Improved Classification of Medical Data Using Abductive Network Committees Trained on Different Feature Subsets

R. E. Abdel-Aal
Department of Computer Engineering,
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

Summary

This paper demonstrates the use of abductive network classifier committees trained on different features for improving classification accuracy in medical diagnosis. In an earlier publication, committee members were trained on different subsets of the training set to ensure enough diversity for improved committee performance. In situations characterized by high data dimensionality, i.e. a large number of features and a relatively few training examples, it may be more advantageous to split the feature set rather than the training set. We describe a novel approach for tentatively ranking the features and forming subsets of uniform predictive quality for training individual members. The abductive network training algorithm is used to select optimum predictors from the feature set at various levels of model complexity specified by the user. Using the resulting tentative ranking, the features are grouped into mutually exclusive subsets of approximately equal predictive power for training the members. The approach is demonstrated on three standard medical diagnosis datasets (Breast Cancer, Heart Disease, and Diabetes). Three-member committees trained on different feature subsets and using simple output combination methods reduce classification errors by up to 20% compared to the best single model developed with the full feature set. Results are compared with those reported previously with members trained through splitting the training set. Training abductive committee members on feature subsets of approximately equal predictive power achieves both diversity and quality for improved committee performance. Ensemble feature subset selection can be performed using GMDH-based learning algorithms. The approach should be advantageous in situations characterized by high data dimensionality.

Keywords:

Abductive networks, neural networks, network ensemble, network committee, feature selection, classification accuracy, medical diagnosis, breast cancer, heart disease, diabetes.

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 and improved diagnosis, the concept of committee (ensemble) classifiers has been adopted in medicine, e.g. [8-11]. With this approach, a number of committee members (base classifiers) are trained on different aspects of the problem and then interrogated simultaneously, with their outputs combined to produce the final predicted committee output. Simple methods of combining the outputs of individual members, such as majority voting or weighted averaging, often lead to useful gains in classification performance. When member classifiers are independent, the resulting diversity in the decision making process is expected to boost generalization performance, thus improving the accuracy, robustness, and reliability of classification. Individual member models are expected to be simpler, and therefore train and execute faster than a single monolithic model. Since the members train and execute

independently in parallel, this approach can also achieve a reduction in the overall training and classification times compared to the monolithic approach. Obviously, combining the outputs of several identical classifiers produces no gain, and improvement is expected only when members err in different ways so that errors may cancel out [12]. It can be shown [13] that the mean squared error in the averaged committee output contains as a component the covariance of the outputs of individual committee members, therefore individual members should ideally be uncorrelated or even negatively correlated. Krogh and Vedelsby [14] have shown that the committee error can be expressed as two terms, one measuring the average generalization error of individual members and the other measuring the diversity or disagreement among the members. An ideal committee would therefore consist of highly accurate classifiers that disagree as much as possible.

While neural network committees have been widely reported in the literature, there appears to be little mention of GMDH-based abductive (or polynomial) network committees. Due to the self-organizing and self-stopping nature of such networks, the absence of initial random weights and the little room for user intervention during training, there is less scope for introducing diversity in member models synthesized using the same training data. In [15] we have described abductive network committees that train on mutually exclusive subsets of the training set to achieve the diversity required for good ensemble performance. Many medical applications are characterized by a large number of features (attributes) and relatively few training examples. Splitting such a small training set may not provide adequate training for the committee's base classifiers. Moreover, using the full set of features to train individual members with the further reduced training sets may lead to overfitting due to the dimensionality problem. This paper proposes training individual members of the abductive network committee on different subsets of the input features using the full number of training examples available. This approach should improve the quality of the base classifiers and the committee performance in applications characterized by high dimensionality. Splitting the feature

set should also speed up the training of individual members as compared to splitting the training set. Classifier ensembles trained on different feature subsets have been used in medicine for classifying cancer using gene expression data derived from DNA arrays [16], diagnosing acute appendicitis with Bayesian classifiers [17], and diagnosing hearing impairment in children using the multi-dimensional voice profile data [18].

