GMDH-Based Feature Ranking and Selection for Improved Classification of Medical Data

R. E. Abdel-Aal
Physics Department, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

Address for corresponding author and reprints:

Dr. R. E. Abdel-Aal
P. O. Box 1759
KFUPM
Dhahran 31261
Saudi Arabia

e-mail: radwan@kfupm.edu.sa
Phone: +966 3 860 4320
Fax: +966 3 860 4281

Abstract

Medical applications are often characterized by a large number of disease markers and a relatively small number of data records. We demonstrate that complete feature ranking followed by selection can lead to appreciable reductions in data dimensionality, with significant improvements in the implementation and performance of classifiers for medical diagnosis. We describe a novel approach for ranking all features according to their predictive quality using properties unique to learning algorithms based on the group method of data handling (GMDH). An abductive network training algorithm is repeatedly used to select groups of optimum predictors from the feature set at gradually increasing levels of model complexity specified by the user. Groups selected earlier are better predictors. The process is then repeated to rank features within individual groups. The resulting full feature ranking can be used to determine the optimum feature subset by starting at the top of the list and progressively including more features until the classification error rate on an out-of-sample evaluation set starts to increase due to overfitting. The approach is demonstrated on two medical diagnosis datasets (Breast Cancer and Heart Disease) and comparisons are made with other feature ranking and selection methods. Receiver operating characteristics (ROC) analysis is used to compare classifier performance. At default model complexity, dimensionality reduction of 22% and 54% could be achieved for the breast cancer and heart disease data, respectively, leading to improvements in the overall classification performance. For both datasets, considerable dimensionality reduction introduced no significant reduction in the area under the ROC curve. GMDH-based feature selection results have also proved effective with neural network classifiers.

Keywords:

Abductive networks, neural networks, feature ranking, feature selection, dimensionality reduction, classification accuracy, ROC characteristics, medical diagnosis, breast cancer, heart disease.

1. Introduction

Machine learning classification techniques provide support for the decision-making process in many areas of health care, including screening, diagnosis, prognosis, monitoring, therapy, survival analysis, and hospital management. Tools used include Bayesian and nearest-neighbor classifiers, rule induction methods, decision trees, fuzzy logic, artificial neural networks, and abductive networks [1] based on the group method of data handling (GMDH) algorithm [2]. Compared to neural networks, abductive networks allow easier model development and provide more transparency and greater insight into the modeled phenomena, which are important advantages in medicine. Medical applications of GMDH-based techniques include modeling obesity [3], analysis of school health surveys [4], drug detection from EEG measurements [5], medical image recognition [6], and screening for delayed gastric emptying [7]. Accuracy is very important in classifiers used for medical applications. A high percentage of false negatives in screening systems increases the risk of real patients not receiving the attention they need, while a high false alarm rate causes unwarranted worries and increases the load on medical resources. In quest for higher classification accuracies, feature subset selection has been used for data reduction in areas characterized by high dimensionality due to the large number of available features, e.g. in remote sensing [8], seismic data processing [9], speech recognition [10], drug design [11], and image segmentation [12]. This approach attempts to select a small subset of optimum features that ideally is necessary and sufficient to describe the phenomenon being modeled [13]. Feature subset selection is expected to improve classification performance, particularly in situations characterized by the high data dimensionality problem caused by relatively few training examples compared to a large number of measured variables. This situation arises frequently in medicine where considerations of risk, time, difficulty, cost, and inconvenience may limit the number of training examples, while the number of disease markers increases rapidly over the years [14]. Even if no significant

improvements in classification performance are achieved, feature reduction has many practical advantages in reducing the number of measurements required, shortening training and execution times, and improving model compactness, transparency and interpretability. Fewer model inputs result in simpler models that train and execute faster, and allow training on smaller datasets without the risk of overfitting. Reducing the number of attributes to be measured for model implementation makes screening tests faster, more convenient and less costly. Simpler models with fewer inputs are also more transparent and more comprehensible, providing better explanation of suggested diagnosis, which is an important requirement in medical applications. Discarding irrelevant and redundant features reduces noise and spurious correlations with the output, and avoids the problems of colinearity between inputs, e.g. instability of least squares estimates and removal of solution uniqueness [15]. Feature reduction has been applied to several areas in medicine, including: classification of EEG signals for operating brain-computer interfaces [16], classification of hepatic lesions from computed tomography images [17], detection of mass lesions in digital mammograms [18], segmenting digital chest radiographs [19], processing of ECG signals for the detection of obstructive sleep apnea [20], classification of ultrasound liver tissues using the wavelet transform [21], and detection of seizure events in newborn children using EEG data [22].

Techniques for feature subset selection can be classified into three main categories: embedded, filter (open-loop), and wrapper (closed-loop) techniques [23]. With embedded techniques, feature selection is performed as part of the induction learning itself. By testing the values of certain features, decision tree algorithms seek to split the training data into subsets, each containing a strong majority of one class. Both filter and wrapper techniques perform feature selection as a preprocessing step prior to the modeling application, with the objective of selecting an optimum feature subset that serves as an input to the learning algorithm. Filter techniques do not use the learning mechanism for feature selection. They filter out undesirable and redundant features

through checking data consistency and eliminating features whose information content is represented by others. Examples of filter techniques for feature selection include Relief [13], which ranks individual features according to a feature relevance score. The correlation-based feature selection (CFS) technique [24] scores and ranks subsets of features, rather than individual features. It uses the criterion that a good feature subset for a classifier contains features that are highly correlated with the class variable but poorly correlated with each other. Information theoretic measures, such as the mutual information criterion, were used for feature selection to avoid mistakes introduced by linear measures such as correlation [25]. The Bhattacharyya probabilistic distance and other statistical measures were used to select feature subsets that maximize class separability [26]. Since filter methods do not use the learning algorithm, they are fast and therefore suitable for use with large databases. Also, resulting feature selections are applicable to various learning techniques. Wrapper techniques [27] search for an optimal feature subset through testing the performance of candidate subsets using the learning algorithm. As the learning algorithm is called repeatedly, wrapper methods are slower than filter methods and do not scale up well to large, high-dimensional datasets, particularly with neural networks, which require long training times. To overcome this limitation, feature subset evaluation could use a simpler learning algorithm, e.g. nearest-neighbour classifier, that is closely related to the target neural network architecture [28]. Wrapper feature selections are unique to the learning algorithm used, and the process should be repeated for a different learning algorithm. Strategies used for searching the feature space include sequential feature selection (SFS) methods [29], either with forward sequential search (FSS) or backward sequential search (BSS). FSS starts with an empty set, adding single features that best improve performance criteria. BSS starts with the full feature set and sequentially removes features whose removal leads to maximum gain in performance. Genetic algorithm (GA) search methods

have been used with both filters [12] and wrappers [28]. Feature selection techniques based on the rough set theory have also been proposed [30].

This paper describes a novel technique for feature ranking and selection with GMDH-based abductive network classifiers. The method relies on the property of the GMDH learning algorithm [1,2] of automatically selecting optimum predictors [31] at various levels of model complexity specified by the user. Information gathered in this way is used to rank the available features according to their predictive quality. In a previous publication [32], we described an approach for the rough ranking of features into groups, and using the resulting ranking for assigning feature subsets of uniform predictive quality to members of a network ensemble. In this paper, the approach is refined by adopting a 2-stage hierarchical ranking procedure to achieve full ranking of individual features for use in the different application area of dimensionality reduction. The resulting ranking can be used to select a given number of features, starting at the top, to build a classifier with a reduced subset of input features. An optimum feature subset can be derived by successively including ranked features one by one, starting with the best feature at the top of the ranking list, and evaluating the resulting classifier on an out-of-sample evaluation dataset. The process is continued as long as the classification error rate on the evaluation set decreases, stopping when the error rate starts to rise due to overfitting. Feature ranking according to predictive quality gives insight into the most effective markers for the diagnosis problem, which should be of interest to medical practitioners. The technique is demonstrated using two standard medical diagnosis datasets from the UCI Machine Learning Repository [33]. Section 2 gives a brief introduction to the GMDH algorithm, the abductive network modeling tool used, and the approach adopted for feature ranking. Section 3 gives a brief outline of the two medical datasets used in the investigation. Section 4 presents the results obtained. In all cases, classifiers trained on the optimum feature sets selected outperform those trained on the full feature sets. Improvements are greater for a dataset that is more

prone to high dimensionality problems. Receiver operating characteristics (ROC) analysis is used to investigate the effect of dimensionality reduction on classifier performance and to compare the proposed method with other feature ranking and selection techniques. Conclusions are made and suggestions for future work given in Section 5.

