambda_j x_{ij} \le x_i^{\text{new}} \qquad \forall i, j, x_{ij} \in E^* \tag{13}$$

$$\sum_j \lambda_j = 1 \tag{14}$$

$$\lambda_j \ge 0 \qquad \forall j. \tag{15}$$

3.2. Back-propagation neural networks

Artificial neural networks (ANNs) were invented to mimic some of the phenomenon observed in Biology. The biological metaphor for ANNs is the human brain. An ANN consists of different sets of neurons or nodes and the connections between one set of neurons to the other. Each connection between two nodes in different sets is assigned a weight that shows the strength of the connection. Connections with positive weights are called excitatory connections and connections with negative weights are called inhibitory connections (Jain & Dubes, 1988; Rumelhart et al., 1986; Weiss & Kapouleas, 1989). The network of neurons and their connections together are referred to as the architecture of the ANN. Let $A = \{N_1, N_2, N_3\}$, $B = \{N_4, N_5, N_6\}$, and $C = \{N_7\}$ be three sets of nodes for an ANN.

The set A be the set of input nodes, set B the hidden set of nodes, and set C the output node. The cardinality of set A is equal to the number of input variables, and the cardinality of set C is equal to number of output variables. Each connection can be view as a mapping from either an input node to a hidden node or from a hidden node to an output node. The general architecture of three sets of nodes is referred to as a three-layer (of nodes) ANN. Fig. 1 illustrates a three-layer ANN. Notice that when the property $A \cap B \cap C = \Phi$ is always true, the network is referred to as a feed-forward network. The connections in a feed forward network are in one direction and the connections from A to B and from B to C are onto. The w_{ij} are the weights that denote the strength of connection from node i to node j.

Information is processed at each node in an ANN. For example, at hidden node N_4 the incoming signal vector (input) from the three nodes in the input set is multiplied by the strength of each connection and is then summed up. The result is passed through an activation function and the outcome is the activation of the node. If x represents the sum of the product of the incoming signal vector and the strength of connection, then the activation, using logistic sigmoid activation function, can be represented by,

$$f(x) = \frac{1}{1 + \mathrm{e}^{-x}}$$

In the back-propagation algorithm based learning, the strengths of connections are randomly chosen. Based on the initial set of randomly chosen weights, the algorithm tries to minimize the following root-mean-square error (RMS):

$$E = \tfrac{1}{2} \sum_{n=0}^{n=N} \|t_n - o_n\|^2$$

where N is number of patterns in the training set, t_n the target output of the nth pattern and o_n the actual output for nth pattern. In each subsequent training step, the initial set of random connection weights (strength of connections) is adjusted towards the direction of maximum decrease of E which is scaled by a learning rate λ. Mathematically, an old weight w_{old} is updated to its new value w_{new} using following equation:

$$w_{new} = w_{\text{old}} - \lambda \nabla E$$

where

$$\nabla E = \left(\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \ldots, \frac{\partial E}{\partial w_n} \right)$$

One useful property of the sigmoid function is that:

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = f(x)(1 - f(x))$$

This means that the derivative (gradient) of the sigmoid function can be calculated by applying the simple multiplication and subtraction operator on the function itself.

For any neuron n_k, its output is determined by

$$o_k = f(\sum_i w_{ik} o_k + \theta_k)$$

where w_{ik} is the weight on the arc connecting neuron n_k with the neuron n_i from the previous layer and with $f(\cdot)$ representing an activation function, usually the logistic function

$$f(x) = \frac{1}{1 + e^{-x}}$$

and θ_k is a bias associated with each neuron, effecting a translation of the logistic function that allows a better fit to the data. At the output layer node, an output value less than 0.5 is considered a categorization into Group 0 and values greater than 0.5 imply Group 1. The network is initialized with small random values for the weights, and the back-propagation learning procedure is used to update the weights as the data is iteratively presented to the input-layer neurons. For any neuron n_k, its output is determined by one of the following formulas:

$$h_k = \frac{1}{1 + e^{-\sum_{i=0}^{A} w_{1ik} x_i}} \forall k$$

$$= 1, \ldots B \text{ if the neuron is in the hidden layer}$$

$$o_k = \frac{1}{1 + e^{-\sum_{i=0}^{B} w_{2ik} h_i}} \forall k$$

$$= 1, \ldots C \text{ if the neuron is in the output layer}$$

where, A is number of input nodes and x_i the ith input, w_{1ik} the strength of connection from the ith input node to the kth hidden node, B number of hidden nodes, w_{2ik} the strength of connection from the ith hidden node to the kth output node, and C the number of output nodes. The weights w_{10k} and w_{20k} are the threshold weights and x_0 and h_0 are both equal to one. The network is initialized with small random values for the weights, and the back-propagation learning procedure is used to update the weights as the data is iteratively presented to the input-layer neurons. At each iteration the weights are updated by back-propagating the error as follows:

$$\Delta w1_{ik} = \eta \delta_k x_i \text{ and } \Delta w2_{ik} = \eta \delta_k h_k$$

where

$$\delta_k = \begin{cases} o_k(1 - o_k)(y_k - o_k), & \text{if } n_k \text{ is an output neuron} \\ h_k(1 - h_k) \sum_j w_{kj} \delta_j, & \text{if } \mathrm{n_k} \text{ is a hidden layer neuron.} \end{cases}$$

Here, η is the learning rate and y_k the actual output value.

For the experiments in this study, a three-layer network was used with four input nodes corresponding to the data attributes and a single output node. The number of hidden layer neurons chosen was twice the number of data inputs, a commonly used heuristic in the literature (Bhattacharyya & Pendharkar, 1998).

4. Data preparation, experiments and results

We used a data set provided by the Department of Surgery, Laurel Highlands Cancer Program at the Cone-

Table 1
The description of the fields in the data set

Variable	Description
Patient ID	The patient's unique identification number
First name	Patient's first name
Last name	Patient's last name
Cause of death	Cause of patient's death (cancer/other/unknown)
Age	Patient's age in years
Grade	Medical code for cancer severity (if any)
Size of tumor	Size of tumor in centimeters (if any)
Node positivity	Diagnosis of a patient with cancer (1 = yes, 0 = no)
Surgeon ID	The surgeon's unique identification number
Type of surgery	Code for type of surgery (if any)
Received radiation	Type of any radiation received (beam/radioactive isotopes)
Received Chemotherapy	Type of any chemo. Received (single/multiple agents)
Hormonal therapy	Type of any hormonal therapy received
Survival	Survival of the patient in weeks
Menopausal status	Cessation of menstruation in human female (1 = menopause, 0 = otherwise)
Estrogen receptor	The female sex hormone (1 = hormone, 0 = no hormone)
Progesterone Receptor	The hormone $C_{21}H_{30}O_2$ whose function is to prepare uterus for the reception and development of fertilized ovum by inducing secretion in proliferated glands (1 = hormone, 0 = no hormone)

Table 2
Input factors

Age
Menopausal status
Estrogen receptor positivity
Progesterone receptors positivity

maugh’s Memorial Medical Center of Johnstown, Pennsylvania. The raw data set was comprised of 479 patients and includes, for each record of data, the following fields, listed in Table 1:

The primary goal of the research was to analyze a limited list of potential factors in an attempt to predict the node positivity of any tumors. We used the set of factors listed in Table 2 as input variables. The output variable was node positivity of the tumor that was a value of 1 for a correct diagnosis, i.e. the patient was diagnosed with breast cancer.

To establish the final mining data set, a screening process, described below, was conducted. After the data was screened, the set was then divided into various sample sets and prepared for execution.

Screening process: data screening was performed to ensure the data set was complete and valid. The primary conditions, which caused a record to be flagged for omission, are listed in Table 3.

After an initial screening, 454 useable records remained. The range of ages (illustrated in Fig. 2) included was 29–88.

Division of the sample sets: the final data set was divided into two data sets; one for learning the patterns and the other for testing the predictive performance. As sample size plays an important role for the learning and predictive performance of the techniques, the 454 records were randomly split into two sets for three different times (depending on the proportion of cases in learning set and test set). Care was taken to ensure that a proportional number of cases with a node positivity of one (1) existed in each part of each of the sample set. The sample sets utilized are listed in Table 4:

Once the records were divided, extraneous data associated with each record was eliminated. Thus, the data set was modified to include only items that were the inputs or the output for the study.

For learning associations, all records with node positivity equal to one to learn associations were included. Using the A PRIORI (Agrawal & Srikant, 1995) algorithm, described in Section 2, we learned the associations between different female hormones and the occurrence of breast cancer. Age being a continuous variable was not used for mining associations. Table 5 illustrate the results of the experiments.