2. Background

In many medical applications, enough features may be available to allow training of all base classifiers on mutually exclusive feature subsets. However, the feature subsets used should be carefully selected to ensure good quality of the individual classifiers, as well a high degree of diversity and independence amongst them to encourage constructive disagreements that enhance ensemble performance [19]. Feature selection for this purpose is somewhat different from traditional feature selection used for data reduction in areas characterized by high dimensionality due to the large number of available features, e.g. in remote sensing [20], seismic data processing [21], speech recognition [22], and drug design [23]. In single classifier applications, dimensionality reduction attempts to select a smaller subset of optimum features by excluding irrelevant and redundant features in order to avoid overfitting, improve performance, and speedup both training and prediction for the classifier. This form of feature reduction has been applied to several areas in medicine, including: classification of EEG signals for operating brain-computer interfaces [24], detection of mass lesions in digital mammograms [25], segmenting digital chest radiographs [26], and detection of seizure events in newborn children using EEG data [27].

Techniques for feature subset selection can be classified into three main categories: embedded, filter, and wrapper techniques [28]. With embedded techniques, feature subset selection is part of the induction learning algorithm itself, as is the case with the CART classification tree. In filter techniques [29], subset selection is performed prior to induction and the selected subset serves as an

input to the induction algorithm. Wrapper techniques [30] search for an optimal subset of features by starting with an empty set and adding or removing features depending on the performance of the induction algorithm, which forms part of the feature search engine. Feature selection techniques based on the rough set theory have also been proposed [31]. The random subspace method (RSM) [32] was used to build decision tree ensemble classifiers based on different features. Tsymbal and Puuronen [17] use the RSM method to randomly select subsets of features, but follow this with a hill-climbing refinement search to optimize the subsets selected. Skrypnyk et. al. [18] use correlation-based feature selection [33] to select optimum subsets of features, each subset excelling in distinguishing one class of the population from the others. Each class is represented by an ensemble member trained on the relevant feature subset. Cho and Ryu [16] use a number of measures based on correlation, distance, information gain, mutual information, and signal-to-noise ratio to extract a set of features. Out of this set, correlation analysis is used to determine two negatively correlated features for training the base classifiers. Genetic algorithms have been used to search among a set of randomly selected feature subsets to maximize a fitness function based on accuracy and diversity for the resulting base classifiers [34]

This paper describes the development of abductive network classifier committees where members train on different feature subsets. A novel technique is presented for feature subset selection, which uses the GMDH learning algorithm [1,2] to automatically select optimum predictors [35] at various levels of model complexity specified by the user. Information collected in this way is used to tentatively rank the available features according to their predictive quality. This ranking is then used to assign features to individual committee members such that the overall quality of the features used is approximately the same for all members. This should avoid large variations in the classification accuracy amongst individual members, thus improving the ensemble performance. Training on mutually exclusive feature subsets should ensure diversity, as each subset represents a different

view of the problem space. The technique is demonstrated using three standard medical datasets from the UCI Machine Learning Repository [36]. Feature ranking according to quality gives insight into the most effective markers for the diagnosis problem. Section 3 gives a brief introduction to the GMDH algorithm and the abductive network modeling tool used. It describes the approaches adopted for constructing the abductive network committees and selecting the feature subsets. Section 4 gives a brief outline of the three medical datasets used in the investigation. Section 5 presents the results obtained. In all cases, the resulting committees outperformed both the best monolithic model utilizing all features and the best committee member. Conclusions are made and suggestions for future work given in Section 6.