2. Methods

2.1 GMDH and AIM Abductive Networks

AIM (abductory inductive mechanism) [34] is a supervised inductive machine-learning tool for automatically synthesizing abductive network models from a database of inputs and outputs representing a training set of solved examples. As a GMDH algorithm, the tool can automatically synthesize adequate models that embody the inherent structure of complex and highly nonlinear systems. Automation of model synthesis not only lessens the burden on the analyst but also safeguards the model generated against influence by human biases and misjudgments. The GMDH approach is a formalized paradigm for iterated (multi-phase) polynomial regression capable of producing a high-degree polynomial model in effective predictors. The process is 'evolutionary' in nature, using initially simple (myopic) regression relationships to derive more accurate representations in the next iteration. To prevent exponential growth and limit model complexity, the algorithm selects only relationships having good predicting powers within each phase. Iteration is stopped when the new generation regression equations start to have poorer prediction performance than those of the previous generation, at which point the model starts to become overspecialized and therefore unlikely to perform well with new data. The algorithm has three main elements: representation, selection, and stopping. It applies abduction heuristics for making decisions concerning some or all of these three aspects.

To illustrate these steps for the classical GMDH approach, consider an estimation data base of n_e observations (rows) and $m+1$ columns for m independent variables (x_1, x_2, ..., x_m) and one

dependent variable y. In the first iteration we assume that our predictors are the actual input variables. The initial rough prediction equations are derived by taking each pair of input variables (x_i, x_j ; $i,j = 1,2,...,m$) together with the output y and computing the quadratic regression polynomial [2]:

$$y = A + B\,x_i + C\,x_j + D\,x_i^2 + E\,x_j^2 + F\,x_i\,x_j \tag{1}$$

Each of the resulting $m(m-1)/2$ polynomials is evaluated using data for the pair of x variables used to generate it, thus producing new estimation variables (z_1, z_2, ..., $z_{m(m-1)/2}$) which would be expected to describe y better than the original variables. The resulting z variables are screened according to some selection criterion and only those having good predicting power are kept. The original GMDH algorithm employs an additional and independent selection set of n_s observations for this purpose and uses the regularity selection criterion based on the root mean squared error r_k over that dataset, where:

$$r_k^2 = \sum_{\ell=1}^{n_s}(y_\ell - z_{k\ell})^2 \Big/ \sum_{\ell=1}^{n_s} y_\ell^2; \quad k = 1,2,...,m(m-1)/2 \tag{2}$$

Only those polynomials (and associated z variables) that have r_k below a prescribed limit are kept and the minimum value, r_{min}, obtained for r_k is also saved. The selected z variables represent a new database for repeating the estimation and selection steps in the next iteration to derive a set of higher-level variables. At each iteration, r_{min} is compared with its previous value and the process is continued as long as r_{min} decreases or until a given model complexity is reached. An increasing r_{min} is an indication of the model becoming overly complex, thus overfitting the estimation data and performing poorly on the new selection data. Keeping model complexity checked is an important aspect of GMDH-based algorithms, which keep an eye on the final objective of constructing the model, i.e. using it with new data previously unseen during training. The best model for this purpose is that providing the shortest description for the data available [35]. Computationally, the

resulting GMDH model can be seen as a layered network of partial quadratic descriptor polynomials, each layer representing the results of an iteration.

A number of GMDH methods have been proposed which operate on the whole training dataset thus eliminating the need for a dedicated selection set. The adaptive learning network (ALN) approach, AIM being an example, uses the predicted squared error (PSE) criterion [35] for selection and stopping to avoid model overfitting, thus solving the problem of determining when to stop training in neural networks. The criterion minimizes the expected squared error that would be obtained when the network is used for predicting new data. AIM expresses the *PSE* as:

$$PSE = FSE + CPM\,(2K/N)\sigma_p^{\,2} \tag{3}$$

where *FSE* is the fitting squared error on the training data, *CPM* is a complexity penalty multiplier selected by the user, *K* is the number of model coefficients, *N* is the number of samples in the training set, and $\sigma_p^{\,2}$ is a prior estimate for the variance of the error obtained with the unknown model. This estimate does not depend on the model being evaluated and is usually taken as half the variance of the dependent variable *y* [35]. As the model becomes more complex relative to the size of the training set, the second term increases linearly while the first term decreases. *PSE* goes through a minimum at the optimum model size that strikes a balance between accuracy and simplicity (exactness and generality). The user may optionally control this trade-off using the *CPM* parameter. Larger values than the default value of 1 lead to simpler models that are less accurate but may generalize well with previously unseen data, while lower values produce more complex networks that may overfit the training data and degrade actual prediction performance.

AIM builds networks consisting of various types of polynomial functional elements. The network size, element types, connectivity, and coefficients for the optimum model are automatically determined using well-proven optimization criteria, thus reducing the need for user intervention

compared to neural networks. This simplifies model development and considerably reduces the learning/development time and effort. The models take the form of layered feed-forward abductive networks of functional elements (nodes) [34], see Fig. 1. Elements in the first layer operate on various combinations of the independent input variables (x's) and the element in the final layer produces the predicted output for the dependent variable y. In addition to the main layers of the network, an input layer of normalizers convert the input variables into an internal representation as Z scores with zero mean and unity variance, and an output unitizer unit restores the results to the original problem space. AIM supports the following main functional elements:

(i) A white element which consists of a constant plus the linear weighted sum of all outputs of the previous layer, i.e.

$$\text{"White" Output} = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + ... + w_nx_n \quad (4)$$

where $x_1, x_2,..., x_n$ are the inputs to the element and $w_0, w_1, ..., w_n$ are the element weights.

(ii) Single, double, and triple elements which implement a third-degree polynomial expression with all possible cross-terms for one, two, and three inputs respectively; for example,

$$\text{"Double" Output} = w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2 + w_5x_1x_2 + w_6x_1^3 + w_7x_2^3 \quad (5)$$

2.2 GMDH-Based Feature Ranking and Selection

This paper describes a hierarchical approach to perform complete ranking of the input features according to their predictive quality by using the GMDH-based AIM learning algorithm to automatically select optimum predictors at various stages of model complexity. In the first stage of the procedure, features are ranked in groups of different predictive quality. With all input features available for use by the model, we start by using a large CPM value to synthesize a simple model consisting of a single White or Triple element using a group of only three input features that are automatically selected by the learning algorithm. Such features would be those having the best

predictive quality among the feature set. The modeling process is then repeated with a lower CPM value to allow the synthesis of a slightly more complex model that selects in another group of features that will have a lower predictive power compared to the first group. The process continues until all features are selected. Recommended minimum value for the CPM parameter is 0.1 for the most complex model [34]. If the most complex model still leaves some features unselected, all features selected thus far are disabled as inputs to enforce selection from the remaining features and allow completion of the feature ranking procedure. In the second stage, features belonging to each group are ranked within the group by repeating the modeling process using only those features belonging to the group as model inputs. In this way, full ranking of all individual features is obtained. The first selected feature has the highest predictive value and is located at the top of the ranking list, followed by the second feature to be selected, etc. Two approaches can be adopted for selecting a feature subset from the ranking list. In the first approach, a compact *m*-feature subset can be obtained by taking the first *m* features starting from the top of the ranking list. In the second approach, the optimum subset of features is determined by repeatedly forming subsets of *k* features, $k = 1, 2, 3, \ldots, n$, where *n* is the total number of available features, starting from the top of the ranking list. A classifier is trained on each of the formed subsets. As *k* increases, classification error rate for the resulting models on the training set is expected to monotonically decrease as the models fit the training data more accurately. However, performance on an out-of-sample evaluation dataset would first improve and then starts to deteriorate due to the model overfitting the training data. The optimum model corresponding to the optimum feature subset would correspond to the smallest value for *k* where the minimum classification error rate is reached on the evaluation set. Overfitting is expected to occur earlier, i.e. at lower *k* values, for smaller training sets and with more complex models.