Table 3
Conditions for data ommission

Undefined node positivity value
Missing patient first &/or last name
Null value in a required field[a]

[a] Required fields are defined to be an input or output factor.

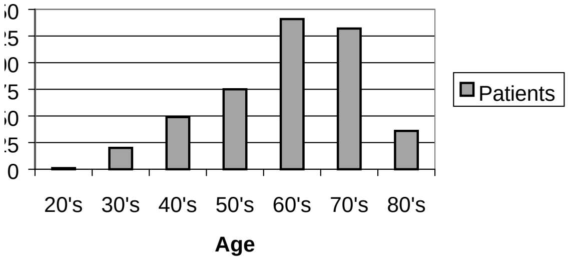

Fig. 2. Age distribution.

After learning associations, the predictive accuracy of the DEA, ANN and discriminant models was examined using the three different scenarios illustrated in Table 4. The SAS PROC DISCRIM procedure was used to run discriminant analysis and a commercial software implementing backpropagation algorithm was used to run ANN for our experiments. For the ANN experiments, we changed the network architecture (number of hidden nodes) to monitor its impact on the predictive and learning performance. The two architectures that we used were one with number of hidden nodes equal to 6 and the other with the number of hidden nodes equal to 11.

As DEA classifies all the training set cases correctly and uses information about one class only, the training performance of DEA was 100%. For testing the performance of DEA on the holdout sample, we took the test set examples and tested each example, one at a time, using commercial DEA software. The strategy was to add an additional test example to the existing training and also to set the node positivity attribute equal to 1. The DEA model, containing all training examples and one test example, was then executed. After the analysis, the following heuristic was employed to determine the node positivity of test example:

$$\text{If test example efficiency}$$

$$= 100\% \ \textit{then node positivity} \neq 1$$

$$\text{ELSE node positivity} = 1$$

Different tests, equal to the number of examples in our test data set, were also run. The results were compared to the actual value of the node positivity in the test cases. Table 6 summarizes the results of each of the algorithms used along with the percentage of accurate patterns.

The results indicate that non-linear, non-parametric techniques ANN and DEA outperform the linear and statistical approach of discriminant analysis. As the training sample size increases, learning and prediction accuracy of linear discriminant analysis increases. The prediction accuracy of DEA increases with increases in the training sample size. The learning and prediction accuracy of ANN remains relatively unchanged with an increase in the training set sample size.

While discriminant analysis techniques play a valuable

Table 4
Sample set division

Sample set label	Pattern set			Test set		
	% of records	Number of records	Node positivity = 1	% of records	Number of records	Node positivity = 1
25/75	25	114	21	75	340	63
50/50	50	227	42	50	227	42
75/25	75	340	63	25	114	21

role in predicting if the patient has breast cancer or not, these techniques provide little information about the "risk" that a patient is likely to have breast cancer. The prediction of the likelihood of breast cancer is important for several reasons. First, frequent check-ups can be recommended for a patient with a high risk of cancer.. Early detection of cancer can not only save the patient's life, but also helps the physician remove the cancer through radiation treatment and painful procedures such as, surgery can be avoided. Association rules can play an important role in predicting the patient's "risk" of cancer. We believe that an expert system can be created that will help a physician assess the patient's risk of cancer. For example, the association rules, illustrated in Table 5, can be easily converted into following rules (expressed in predicate logic):

$$\mathrm{R1}: \forall x : \mathrm{Estrogen}(x) \longrightarrow \mathrm{Breast\_Cancer}(x)\mathrm{CF} = 0.74$$

$$\mathrm{R2}: \forall x : \mathrm{Progesterone}(x) \longrightarrow \mathrm{Breast\_Cancer}(x)\mathrm{CF} = 0.62$$

$$\mathrm{R3}: \forall x : \mathrm{Menopause}(x) \longrightarrow \mathrm{Breast\_Cancer}(x)\mathrm{CF} = 0.59$$

$$\mathrm{R4}: \forall x : \mathrm{Menopause}(x) \wedge \mathrm{Estrogen}(x) \longrightarrow \mathrm{Breast\_Cancer}(x)\mathrm{CF} = 0.50$$

$$\mathrm{R5}: \forall x : \mathrm{Menopause}(x) \wedge \mathrm{Progesterone}(x) \longrightarrow \mathrm{Breast\_Cancer}(x)\mathrm{CF} = 0.47$$

$$\mathrm{R6}: \forall x : \mathrm{Estrogen}(x) \wedge \mathrm{Progesterone}(x) \longrightarrow \mathrm{Breast\_Cancer}(x)\mathrm{CF} = 0.56$$