3. Computational Methods

3.1 GMDH and AIM Abductive Networks

AIM (abductory inductive mechanism) [37] 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]:

$$y = A + B\,x_i + C\,x_j + D\,x_i^2 + E\,x_j^2 + F\,x_i\,x_j \qquad (1)$$

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:

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

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 over-fitting 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 [38]. 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 [38] 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:

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

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* [38]. 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) [37], 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,

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

3.2 Abductive Network Committees

Fig. 2 is a schematic of the abductive network committees developed. All committees consisted of only three members (n = 3). The full set of m features available is split into three subsets of comparable predictive quality, each containing approximately m/3 features, using the feature selection technique described in Section 3.3 below. Each of the individual members is trained on the full training set available using only the designated feature subset. During classification, the trained member networks are simultaneously interrogated, each with the relevant subset of the input vector, and their continuous outputs $y_{i;}$ i = 1,2,3 are combined to generate the final predicted committee output z_c. The committees adopt cooperation schemes in the form of simple combination rules implemented by the output combination module in Fig. 2. Such rules include simple majority voting of categorical member outputs and simple averaging of continuous member outputs [15]. Categorical (classification) outputs are derived from corresponding continuous values by simple rounding (thresholding at 0.5).

3.3 Selection of the Feature Subsets

To ensure good classification performance by all committee members, the members need to train on feature subsets of comparable overall predictive quality. One way to achieve this is to assign features to the n members in groups of n features of similar quality, one feature for each member. The process is then repeated with the next group of n features, and so on until all features are assigned to the members. Fig. 3 shows a scheme where all available m features are arranged according to predictive quality, with feature 1 having the highest quality. Cho and Ryu [16] determine such ranking using performance scoring on a set of statistical, similarity, and information-theoretical measures. With n = 3 as depicted in the figure, the top-ranking three features (numbered 1, 2, and 3) are assigned to the committee members # 1, 2, and 3, respectively, with member # 1 receiving feature 1. The second best group of three features (numbered 4, 5, and 6) are assigned to members # 2, 3, and 1, respectively. The number for the committee member

receiving the best feature among the group of three features is shifted by one, in a modulo *n* manner, each time a new group is assigned. This ensures that member # 1 does not always get the first (and best) feature in each group. The assignment procedure continues until all features are exhausted.

In this paper we perform tentative ranking of the input features according to predictive quality by using the GMDH-based learning algorithm of the AIM tool to automatically select optimum predictors at various stages of model complexity. 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 of only three input features. 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 that selects in three more features that will have lower predictive quality than the first three. The process continues until all features are selected. In this way, features will be determined in groups of three. Features within each group will be arbitrarily assigned to the three committee members, since no information exists on the ranking of predictive quality within each group. Recommended values for the CPM parameter range from 10 for the simplest model down to 0.1 for the most complex model [37]. If the most complex model still leaves some features unselected, all features selected thus far can be disabled as inputs to enforce selection from the remaining features and allow completion of the feature ranking process.

3.4 Software and Hardware Implementations

Results reported in this paper were obtained using the AIM abductive network software version 1.0 for Macintosh computers [37]. A version of the software has later been developed for PCs running Windows [39]. Contrary to neural networks where hardware implementations have been available for some time in the form of VLSI chips from several vendors, there appears to be no integrated circuit implementations of abductive or polynomial networks available at present.

4. Data Sets for Training and Evaluation

Three standard medical diagnosis datasets from the UCI Machine Learning Repository [36] were used for this study. These include the Wisconsin breast cancer dataset, the Cleveland heart disease dataset, and the Pima Indian diabetes dataset. Out of the 699 cases for the breast cancer dataset, 16 records containing missing attributes were omitted, 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%. Table 1 lists important statistics on the datasets, including the percentage prevalence of positives in the total, training, and evaluation sets. 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:

4.1 The Wisconsin Breast Cancer Dataset (WBCD)

This dataset [40] 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 [40]. 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 [41].

4.2 The Cleveland Heart Disease Dataset

This dataset [42] 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 [41].