3. Material

Two standard medical diagnosis datasets from the UCI Machine Learning Repository [33] were used for this study. These include the Wisconsin breast cancer dataset and the Cleveland heart disease dataset. Out of the 699 records for the breast cancer dataset, 16 records containing missing attributes were deleted, leaving 683 for use. In all cases, the dataset was randomly split into a training set comprising approximately 70% of the data and an evaluation set consisting of the remaining 30%. Appendix A lists the row numbers for the evaluation sets used. Remaining rows constitute the training sets. Unless otherwise mentioned, all models were trained on the same training set and evaluated on the same evaluation set. Table 1 lists important statistics on the datasets, including the number of features, number of records and percentage prevalence of positives in the total, training, and evaluation sets. The ratio of training set size to the number of features is 54 for the breast dataset and 15 for the heart dataset. Therefore, the datasets represent two different situations of relatively low and high data dimensionality, respectively. Table 2 lists the names or brief descriptions of the features for each dataset. The feature number used in the table is the column number for the feature in the dataset, and will be used to identify the feature throughout this paper. Following is a brief description of each dataset:

3.1 The Wisconsin Breast Cancer Dataset (WBCD)

This dataset [36] was obtained from Dr. William H. Wolberg of the University of Wisconsin Hospitals, Madison, Wisconsin, USA. The set includes nine features of ordinal variables having integer values in the range of 1 to 10 that describe visually assessed characteristics of fine needle aspiration (FNA) samples. The feature names are listed in the second column of Table 2. The feature number used in the table is the column number for the feature in the dataset after the column containing the sample code number in the original dataset was removed. A binary-valued class

variable indicates diagnosis as malignant (1) or benign (0). A classifier constructed using the multi-surface method (MSM) of pattern separation successfully diagnosed 97% of new cases [36]. 10-fold cross-validation average classification accuracies reported in the literature for a single classifier are 96.9% and 94.7% using backpropagation neural networks and the C4.5 decision tree tool, respectively [37].

3.2 The Cleveland Heart Disease Dataset

This dataset [38] is based on data from the Cleveland Clinic Foundation and consists of 270 records, each having 13 input features (a subset of an original set of 75 features). Brief feature description is shown in the third column of Table 2. A binary-valued class variable indicates the presence (1) or absence (0) of heart disease. 10-fold cross-validation average classification accuracies reported in the literature for a single classifier are 81.8% and 77.1% using backpropagation neural networks and the C4.5 decision tree tool, respectively [37]. With the large number of features relative to the size of the training set, this dataset is suitable for demonstrating benefits of dimensionality reduction brought about by feature selection.

4. Results

4.1 The Breast Cancer Data

Feature ranking was carried out using the training set of 483 cases with all nine features available as inputs. Training was performed in three steps of increasing model complexity corresponding to CPM = 2.5, 2, and 1.5. Models synthesized at these complexity levels are shown in Fig. 2. The left hand half of Table 3 lists features selected for each model, indicating the new group of three features selected by the learning algorithm at each stage. The right hand half of the table shows results for the additional steps taken to rank the features within each of the three 3-feature groups. Two steps are used to determine the ranks for the first and the second features of

each group, with the remaining feature considered third by default. Reading ranked features in this order from the last two columns in Table 3 gives the following ranking for the feature set: {2,6,7,8,1,5,3,4,9}, with feature 2 (Uniformity of cell size) being the best feature, followed by feature 6 (Bare nuclei) and feature 7 (Bland chromatin). Rough set data analysis of the dataset [39] reveals that Uniformity of cell size has a high classification quality and that Bare nuclei with Bland chromatin can account for 100% of the cases considered. Table 4 compares the feature ranking obtained with results reported in the literature using three other feature ranking algorithms [40]. All four methods listed unanimously select feature 9 (Mitoses) as the poorest predictor. The GMDH method alone selects feature 2 as the best feature, as opposed to feature 6 selected by all remaining methods. However, the GMDH method agrees with each of the three other methods in 2-3 out of the four highest ranking features.

Nine models were trained on the full training set with the default complexity penalty parameter, CPM = 1, using only k features; k = 1, 2, 3, …, 9 taken from the top of the ranking list. For example, the model having k = 3 uses the feature subset {2,6,7}. Each model was evaluated both on the training set and the evaluation set, and the resulting classification error rates are plotted in Fig. 3(a). As the number of features increases, the classification error rate monotonically decreases for the training set as the model more accurately fits the training data. The classification error rate on the out-of-sample evaluation set reaches a minimum of 2.5% at k = 7, 8 before it starts to rise as further increase in the number of features causes the model to overspecialize on the training data which affects its ability to generalizes well for the new data in the evaluation set. Therefore, the smallest optimal feature subset is {2,6,7,8,1,5,3} corresponding to k = 7. It is noted that the eighth feature {4} appears to be redundant to that subset, as it does not affect the classification error rate. Using the 7-feature optimum subset reduces the classification error rate from 4% with the full set of 9 features to 2.5%. Reducing the number of features used from 9 to 7

represents a 22% reduction in the size of the feature set used. Table 5(a) gives a detailed performance comparison between models trained on the full feature set and the optimal reduced subset, both at CPM = 1. Results indicate that dimensionality reduction leads to improvements in sensitivity, specificity, positive and negative predictive values, as well as overall classification accuracy. An increase of 3 percentage points is achieved in the positive predictive value.

As evident from Fig. 3(a) for CPM = 1, signs of overfitting occur quite late, with the classification error rate on the evaluation set leveling off at $k = 7$ and starting to increase at $k = 9$. This is due to the relatively large number of training examples compared to the number of features used. If more complex models were to be synthesized, overfitting would occur earlier, and smaller optimum feature subsets would be obtained. To verify this, the procedure used to obtain the plot in Fig. 3(a) was repeated with all models synthesized at the lower value of CPM = 0.5 for the complexity penalty multiplier, which produces more complex models, and the results are plotted in Fig. 3(b). Overfitting now occurs at $k = 5$, and the optimum feature subset at the new level of model complexity contains only four features {2,6,7,8}. Table 5(b) gives a detailed performance comparison between models trained on the full feature set and the optimal reduced subset, both at CPM = 0.5, showing improvements in sensitivity, specificity, positive and negative predictive values, as well as overall classification accuracy. Feature reduction reduces the classification error rate from 3.5% to 2.5% and increases the positive predictive value by 3 percentage points.

The ROC Characteristics [41] were used to compare the performance of the model using the optimum feature subset with that of the model using the full feature set, as well as models developed using reduced feature subsets obtained by different feature selection methods. The ROC curve is a plot of the sensitivity (true positive rate) versus the false positive rate (= 1 – specificity) for various values of the threshold used to sort a continuous classifier output into normal or abnormal classes. The area under the ROC curve (AUC) is a useful measure for determining the

quality of classification schemes and diagnostic tests, and statistically comparing their performance. This parameter is ideally 1.0 for an ideal classifier which has an ROC curve that passes through the point (0,1), thus giving 100% sensitivity at 100% specificity. Practically useful classifiers would have AUC values in the range ($0.5 < AUC \leq 1.0$). We used the Analyse-it statistical software package [42] which employs the Hanley and McNeil method [43] for performing the ROC analysis. Fig. 4(a) shows a plot of the two ROC curves as well as values of the AUC parameter and its standard error (SE) for two abductive network classifiers at CPM = 0.5, one trained on the highest ranking four features as determined above by the GMDH-based method, and the other trained on the full set of nine features. Results indicate that feature reduction does not lead to any loss in the area under the ROC curve. Fig. 4(b) compares the ROC curves and the AUC parameter for two abductive network classifiers developed by training on the highest ranking four features as determined by the GMDH method and the discriminant analysis (DA) feature selection method (method 2 in Table 4). Both models were trained with CPM = 0.5. The results indicate very similar ROC characteristics. The AUC parameter is slightly larger for the model using GMDH-selected features, but the difference is not statistically significant.