Table 5
Female hormones and their associations with breast cancer

Association rule	Confidence factor
Menopause → breast cancer	0.588
Estrogen → breast cancer	0.735
Progesterone → breast cancer	0.618
Menopause ∧ estrogen → breast cancer	0.500
Menopause ∧ estrogen ∧ progesterone → breast cancer	0.441
Estrogen ∧ progesterone → breast cancer	0.559
Menopause ∧ progesterone → breast cancer	0.471

The above rules, when coded in an expert system, can help a physician assess a patient's risk of contracting breast cancer. For example, if a patient, Mary, has reached the age of menopause and is taking Estrogen then using certainty factors, Mary's risk can be calculated as follows:

$$\mathrm{CF}_{\mathrm{Mary}} = \mathrm{CF}_{\mathrm{R4}} + (1 - \mathrm{CF}_{\mathrm{R4}})(\mathrm{CF}_{\mathrm{R1}} + (1 - \mathrm{CF}_{\mathrm{R1}})\mathrm{CF}_{\mathrm{R3}})$$

$$\mathrm{CF}_{\mathrm{Mary}} = 0.947$$

The certainty factor for Mary's risk of breast cancer is about 0.95 or 95%. Although the risk for breast cancer appears high, it is important to note that above factors are not the only factors that determine the occurrence of breast cancer. Thus, even though Mary's risk of getting cancer, as determined by the expert system, is 95%, Mary may not necessarily contract breast cancer in the future. In an event, frequent check-ups will definitely save Mary's life by early detection.

4.1. Comparison to previous studies

Comparing the results of our study with other studies on an equitable basis is difficult for the following reasons:

1. DEA uses information about one class to determine the discriminant function whereas, other techniques use information about 2 classes to determine the discriminant function.
2. The performance of DEA is likely to vary if DEA and its

Table 6
The DEA, ANN and LDA comparison results

Hidden nodes	Data ratio	*25/75* (%)	*50/50 (%)*	*75/25 (%)*
Artificial neural networks				
n = 6	Good pats (Learning)	82	82	81
	Good pats (Test)	81.5	81.5	81.5
n = 11	Good pats (Learning)	81	81	81
	Good pats (Test)	81.6	81.5	81.5
Data envelopment analysis				
	Test	62.1	66.5	67
Linear discriminant analysis				
	Learning	65	68	66.8
	Test	60.5	66.1	65.6

variant is used for the 2 classes separately (Pendharkar & Kumar, 1998).

3. The attributes considered in this study are different from the attributes considered in other studies.
4. The data sets are different.

The results of the experiments performed by Michalski, Mozetic, Hong and Lavrac (1986) regarding the prognosis of breast cancer recurrence yielded 66% classification accuracy. This particular study utilized a 70/30 split of the pattern versus test data for 286 examples. The study considered nine attributes as input factors. The specific attributes considered are unknown.

Clark and Niblett (1987a,b) performed a study to attempt to predict the recurrence of breast cancer within five years. This study utilized a 70/30 split of the data set comprised of 286 patients. The study considered nine attributes, unknown from the article, as the input factors. The resulting accuracy achieved was a range between 65 and 72% based on the various algorithms tested.

These studies illustrate the capability of the various types of algorithms to convey accurate classifications in the medical domain of breast cancer. In each of the above listed cases, the data set size was significantly smaller than the set used for the current research.

5. Summary and future work

We have used DEA and ANN as a tool for mining breast cancer patterns. Our results indicate that DEA is a competitive tool for binary classification problems. In the cases where the number of examples in one class is significantly greater than the number of examples in the second class, DEA may be an appropriate tool to use. The results of our study indicate that neural networks outperform DEA in terms of prediction accuracy. One of the reasons for superior performance of neural networks over DEA is that DEA assumes the convexity of the acceptable cases and neural networks relax this assumption. Both DEA and neural networks outperform the traditional statistical discriminant analysis.