4.3 The Pima Indians Diabetes Dataset

This dataset [43] was donated by Vincent Sigillito of the Johns Hopkins University, Baltimore, Maryland, USA. It consists of 768 records of female patients at least 21 years old of Pima Indian heritage. There are eight numerical features representing physiological measurements and medical test results. Brief feature description is shown in the fourth column of Table 2. A binary-valued class variable indicates whether the patient shows signs of diabetes according to World Health Organization criteria (1) or not (0). This dataset is particularly difficult to classify, with 10-fold cross-validation average classification accuracies reported in the literature for single classifiers being 76.4% and 74.6% for backpropagation neural networks and the C4.5 decision trees tool, respectively [41].

5. Results

5.1 The Breast Cancer Data

Feature subset selection for a 3-member committee 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. 4. Table 3(a) lists features selected for each model, indicating the new

group of three features introduced at each stage. All the resulting models were evaluated on the same evaluation set of 200 cases, and the corresponding percentage classification errors are shown. The model at CPM = 2 is the optimum monolithic model encountered, with a classification error of 2.5%. Table 3(b) lists the feature subsets, each comprising three features, assigned to the three committee members. Each member was trained on the full training set at the CPM value given in the table. Shown also are the percentage classification errors obtained when the individual members were evaluated on the evaluation set. When the member outputs were combined, the committee achieved a classification error of 2% using either simple majority voting of categorical member outputs or simple averaging of continuous member outputs prior to thresholding. This value is 50% lower than that for the best committee member and 20% lower than that for the best monolithic module developed with the full feature set. 10-fold cross validation classification error rates on the same dataset for 10-member neural networks using the bagging and boosting resampling techniques are 3.3% and 3.9%, respectively [41]. Table 3(c) gives a detailed performance comparison between that committee and the default monolithic output (CPM = 1), showing improvements in overall classification accuracy as well as sensitivity, specificity, and positive and negative predictive values.

Results in Table 3(a) suggest that feature numbers 2, 6, and 7 are the best markers for diagnosing breast cancer from the given dataset, while the poorest features are 3, 4, and 9, see Table 2 for the corresponding feature names. To verify these results, two models were synthesized at the same level of default model complexity (CPM = 1): one using only features {2, 6, 7} as inputs and the other using only features {3, 4, 9} as inputs. When run on the evaluation set, the first model gave 7 errors while the second gave 10 errors. Referring to Table 2, the best features are: Uniformity of cell size, Bare nuclei, and Bland chromatin. Rough set data analysis of the breast cancer dataset 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 [44].

5.2 The Heart Disease Data

Feature subset selection for a 3-member committee 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. Table 4(a) lists features selected for each model, indicating the new group of features introduced at each stage. The table also shows the corresponding percentage classification errors obtained when the models were evaluated on the same evaluation set of 80 cases. The model at CPM = 2 is the optimum monolithic model encountered, with a classification error of 15%. Table 4(b) lists the feature subsets, each containing 4 to 5 features, assigned to the three committee members. Each Member was trained on the full training set at the CPM value given in the table. Shown also are the percentage classification errors obtained when the individual members were evaluated on the evaluation set. When the member outputs were combined using simple majority voting of categorical member outputs, the committee achieved a classification error of 12.5%. This value is 28.6% lower than that for the best committee member and 16.7% lower than that for the best monolithic model developed with the full feature set. 10-fold cross validation classification error rates on the same dataset for 10-member neural networks using the bagging and boosting resampling techniques are 16.7% and 19.1%, respectively [41]. Table 4(c) gives a detailed performance comparison between the committee and the default monolithic output (CPM = 1) showing improvements in overall classification accuracy as well as sensitivity, specificity, and positive and negative predictive values. The committee classifier increases the sensitivity and the positive predictive value by approximately 7 percentage points. The overall classification accuracy of 87.5% is significantly better than the best value of 80%

reported previously for a 2-member committee trained on a split training set [15]. The relatively small dataset makes training on different feature subsets a more viable option in this case.