It is expected that feature ranking and minimization results would generally be unique to the learning algorithm used to derive them. However, we have investigated if the GMDH-based results on optimum feature subsets would be applicable to other learning paradigms, e.g. neural networks. The Pathfinder neural network software for Windows [44] was used to develop multilayer perceptron networks trained by error back propagation using both the full feature set and the two optimal subsets derived above for the breast cancer data. The networks had one hidden layer of neurons using the sigmoid transfer function and were trained and evaluated using the same data used to develop the corresponding abductive network models, with 20% of the training data reserved for cross validation. The number of neurons in the hidden layer was progressively reduced

to match the reduction in the number of model inputs used. Table 6 compares the performance of a model developed using all 9 features with that of two models using the two optimum feature subsets determined by the GMDH approach described above. Both the 7-feature and the 4-feature subsets reduce the classification error rate by approximately 29% compared to the full-feature set. Improvements of up to 3 percentage points are achieved in sensitivity and the positive predictive value. This indicates that feature ranking and selection results obtained using the proposed GMDH-based procedure may also prove useful with other learning algorithms.

4.2 The Heart Disease Data

Feature ranking was carried out using the training set of 190 cases with all 13 features available as inputs. Training was performed in four steps of increasing model complexity corresponding to CPM = 4, 2, 1.5, and 1. The left hand half of Table 7 lists features selected for each model, indicating the new group of three features selected at each stage. The right hand half of the table shows results for the additional steps taken to rank the features within each feature group. Reading features from the last two columns in Table 7 in the order they were selected gives the following ranking: {13,12,9,3,2,10,8,4,5,11,1,7,6}, with feature 13 being the highest predictive feature. Referring to Table 2, the highest ranking three features (13, 12, and 9) correspond to: Exercise induced angina (EXANG), Number of major vessels colored by fluoroscopy (CA), and Thal, respectively. Duch, Adamczak, and Grabczewski [45] derive the following rule as one of two classification rules that describe the dataset:

R_1: CA = 0 AND (Thal = 0 OR EXANG = 0) (6)

Table 8 compares the GMDH-based feature ranking obtained with results reported in the literature using three other feature ranking/selection algorithms. The conditional probabilities (CP) feature ranking algorithm [46] (method 2 in the table) ranks only the best eight of the 13 features. Method 3 [45] and method 4 [47] are feature selection algorithms, and the selected features were listed in the

table in the order they were reported in the referenced work, which may not reflect exact ranking. According to the given listing, all four methods unanimously select features 13 and 12 as the most important feautures. There is 87.5% agreement (7 out of 8) on the composition of the subset containing the highest ranking eight features as determined by the GMDH and CP feature ranking approaches.

Thirteen models were trained on the full training set with the default complexity penalty parameter, CPM = 1, using only k features; k = 1, 2, 3, …, 13 taken from the top of the ranking list. For example, the model having k = 3 uses the feature subset {13,12,9}. Each model was evaluated both on the training set and the evaluation set, and the resulting classification error rates are plotted in Fig. 5. As the number of features used increases, the classification error rate decreases for the training set as the model more accurately fits the training data. Classification error rate on the evaluation set reaches a global minimum of 15% at k = 6, 7. It then starts to rise as further increase in the number of features causes the model to overspecialize on the training data which affects its ability to generalizes well with new data in the evaluation set. Therefore, the smallest optimal feature subset is {13,12,9,3,2,10} corresponding to k = 6. Using the 6-feature optimum subset reduces the classification error rate from 17.5% with the full set of 13 features to 15%. Lowering the number of features used from 13 to 6 represents a 54% reduction in the size of the feature set. Table 9 gives a detailed performance comparison between abductive models trained on the full feature set and the optimal reduced subset at CPM = 1, showing improvements in sensitivity, specificity, positive and negative predictive values, as well as overall classification accuracy. Increases of approximately 6 and 3 percentage points are achieved in the sensitivity and the positive predictive value, respectively. Comparison of Fig. 5 with Fig. 3(a) for the breast cancer data at the same level of model complexity (CPM = 1) indicates that overfitting occurs earlier with the heart

dataset because of the lower ratio of the number of training examples to the number of features, and therefore greater susceptibility to high data dimensionality problems.

Fig. 6(a) shows a plot of the two ROC curves as well as values for the AUC parameter and its standard error (SE) for two abductive network classifiers at CPM = 1, one trained on the highest ranking six features as determined above using the GMDH method, and the other trained on the full set of 13 features. Results indicate that feature reduction does not lead to any loss in the area under the ROC curve. The AUC is slightly higher for the reduced feature subset, but the increase is not statistically significant. Fig. 6(b) compares the ROC curves and the AUC parameter for two abductive network classifiers developed by training at CPM = 1 on the highest ranking six features as determined through ranking by the GMDH method and the CP algorithm (method 2 in Table 8). The results indicate that the ROC curve for the GMDH-based ranking lies generally above the curve for the CP-based ranking and therefore is closer to that of the ideal classifier, which indicates better feature ranking by the GMDH method. The AUC parameter is slightly larger for GMDH-based ranking, but again the increase is not statistacilly significant.

The Pathfinder neural network software was used to develop multilayer perceptron networks trained by error back propagation using both the full feature set and the optimal subsets derived above for the heart disease data. The networks had one hidden layer of neurons using the sigmoid transfer function and were trained and evaluated using the same data used to develop the corresponding abductive network models, with 20% of the training data reserved for cross validation. Table 10 compares the performance of the neural network model developed using all 13 features of the data with that developed using the optimum 6-feature subset determined by the GMDH-based ranking approach. The 6-feature model gives the same overall classification accuracy

as the full-feature model. As the former requires less than half the number of input features required by the latter, models using the reduced feature set will still be more efficient.

5. Conclusions

Automatic input selection by GMDH type learning algorithms can be utilized for feature ranking and subsequent selection of optimum feature subsets for improved implementation and performance of classifiers for medical screening and diagnosis. Feature ranking according to the predictive quality should also be of interest to medical practitioners as it provides insight into the diagnostic value of various disease markers collected. Feature reduction is particularly useful with high-dimensional data characterized by a large number of features and a relatively few training examples, which is the case in many medical applications. We have described a 2-stage hierarchical approach to perform complete ranking of individual features. In the first stage, features are ranked in groups by the order they are selected in by a GMDH type learning algorithm as the complexity level specified for the model is gradually increased. Features within each group are then ranked by repeating the procedure with only the features within the group used as model inputs. The feature ranking list thus obtained can be used to determine the contents of an optimum feature subset that minimizes classification error rate on a dedicated evaluation set. Feature ranking results are comparable with those reported in the literature using other techniques. With the heart disease dataset, an optimal subset giving 54% feature reduction improves the overall classification accuracy from 82.5% to 85%. Larger improvements may be possible with other datasets. GMDH-based ranking compares favorably with that by other techniques reported in the literature, and ROC curves for resulting optimum classifiers more closely approach that of an ideal classifier. We have demonstrated that feature reduction results obtained from this GMDH-based approach could be applied to other learning algorithms. For example, the optimal feature subset giving 56% feature

reduction with the breast cancer data improves the classification accuracy of a neural network classifier from 96.5% to 97.5% while increasing sensitivity from 94% to 97%. Future work would explore applying the feature reduction results to other learning algorithms and using the technique with other medical datasets.

Acknowledgement

The author wishes to acknowledge the support of the Physics Department and the Research Committee at King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.

Appendix A

List of data rows used for model evaluation (remaining rows used for training)

a. Breast Cancer Data: (200 records)

Note: Numbering was done after deleting the 16 rows containing missing attributes.

4,5,6,7,8,9,11,14,16,19,21,23,24,25,26,30,32,36,38,39,40,41,42,43,45,46,48,51,54,59,65,67,70,73
,74,76,78,79,80,81,84,88,89,92,94,95,97,98,100,101,102,103,105,107,108,109,112,113,120,123
,124,127,129,138,140,142,143,144,146,147,148,149,151,152,153,154,155,156,163,165,166,169
,171,173,174,176,177,178,187,188,189,191,194,196,197,198,201,204,205,207,208,210,213,217
,220,222,227,233,238,241,245,246,248,251,256,261,269,271,272,278,280,281,285,291,306,312
,313,314,316,318,319,321,322,324,325,326,329,332,333,342,345,353,357,359,362,367,368,370
,375,386,387,391,394,396,408,410,415,417,419,434,441,449,457,460,464,465,467,469,470,500
,503,506,523,524,527,529,535,537,546,554,558,570,571,576,583,593,594,602,608,609,611,615
,616,619,626,633,636,654,667,681

b. Heart Disease Data (80 records):

3,5,10,11,17,18,19,20,21,24,25,31,34,37,40,50,58,59,60,62,63,64,66,67,68,69,70,71,72,73,75,80,81
,84,87,90,95,96,98,99,103,106,110,112,113,114,119,129,141,142,150,156,157,161,165,166,171
,177,185,192,193,201,202,203,206,208,212,222,225,226,227,228,235,236,240,248,249,256,264,266

References

[1] Montgomery GJ, Drake KC. Abductive networks. Proceedings of the SPIE Conference on the Applications of Artificial Neural Networks, 1990, 56-64.