While classification approaches help a physician diagnose breast cancer, our experiments with learning association rules show that risk assessment expert systems can be developed. In the fight against breast cancer, we believe that the combination of association rules and classification approaches will provide an effective means to accurate and economical breast cancer diagnosis.

Several other approaches can be used for learning about breast cancer patterns. For example, in our study, we didn't use a popular machine learning technique called ID3 (Quinlan, 1986). ID3 is a non-parametric machine learning technique that uses an information-theoretic induction based approach to construct decision tree from training data. Although, ID3 provides the classification rules, it does not output certainty factors. We selected association rules for this reason. ID3, however, provides information about the most important discriminatory attribute (based on the information measure) which makes it a candidate for future investigation. Future work may focus on using ID3 for learning decision tree from the existing breast cancer data.

Acknowledgements

The research done in this study was funded by a research grant from University of Pittsburgh.

References

Agrawal, R., Faloutsos, C., Swami, A. (1993). Efficient similarity search in sequence databases. *Proceedings of Fourth International Conference Foundations of Data Organization and Algorithms*, October.

Agrawal, R., Lin, K.I., Sawhney, H.S., Shim, K. (1995). Fast similarity search in the presence of noise, scaling, and translation in time series databases. *Proceedings of 21st International Conference on Very Large DataBases* (pp. 490–501).

Agrawal, R., Srikant, R. (1995). Fast algorithms for mining association rules in large databases. *Proceedings of 20th International Conference on Very Large Database* (pp. 478–499)

Altman, E. L., Eisenbeis, R. A., & Sinkey, J. (1981). *Application of classification techniques in business, banking and finance*, Greenwich, CT: JAI Press.

Atlas, L., Cole, R., Connor, J., El-Sharkawi, M., Marks II, R.J., Muthusamy, Y., Barnard, E. (1990). Performance comparisons between backpropagation networks and classification trees on three real-world applications. *Advances in Neural Information Processing Systems* (vol. 2). Denver, CO.

Bhattacharyya, S., & Pendharkar, P. C. (1998). Inductive, evolutionary and neural techniques for discrimination: a comparative study. *Decision Sciences*, *28* (4), 000.

Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. (1984). *Classification and regression trees*, Monterey, CA: Wadsworth.

Cheeseman, P., & Stutz, J. (1996). Bayesian classification (AutoClass): theory and results. In U. M. Fayyad & G. Piatetsky-Shapiro & R. Uthurusamy (Eds.), *Advances in knowledge discovery and data mining*, (pp. 153–180). Cambridge, MA: AAAI/MIT Press.

Chen, M.S., Park, J.S., Yu, P.S. (1996). Data mining for path traversal patterns in a web environment. *Proceedings of 16th International Conference on Distributed Computing Systems* (pp. 385–392).

Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: an overview from database perspective. *IEEE Transactions on Knowledge and Data Engineering*, *8* (2), 866–883.

Cheung, D. W., Ng, V. T., Fu, A. W., & Fu, Y. (1996). Efficient mining of association rules in distributed databases. *IEEE Transactions on Knowledge and Data Engineering*, *8* (6), 911–922.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*, 1978.

Chung, H. M., & Silver, M. S. (1992). Rule-based expert systems and linear models: an empirical comparison of learning-by-examples methods. *Decision Sciences*, *23*, 687–707.

Clark, P., & Niblett, T. (1987). *Progress in machine learning (from the Proceedings of the Second European Working Session on Learning)*, (pp. 11–30). *Induction in noisy domains*Bled, Yogoslavia: Sigma Press.

Clark, P., & Niblett, T. (1987). The CN2 induction algorithm. *Machine Learning*, *3* (4), 261–283.

Elmore, J., Wells, M., Carol, M., Lee, H., Howard, D., & Feinstein, A. (1994). Variability in radiologists' interpretation of memograms. *New England Journal of Medicine*, *331* (22), 1493–1499.

Ester, M., Kriegel, H.P., Xu, X. (1995). Knowledge discovery in large spatial databases: focusing techniques for efficient class identification. *Proceedings Fourth International Symposium of Large Spatial Databases* (pp. 67–82). Portland, Maine.

Faloutsos, C., & Lin, K. I. (1995). FastMap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. *Proceedings ACM Sigmoid*,, 163–174.

Faloutsos, C., Ranganathan, M., & Manolopoulos (1994). Fast subsequence matching in time-series databases. *Proceedings ACM Sigmoid*,, 419–429 Minneapolis.