Results in Table 4(a) suggest that feature numbers 9, 12, and 13 are the best markers for diagnosing heart disease from the given dataset, while the poorest features are 6, 7, 11, and 1. To verify these results, two models were synthesized at the same default level of model complexity (CPM = 1): one using only features {9, 12, 13} as inputs and the other using only features {6, 7, 11, 1}as inputs. When run on the evaluation set, the first model gave 16 errors while the second gave 25 errors. Referring to Table 2, the best features are: Exercise induced angina (EXANG), Number of major vessels colored by fluoroscopy (CA), and Thal. Duch, Adamczak, and Grabczewski [45] derive the following rule as one of three classification rules that describe the dataset:

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

5.3 The Diabetes Data

Feature subset selection for a 3-member committee was carried out using the training set of 518 cases with all 8 features available as inputs. Training was performed in three steps of increasing model complexity corresponding to CPM = 3, 1.8, 0.5. Table 5(a) lists features selected for each model, indicating the new group of features introduced at each stage. The table also shows the corresponding percentage classification errors obtained when the models were evaluated on the same evaluation set of 250 cases. Both models with CPM = 3 and CPM = 0.5 give optimum performance for monolithic models, with a classification error of 24.4%. Table 5(b) lists the feature subsets, each containing 2 to 3 features, assigned to the three committee members. Each Member was trained on the full training set at the CPM value given in the table. Shown also are the percentage classification errors obtained when the individual members were evaluated on the evaluation set. With the linear outputs of the three committee member combined using simple

averaging, the committee achieved a classification error of 22.4%. This value is 12.5% lower than that for the best committee member and 8.2% lower than that for the best monolithic module developed with the full feature set. 10-fold cross validation classification error rates on the same dataset for 10-member neural networks using the bagging and boosting resampling techniques are 23.2% and 22.8%, respectively [41]. Table 5(c) gives a detailed performance comparison between the committee and the default monolithic output (CPM = 1) showing improvements in overall classification accuracy as well as specificity and positive and negative predictive values. About 5 and 10 percentage points are gained on specificity and positive predictive value, respectively, at the expense of approximately 2 points lost on sensitivity. The relatively few number of features available in this dataset makes ensembling using different feature subsets less attractive, and it may be responsible for the comparatively poorer performance of the resulting committee. The overall classification accuracy of 77.6% is lower than the value of 78.8% reported previously for a 3-member committee trained on a split training set [15]. The relatively large dataset and small feature set in this case makes training on different subsets of the training data a more viable option.

Results in Table 5(a) suggest that feature numbers 2, 6, and 8 are the best markers for diagnosing diabetes from the given dataset, while the poorest is feature 4. Features 7, 5, 1, and 3 have intermediate predictive quality. To verify these results, two models were synthesized at the same default level of model complexity (CPM = 1): one using only features {2, 6, 8} as inputs and the other using only features {1, 3, 4} as inputs. When run on the evaluation set, the first model gave 60 errors while the second gave 80 errors. Referring to Table 2, the best three features are: Plasma glucose concentration in an oral glucose tolerance test, Body mass index, and Age. Zhu and Guan [46] have found that the first two of these three features score the highest values for the relative importance factor (RIF), respectively, among all features of the dataset.

6. Conclusions

Abductive network committees with members trained on different feature subsets can help improve classification performance compared to a single model trained on the full feature set. Committee classifiers also train and executes faster. Splitting the feature set, rather than the training examples, among the committee members is a more effective way of introducing diversity while avoiding the high dimensionality problem in situations where a few training examples but a large number of features exist. A novel approach for ensemble feature subset selection is introduced which assigns subsets of approximately equal overall predictive quality to the committee members. The approach relies on ranking the features according to their predictive power. Feature ranking is performed using the property of the GMDH-based abductive network learning algorithm of automatically selecting optimum predictive features at various levels of model complexity imposed by the user. Committees comprising only three members and training on as few as 8 features in total achieve an appreciable gain in classification performance compared to the best single models and the best committee members. Performance is comparable with results reported in the literature using other ensembling techniques including bagging and boosting. Results were compared with those reported earlier on two of the datasets using abductive networks trained by splitting the training set. The comparison indicates that splitting the feature subset would be particularly advantageous for small datasets having an adequate number of features, e.g. the heart disease dataset. As a by product, the feature ranking performed gives an insight into the contribution of the various disease markers to the diagnosis problem at hand, which should be of interest to medical practitioners. Information gained in this regard match findings reported in the literature using other methods for feature ranking and selection. Future work would attempt to further refine the feature ranking procedure and apply the technique for reducing the dimensionality of single classifiers.