[2] Farlow SJ. The GMDH algorithm. In: Farlow SJ, ed. Self-Organizing Methods in Modeling: GMDH Type Algorithms. New York: Marcel-Dekker, 1984:1-24.

[3] Abdel-Aal RE, Mangoud AM. Modeling obesity using abductive networks. Comput Biomed. Res. 1997;30:451-71.

[4] Abdel-Aal RE, Mangoud AM. Abductive machine learning for modeling and predicting the educational score in school health surveys. Methods Inf Med 1996;35:265-71.

[5] Echauz J, Vachtsevanos G. Neural network detection of antiepileptic drugs from a single EEG trace**.** Proceedings of the Electro/94 International Conference, 1994, 346-51.

[6] Kondo T, Pandya AS, Zurada JM. GMDH-type neural networks and their application to the medical image recognition of the lungs**.** Proceedings of the 38th IEEE SICE Annual Conference, 1999, 1181-6.

[7] Cheung J, Lin ZY, McCallum RW, Chen JDZ. Screening of delayed gastric emptying using electrogastrography and abductive networks**.** Gastroenterology Suppl. S 1997;112:A711.

[8] Yu S, De Backer S, Scheunders P. Genetic feature selection combined with composite fuzzy nearest neighbor classifiers for high-dimensional remote sensing data**.** IEEE International Conference on Systems, Man, and Cybernetics, 2000, 1912-6.

[9] Hoffman AJ, Hoogenboezem C, van der Merwe NT, Tollig CJA. Seismic buffer recognition using mutual information for selecting wavelet based features. IEEE International Symposium on Industrial Electronics, 1998, 663-7.

[10] Abdulla WH, Kasabov N. Reduced feature-set based parallel CHMM speech recognition systems**.** Information Sciences 2003;156:21-38.

[11] Ozdemir M, Embrechts MJ, Arciniegas F, Breneman CM, Lockwood L, Bennett KP. Feature selection for in-silico drug design using genetic algorithms and neural networks**.** IEEE Mountain Workshop on Soft Computing in Industrial Applications, 2001, 53-7.

[12] Matsui K, Kosugi Y. Image segmentation by neural-net classifiers with genetic selection of feature indices. IEEE International Conference on Image Processing, 1999, 524–8.

[13] Kira K, Rendell LA. A practical approach to feature selection. Proceedings of the Ninth International Conference on Machine Learning, 1992, pp. 249–56.

[14] Lewin DI. Getting clinical about neural networks. IEEE Intelligent Systems 2000;15:2-3.

[15] Abbott DW. A two-stage approach to feature downselection for pattern recognition IEEE International Conference on Systems, Man and Cybernetics, 1995, 1527–32.

[16] Garrett D, Peterson DA, Anderson CW, Thaut MH. Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering 2003;11:141–4.

[17] Gletsos M, Mougiakakou SG, Matsopoulos GK, Nikita KS, Nikita AS, Kelekis D. A computer-aided diagnostic system to characterize CT focal liver lesions: design and optimization of a neural network classifier. IEEE Transactions on Information Technology in Biomedicine 2003; 7:153–62.

[18] Kupinski MA, Giger ML. Feature selection and classifiers for the computerized detection of mass lesions in digital mammography. International Conference on Neural Networks, 1997, 2460-3.

[19] McNitt-Gray MF, Huang HK, Sayre JW. Feature selection in the pattern classification problem of digital chest radiograph segmentation. IEEE Transactions on Medical Imaging 1995;14:537–47.

[20] de Chazal P, Heneghan C, Sheridan E, Reilly R, Nolan P, O'Malley M. Automated processing of the single-lead electrocardiogram for the detection of obstructive sleep apnoea**.** IEEE Transactions on Biomedical Engineering 2003;50:686–96.

[21] Lee W-L, Chen Y-C, Hsieh K-S. Ultrasonic liver tissues classification by fractal feature vector based on M-band wavelet transform. IEEE Transactions on Medical Imaging 2003; 22 :382–92

[22] Zarjam P, Mesbah M, Boashash B. An optimal feature set for seizure detection systems for newborn EEG signals**.** International Symposium on Circuits and Systems, 2003, V-33-36.

[23] Blum A, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence 1997;97:245-71.

[24] Hall MA. Correlation-based feature selection for discrete and numeric class machine learning. Proceedings of the 17th International Conference on Machine Learning, Stanford University, CA. Morgan Kaufmann Publishers, 2000.

[25] Battiti R**.** Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks 1994;5:537-50

[26] Kittler J. Feature selection and extraction. In: Young TY and Fu KS, eds. Handbook of Pattern Recognition and Image Processing. San Diego, CA: Academic, 1986:59–83.

[27] Kohavi R, John GH. Wrappers for feature subset election. Artificial Intelligence 1997;7:273-323.

[28] Brill FZ, Brown DE, Martin WN. Fast generic selection of features for neural network classifiers. IEEE Transactions on Neural Networks 1992;3:324-8.

[29] Aha DW, Bankert RL. A comparative evaluation of sequential feature selection algorithms. Learning from Data: AI and Statistics V, Springler-Verlag, 1996.

[30] Swiniarski RW, Skowron A. Rough set methods in feature selection and recognition Pattern Recognition Letters 2003;24:833-49.

[31] Pachepsky YA, Rawls WJ. Accuracy and reliability of pedotransfer functions as affected by grouping Soils**.** Soil Sci. Soc. Am. J. 1999;63:1748–57.

[32] Abdel-Aal RE. Improving classifier performance using abductive network committees trained on different feature subsets. Submitted to Computer Methods and Programs in Medicine.

[33] http://www.ics.uci.edu/~mlearn/MLRepository.html

[34] AbTech Corporation, Charlottesville, VA, AIM User's Manual, 1990.

[35] Barron AR. Predicted squared error- a criterion for automatic model selection. In: Farlow SJ, ed. *Self-Organizing Methods in Modeling: GMDH Type Algorithms*. New York: Marcel-Dekker, 1984:87-103.

[36] Wolberg WH, Mangasarian OL. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceedings of the USA National Academy of Sciences 1990;87:9193-6.

[37] Opitz DW, Maclin RF. An empirical evaluation of bagging and boosting for artificial neural networks. International Conference on Neural Networks, 1997, 1400–5.

[38] Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid J, Sandhu S, Guppy K, Lee S, Froelicher V. International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology 1989;64:304-10.

[39] Hassanein AE, Ali JMH. Rough set approach for generation of classification rules of breast cancer data. Informatica 2004;15:23-38.

[40] Verikas A, Bacauskiene M. Feature selection with neural networks. Pattern Recognition Letters 2002;23:1323-35.

[41] DeLong ER, DeLong DM, Clarke-Pearson DL, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44:837-45.

[42] Analyse-it Software Ltd, PO Box 77, Leeds, LS12 5XA, UK.

[43] Hanley JA, McNeil BJA, Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases. Radiology 1983;148:839-43.

[44] Z Solutions, Inc., 6595G Roswell Road, Suite 662, Atlanta, Georgia 30328, USA.

http://www.zsolutions.com/index.htm

[45] Duch W, Adamczak R and Grabczewski K. A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. IEEE Transactions on Neural Networks 2001;12:277-306.

[46] Ahmad A, Dey L. A feature selection technique for classificatory analysis. Pattern Recognition Letters 2005;26:43-56.

[47] Duch W, Grudziński K. Search and global minimization in similarity-based methods. International Joint Conference on Neural Networks, IJCNN'99, Washington, 1999.

Table 1. Summary statistics for the two datasets.

Dataset	Number of Features	Whole Dataset		Training Set		Evaluation Set	
		Number of Cases	Prevalence, %	Number of Cases	Prevalence, %	Number of Cases	Prevalence, %
Breast	9	683	35	483	35.6	200	33.5
Heart	13	270	44.4	190	44.7	80	43.8

Table 2. Brief description of the features in the two datasets.