Freed, N., & Glover, F. (1981). A linear programming approach to the discriminant problem. *Decision Sciences*, *12*, 68–74.

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

Fisher, D.H., McKusick, K.B. (1989). An empirical comparison of ID3 and back-propagation. *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence* (pp. 788–793). Detroit, MI. Los Altos, CA: Morgan Kaufmann.

Fisher, D. (1987). Improving inference through conceptual clustering. *Proceedings AAAI Conference* (pp. 461–465). Seattle.

Fisher, D. (1995). Optimization and simplification of hierarchical clusterings. *Proceedings First International Conference on Knowledge Discovery and Data Mining (KDD '95)* (pp. 118–123). Montreal, CA.

Han, J., Fu, Y. (1995). Discovery of multiple-level association rules from large databases. *Proceedings of the 21st International Conference on Very Large DataBases* (pp. 420–431).

Hand, D. J. (1981). *Discrimination and classification*, New York: Wiley.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*, Englewood Cliffs, NJ: Prentice-Hall.

Koehler, G. J. (1991). Linear discriminant functions determined through genetic search. *ORSA Journal on Computing*, *3* (4), 345–357.

Kovalerchuck, B., Triantaphyllou, E., Ruiz, J. F., & Clayton, J. (1997). Fuzzy logic in computer-aided breast cancer diagnosis: analysis of lobulation. *Artificial Intelligence in Medicine*, *11*, 75–85.

Li, C. S., Yu, P.S., Castelli, V. (1996). Hierarchy scan: a hierarchical similarity search algorithm for databases of long sequences. *Proceedings of 12th International Conference on Data Engineering*.

Mannila, H., Toivonen, H., Verkamo, I. A. (1994). efficient algorithms for discovering association rules. *Proceedings AAAI Workshop of Knowledge Discovery in Databases* (pp. 181–192).

Michalski, R.S., Mozetic, I., Hong, J., Lavrac, N. (1986). The multi-purpose incremental learning system aq15 and its testing application to three medical domains. *Proceedings of the Fifth National Conference on Artificial Intelligence* (pp. 1041–1045). Philadelphia, PA: Morgan Kauffman.

Ng, R., Han, J. (1994). Efficient and effective clustering method for spatial data mining. *Proceedings of International Conference on Very Large DataBases* (pp. 144–155). Santiago, Chile.

Park, J. S., Chen, M. S., & Yu, P. S. (1995). Efficient parallel data mining for association rules. *Proceedings ACM SIGMOD*,, 175–186.

Pass, S. (1997). Discovering value in a mountain of data. *OR/MS Today*,, 24–28.

Pendharkar, P.C., Kumar, S. (1998). A data envelopment analysis application for marginal cost assignment in certain case based expert system. *Proceedings of Third INFORMS Conference on Information Systems and Technology* (pp. 347–358). Montreal.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, *1*, 81–106.

Rumelhart, D. E., Hinton, G. E., & William, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: exploration in the microstructure of cognition*, *Foundations*, 1. Cambridge, MA: MIT Press.

Savasere, A., Omiecinski, E., Navathe, S. (1995). An efficient algorithm for mining association rules in large databases. *Proceedings of 21st International Conference on Very Large DataBase* (pp. 432–444).

Shavlik, J. W., Mooney, R. J., & Towell, G. G. (1991). Symbolic and neural learning algorithms: an experimental comparison. *Machine Learning*, *6*, 111–143.

Srikant, R., Agrawal, R. (1995). Mining generalized association rules. *Proceedings of 21st International Conference on Very Large DataBase* (pp. 407–419).

Troutt, M. D., Rai, A., & Zhang, A. (1995). The potential use of DEA for credit applicant acceptance systems. *Computers and Operations Research*, *4*, 405–408.

Weiss, S.M., Kapouleas, I. (1989). An empirical comparison of pattern recognition, neural nets, and machine learning classification methods. *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence* (pp. 688–693). Detroit, MI. Los Altos, CA: Morgan Kaufmann.

Wingo, P. A., Tong, T., & Bolden, S. (1995). Cancer statistics. *Ca-A Cancer Journal for Clinicians*, *45* (1), 8–30.

Wolpert, D.H., Macready, W.G. (1995). No free lunch theorems for search. *Santa Fe Institute Technical Report* No. SFI-TR-95-02-010.