Acknowledgement

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

References

[1] G.J. Montgomery, K.C. Drake, Abductive networks, Proceedings of the SPIE Conference on the Applications of Artificial Neural Networks, (1990) 56-64.

[2] S.J. Farlow, The GMDH algorithm, in: S.J. Farlow (Ed.), Self-Organizing Methods in Modeling: GMDH Type Algorithms, Marcel-Dekker, New York, 1984, pp. 1-24.

[3] R.E. Abdel-Aal, A.M. Mangoud, Modeling obesity using abductive networks, Comput. Biomed. Res. 30 (1997) 451-471.

[4] R.E. Abdel-Aal, A.M. Mangoud, Abductive machine learning for modeling and predicting the educational score in school health surveys, Methods Inf. Med. 35 (1996) 265-271.

[5] J. Echauz, G. Vachtsevanos, Neural network detection of antiepileptic drugs from a single EEG trace**,** Proceedings of the Electro/94 International Conference (1994) 346-351.

[6] T. Kondo, A.S. Pandya, J.M. Zurada, GMDH-type neural networks and their application to the medical image recognition of the lungs**,** Proceedings of the 38th IEEE SICE Annual Conference, (1999) 1181-1186.

[7] J. Cheung, Z.Y. Lin, R.W. McCallum, J.D.Z. Chen, Screening of delayed gastric emptying using electrogastrography and abductive networks**,** Gastroenterology Suppl. S 112 (1997) A711.

[8] B.O. Peters, G. Pfurtscheller, H. Flyvbjerg, Automatic differentiation of multichannel EEG signals**,** IEEE Transactions on Biomedical Engineering 48 (2001) 111–116.

[9] P. Gopinath, N.P. Reddy, Toward intelligent Web monitoring: performance of committee neural networks vs single neural network, Proceedings of the IEEE International Conference on Information Technology Applications in Biomedicine (2000) 179-182.

[10] A.J.C. Sharkey, N.E. Sharkey, S.S. Cross, Adapting an ensemble approach for the diagnosis of breast cancer, Proceedings of the International Conference on Artificial Neural Networks (1998) 281-286.

[11] Z.-H. Zhou, Y. Jiang, Y.-B. Yang, S.-F. Chen, Lung cancer cell identification based on artificial neural network ensembles, Artificial Intelligence in Medicine 24 (2002) 25-36.

[12] A. Swann, N. Allinson, Fast committee learning: Preliminary results, Electronics Letters 34 (1998) 1408-1410.

[13] S.-J. Kim, B.-T. Zhang, Combining locally trained neural networks by introducing a reject class**,** IEEE International Joint Conference on Neural Networks (1999) 4043-4047.

[14] J. Krogh, A. Vedelsby, Neural network ensembles, cross validation, and active learning, Proceedings of the Conference on Neural Information Processing Systems (1995) 231-238.

[15] R.E. Abdel-Aal, Abductive network committees for improved classification of medical data, Methods of Information in Medicine 43 (2004) 192-201.

[16] S.-B. Cho, J. Ryu, Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features, Proceedings of the IEEE 90 (2002) 1744-1753.

[17] A. Tsymbal, S. Puuronen, Ensemble feature selection with the simple Bayesian classification in medical diagnostics**,** 15th IEEE Symposium on Computer-Based Medical Systems (2002) 225-230.