Feature Number in Dataset	Feature Description	
	Breast Cancer Dataset	Heart Disease Dataset
1	Clump thickness	Age
2	Uniformity of cell size	Sex
3	Uniformity of cell shape	Chest pain type (4 values)
4	Marginal adhesion	Resting blood pressure
5	Single epithelial cell size	Serum cholesterol in mg/dl
6	Bare nuclei	Fasting blood sugar > 120 mg/dl
7	Bland chromatin	Resting electrocardiographic results (values: 0,1,2)
8	Normal nucleoli	Maximum heart rate achieved
9	Mitoses	Exercise induced angina (EXANG)
10		Oldpeak = ST depression induced by exercise relative to rest
11		Slope of the peak exercise ST segment
12		Number of major vessels (0-3) colored by fluoroscopy (CA)
13		Thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

Table 3. Feature ranking results for the breast cancer data. Left hand side: Ranking of feature groups. Right hand side: Ranking of individual features within groups.

Step No.	Features Used as Inputs	CPM	Feature Groups Automatically Selected			Step No.	Feature Group Used as Inputs	CPM	Individual Features Automatically Selected	
1	1, 2, 3, 4, 5, 6, 7, 8, 9 (All features)	2.5	2, 6, 7			4	2, 6, 7	10	2	
						5		5	2	6
2		2	2, 6, 7	1, 5, 8		6	1, 5, 8	20	8	
						7		10	8	1
3		1.5	2, 6, 7	1, 5, 8	3, 4, 9	8	3, 4, 9	5	3	
						9		2.5	3	4

Table. 4. Comparison of feature ranking for the breast cancer data by the proposed method and three other feature ranking methods [Ref. 40]. Features are listed highest quality first.

Number	Method	Description	Feature Ranking
1	GMDH	Proposed Method	{2,6,7,8,1,5,3,4,9}
2	DA	Filter method based on discriminant analysis	{6,3,2,7,1,8,5,4,9}
3	SNR	Neural network weights method based on signal-to-noise ratio	{6,1,3,2,7,8,4,5,9}
4	FQI	Neural network output method based on a feature quality index	{6,1,8,3,4,7,5,2,9}

Table 5. Performance comparison between two abductive models for the breast cancer data: one using the full feature set and the other using an optimum feature subset determined by the GMDH-based approach. (a) at CPM =1 and (b) at CPM = 0.5.

(a)

Model (CPM = 1)	Sensitivity, %	Specificity, %	Positive Predictive Value, %	Negative Predictive Value, %	Overall Classification Accuracy, %
Using all 9 features	92.5	97.7	95.4	96.3	96
Using optimum 7-feature subset {2,6,7,8,1,5,3}	94.0	99.2	98.4	97.1	97.5

(b)

Model: (CPM = 0.5)	Sensitivity, %	Specificity, %	Positive Predictive Value, %	Negative Predictive Value, %	Overall Classification Accuracy, %
Using all 9 features	96.9	96.3	92.5	98.5	96.5
Using optimum 4-feature subset {2,6,7,8}	97.0	97.8	95.5	98.5	97.5

Table 6. Performance comparison between a neural network model trained on all nine features of the breast cancer data and two neural models trained on optimum feature subsets determined by GMDH-based ranking and selection. Classification performance is improved with up to 56% feature reduction.

Features Used	Number of Hidden Neurons	Sensitivity, %	Specificity, %	Positive Predictive Value, %	Negative Predictive Value, %	Overall Classification Accuracy, %
All 9 features	6	94.0	97.7	95.4	97.0	96.5
Optimum 7-feature subset {2,6,7,8,1,5,3}	5	94.0	99.2	98.4	97.0	97.5
Optimum 4-feature subset {2,6,7,8}	4	97.0	97.8	95.6	98.5	97.5

Table 7. Feature ranking results for the heart disease data. Left hand side: Ranking of feature groups. Right hand side: Ranking of individual features within groups.

Step No.	Features Used as Inputs	CPM	Feature Groups Automatically Selected				Step No.	Feature Group Used as Inputs	CPM	Individual Features Automatically Selected	
1	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 (All features)	4	9, 12, 13				5	9, 12, 13	10	13	
							6		5	13	12
2		2	9, 12, 13	3, 2, 10			7	3, 2, 10	10	3	
							8		6	3	2
3		1.5	9, 12, 13	3, 2, 10	4, 8, 5		9	4, 8, 5	5	8	
							10		2	8	4
4		1	9, 12, 13	3, 2, 10	4, 8, 5	11,6,7,1	11	11,6,7,1	10	11	
							12		4	11	1
							13	6,7	1	7	

Table 8. Comparison of feature ranking for the heart disease data by the proposed method and three other feature ranking/selection methods. Ranked features are listed highest quality first.
* = Sequence may not represent exact ranking.

Number	Method	Description	Feature Ranking/Selection
1	GMDH	Proposed Method	{13,12,9,3,2,10,8,4,5,11,1,7,6}
2	CP [Ref. 46]	Feature ranking method based on conditional probabilities	{13,12,3,9,11,10,8,2}
3	LR [Ref. 45]	Logic rules extraction algorithm for feature selection	{13,12,3,9,11}*
4	SB [Ref. 47]	Similarity-based, feature-dropping, selection algorithm	{13,12,3}*

Table 9. Performance comparison between two abductive models for the heart disease data: one using the full feature set and the other using an optimum subset of the highest ranking six features as determined by the GMDH-based approach. CPM =1.

Model: (CPM = 1)	Sensitivity, %	Specificity, %	Positive Predictive Value, %	Negative Predictive Value, %	Overall Classification Accuracy, %
Using all 13 features	71.4	91.1	86.2	80.4	82.5
Using optimum 6-feature subset {13,12,9,3,2,10}	77.1	91.1	87.1	83.7	85

Table 10. Performance comparison between two neural network models for the heart disease data: one using the full feature set and the other using the optimum subset of the highest ranking six features as determined by the GMDH-based approach. The 54% reduction in the size of the feature set does not degrade overall classification accuracy.

Features Used	Number of Hidden Neurons	Sensitivity, %	Specificity, %	Positive Predictive Value, %	Negative Predictive Value, %	Overall Classification Accuracy, %
All 13 features	8	82.9	84.4	80.6	86.4	83.75
Optimum 6-feature subset {13,12,9,3,2,10}	6	80	86.7	82.4	84.8	83.75

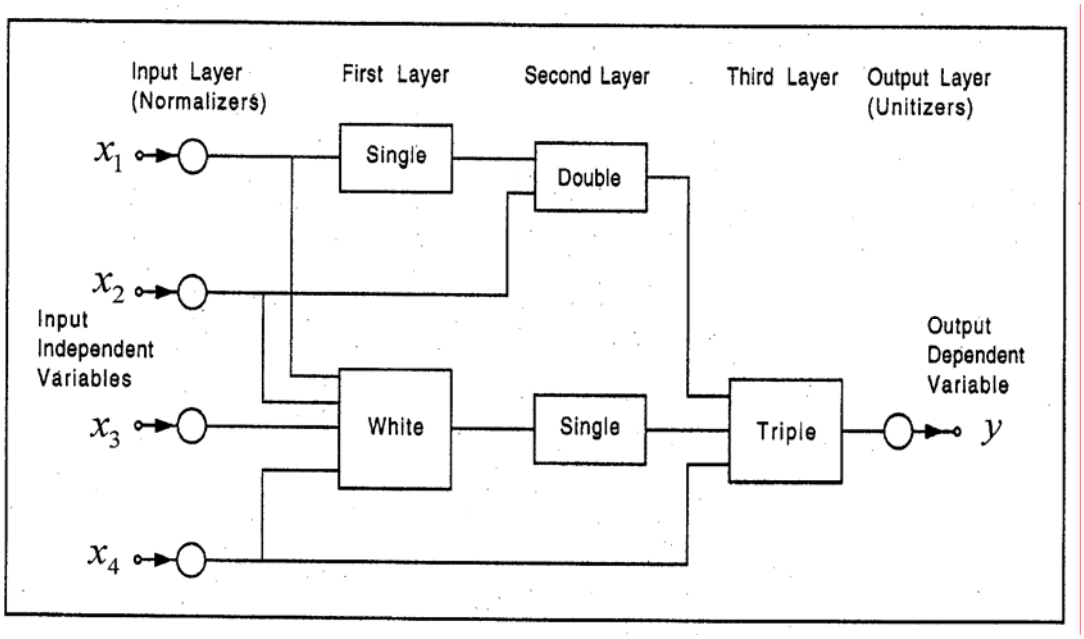


Fig. 1. AIM abductive network showing various types of functional elements.

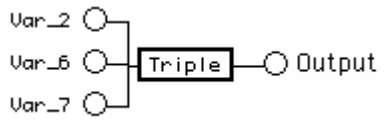
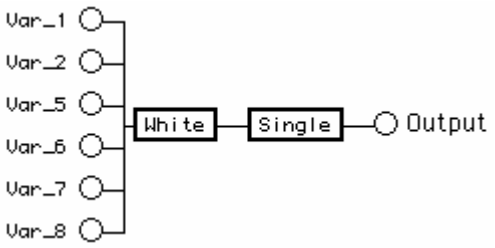
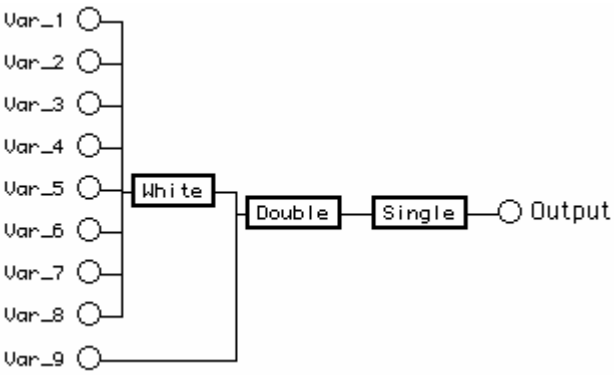
Step	CPM	Resulting Model Structure
1	2.5	
2	2	
3	1.5	

Fig. 2. Models synthesized at three levels of increasing model complexity for the breast cancer data. Numbers at input nodes refer to features automatically selected by the learning algorithm.

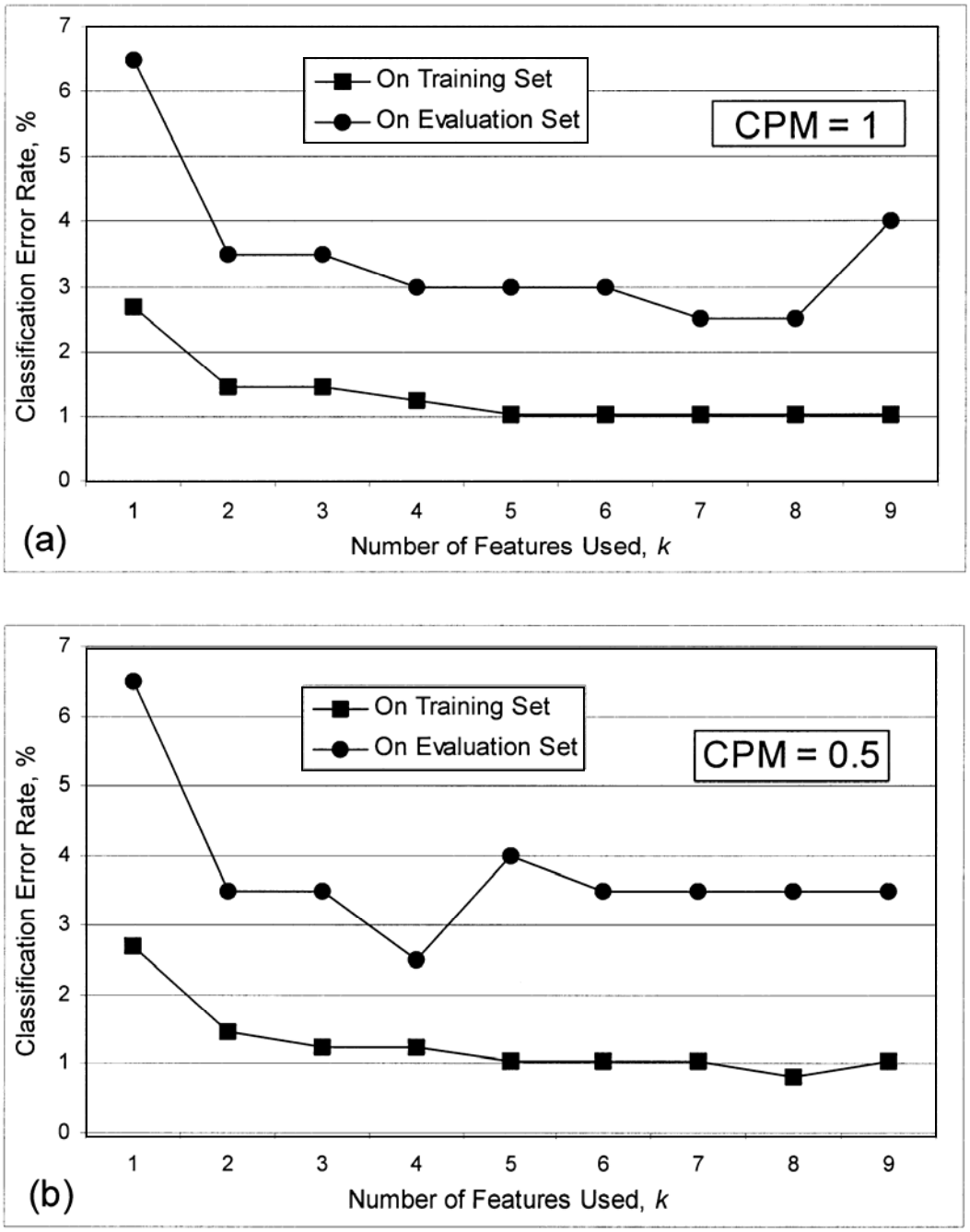


Fig. 3. Plots of the classification error rates on both the training and evaluation sets of the breast cancer data for nine models trained on k features; $k = 1, 2,\ldots,9$ taken from the top of the feature ranking list. (a): With the default model complexity (CPM = 1), (b): With larger model complexity (CPM = 0.5).