[18] I. Skrypnyk, A. Grzanka, S. Puuronen, A. Szkielkowska, Selection of voice features to diagnose hearing impairments of children, 14th IEEE Symposium on Computer-Based Medical Systems (2001) 427-432.

[19] P. Cunningham, J.G. Carney, Diversity versus quality in classification ensembles based on feature selection, 11th European Conference on Machine Learning (2000), in: Lecture Notes in Artificial Intelligence, R. López de Mántaras and E. Plaza, (Eds.), Springer Verlag, 2000, pp. 109-116.

[20] S. Yu, S. De Backer, P. Scheunders, 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) 1912-1916.

[21] A.J. Hoffman, C. Hoogenboezem, N.T. van der Merwe, C.J.A. Tollig, Seismic buffer recognition using mutual information for selecting wavelet based features, IEEE International Symposium on Industrial Electronics (1998) 663-667.

[22] W.H. Abdulla, N. Kasabov, Reduced feature-set based parallel CHMM speech recognition systems**,** Information Sciences 156 (2003) 21-38.

[23] M. Ozdemir, M.J. Embrechts, F. Arciniegas, C.M. Breneman, L. Lockwood, K.P. Bennett, Feature selection for in-silico drug design using genetic algorithms and neural networks**,** IEEE Mountain Workshop on Soft Computing in Industrial Applications (2001) 53-57.

[24] D. Garrett, D.A. Peterson, C.W. Anderson, M.H. Thaut, Comparison of linear, nonlinear, and feature selection methods for EEG signal classification, IEEE Transactions on Neural Systems and Rehabilitation Engineering 11 (2003) 141–144.

[25] M.A. Kupinski, M.L. Giger, Feature selection and classifiers for the computerized detection of mass lesions in digital mammography, International Conference on Neural Networks (1997) 2460-2463.

[26] M.F. McNitt-Gray, H.K. Huang, J.W. Sayre, Feature selection in the pattern classification problem of digital chest radiograph segmentation, IEEE Transactions on Medical Imaging 14 (1995) 537–547.

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

[28] A. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artificial Intelligence 97 (1997) 245-271.

[29] K. Kira, L.A. Rendell, A practical approach to feature selection, Proceedings of the Ninth International Conference on Machine Learning (1992) 249-256.

[30] R. Kohavi, G.H. John, Wrappers for feature subset election, Artificial Intelligence 7 (1997) 273-323.

[31] R.W. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognition Letters 24 (2003) 833-849.

[32] T.K. Ho, The random subspace method for constructing decision forests**,** IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 832-844.

[33] M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, Proceedings of the Seventeenth International Conference on Machine Learning (2000) 359–366.

[34] D. Opitz, Feature selection for ensembles, 16th National Conference on Artificial Intelligence (1999) 379-384.

[35] Y.A Pachepsky, W.J. Rawls, Accuracy and reliability of pedotransfer functions as affected by grouping Soils**,** Soil Sci. Soc. Am. J. 63 (1999) 1748–1757.

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

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

[38] A.R. Barron, Predicted squared error- a criterion for automatic model selection, in: S.J. Farlow (Ed.), Self-Organizing Methods in Modeling: GMDH Type Algorithms, Marcel-Dekker, New York, 1984, pp. 87-103.

[39] AbTech Corporation, Charlottesville, VA, ModelQuest Prospector Software, 1995.

[40] W.H. Wolberg, O.L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology, Proceedings of the USA National Academy of Sciences 87 (1990) 9193-9196.

[41] D.W. Opitz, R.F. Maclin, An empirical evaluation of bagging and boosting for artificial neural networks, International Conference on Neural Networks (1997) 1400–1405.

[42] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, V. Froelicher, International application of a new probability algorithm for the diagnosis of coronary artery disease, American Journal of Cardiology 64 (1989) 304-310.

[43] J.W. Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler, R.S. Johannes, Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, Proceedings of 12th Symposium on Computer Applications in Medical Care (1988) 261-265.