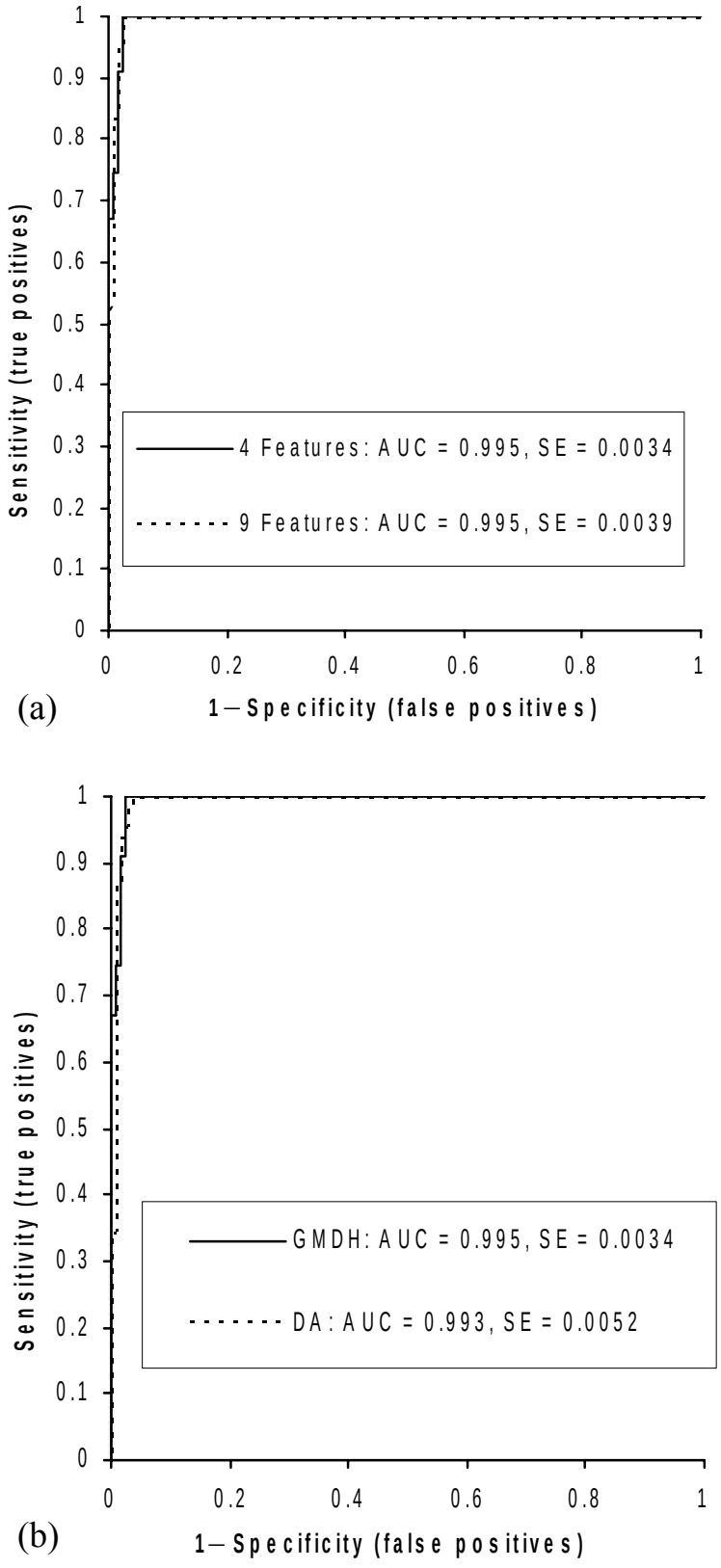


Fig. 4. Comparison of the ROC characteristics for the optimum abductive model using the highest ranking four features for the breast cancer data as determined by the GMDH-based approach and (a): An abductive model using all nine features, (b): An abductive model using the highest ranking four features as determined by a filter feature ranking method based on discrimination analysis (DA) [Ref. 40]. CPM = 0.5.

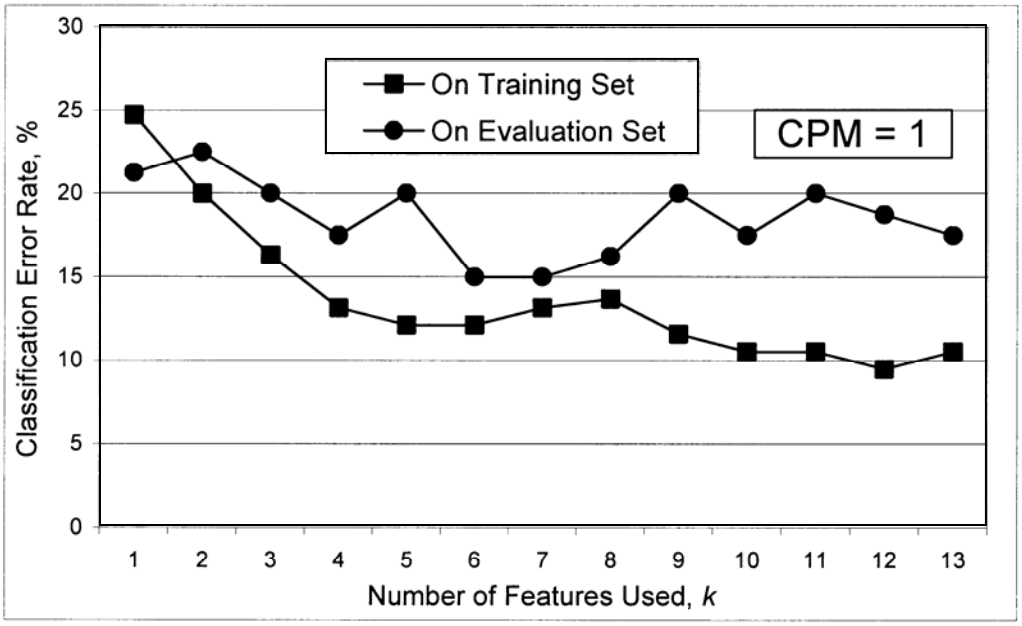
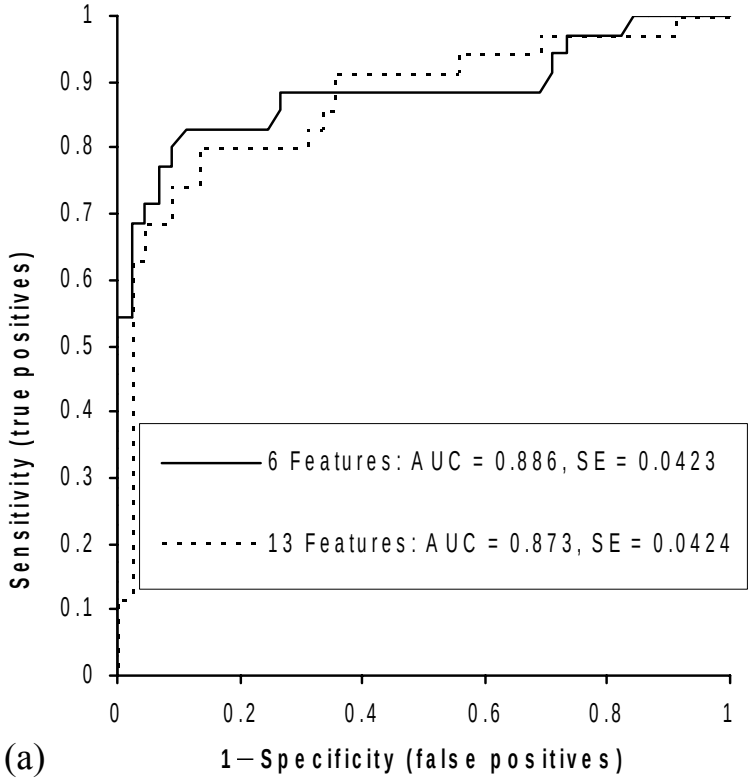
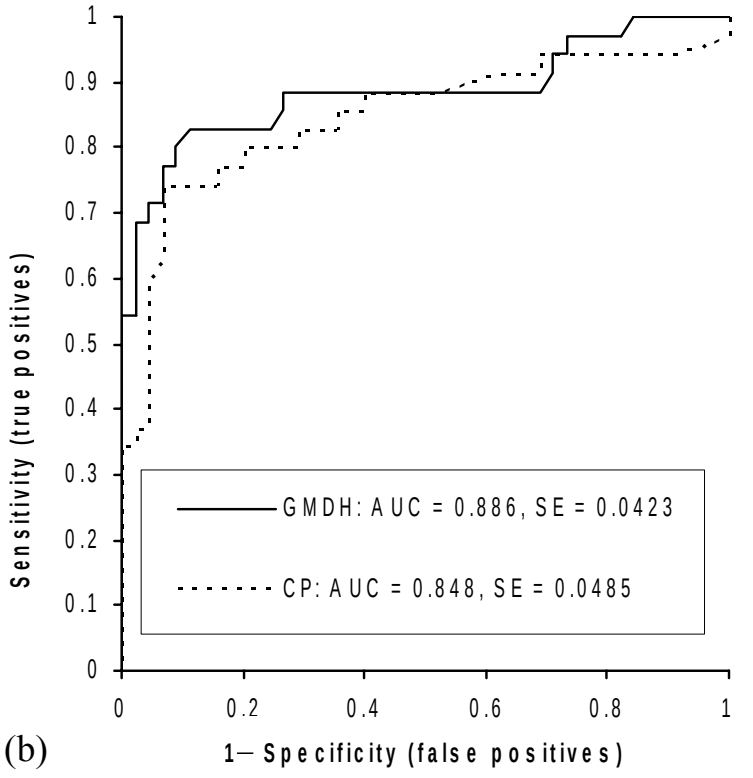


Fig. 5. Plots of the classification error rates on both the training and evaluation sets of the heart disease data for thirteen models trained on k features; k = 1, 2, …,13 taken from the top of the feature ranking list. CPM = 1.



(a)



(b)

Fig. 6. Comparison of the ROC characteristics for the optimum abductive model using the highest ranking six features for the heart disease data as determined by the GMDH-based approach and (a): An abductive model using all 13 features, (b): An abductive model using the highest ranking six features as determined by the conditional probabilities (CP) feature ranking method [Ref. 46]. CPM = 1.