[44] A.E. Hassanein, J.M.H. Ali, Rough set approach for generation of classification rules of breast cancer data, Informatica 15 (2004) 23-38.

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

[46] F. Zhu, S. Guan, Feature selection for modular GA-based classification**,** Applied Soft Computing 4 (2004) 381-393.

Table 1. Summary statistics for the three datasets used.

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
Diabetes	8	768	34.9	518	35.1	250	34.4

Table 2. Brief description of the features in the three datasets used.

Feature Number in Dataset	Feature Description		
	Breast Cancer Dataset	Heart Disease Dataset	Diabetes Dataset
1	Clump thickness	Age	Number of pregnancies
2	Uniformity of cell size	Sex	Plasma glucose concentration in an oral glucose tolerance test
3	Uniformity of cell shape	Chest pain type (4 values)	Diastolic blood pressure (mm Hg)
4	Marginal adhesion	Resting blood pressure	Triceps skin fold thickness (mm)
5	Single epithelial cell size	Serum cholesterol in mg/dl	Two-hour serum insulin (μU/ml)
6	Bare nuclei	Fasting blood sugar > 120 mg/dl	Body mass index
7	Bland chromatin	Resting electrocardiographic results (values: 0,1,2)	Diabetes pedigree function
8	Normal nucleoli	Maximum heart rate achieved	Age (years)
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. Results for the Breast Cancer data. (a): Single models of increased complexity synthesized using the full set of features. (b): Composition and performance of the three member models and the resulting committee. (c): Detailed performance comparison between the committee in (b) and the default single model.

(a)

Step Number	CPM	Features Automatically Selected			Classification Error, %	Remarks
1	2.5	2, 6, 7			4	Simplest model
2	2	2, 6, 7	1, 5, 8		2.5	Best monolithic model encountered
3	1.5	2, 6, 7	1, 5, 8	3, 4, 9	5.5	

(b)

Member Number	Feature Subset Assigned	CPM	Classification Error, %	Remarks	3-Member Committee	
					Output Combining Method	Classification Error, %
1	2, 1, 9	2	5		Simple averaging or Majority voting	2
2	6, 5, 4	2	4.5			
3	7, 8, 3	1	4	Best member		

(c)

Model	Sensitivity, %	Specificity, %	Positive Predictive Value, %	Negative Predictive Value, %	Overall Classification Accuracy, %
Default Monolithic Model (All attributes, CPM = 1)	92.5	98.5	96.9	96.3	96.5
3- Member Committee in (b) above	95.5	99.2	98.5	97.8	98

Table 4. Results for the Heart Disease data. (a): Single models of increased complexity synthesized using the full set of features. (b): Composition and performance of the three member models and the resulting committee. (c): Detailed performance comparison between the committee in (b) and the default single model.

(a)

Step Number	CPM	Features Automatically Selected				Classification Error, %	Remarks
1	4	9, 12, 13				18.75	Simplest model
2	2	9, 12, 13	3, 2, 10			15	Best monolithic model encountered
3	1.5	9, 12, 13	3, 2, 10	4, 8, 5		16.25	
4	1	9, 12, 13	3, 2, 10	4, 8, 5	6, 7, 11, 1	17.5	

(b)

Member Number	Feature Subset Assigned	CPM	Classification Accuracy, %	Remarks	3-Member Committee	
					Output Combining Method	Classification Accuracy, %
1	9, 3, 4, 6	1	21.25		Majority Voting	12.5%
2	12, 2, 8, 7, 1	2	22.25			
3	13, 10, 5, 11	1	17.5	Best member		

(c)

Model	Sensitivity, %	Specificity, %	Positive Predictive Value, %	Negative Predictive Value, %	Overall Classification Accuracy, %
Default Monolithic Model (All attributes, CPM = 1)	71.4	91.1	86.2	80.4	82.5
3- Member Committee in (b) above	77.1	95.6	93.1	84.3	87.5

Table 5. Results for the Diabetes data. (a): Single models of increased complexity synthesized using the full set of features. (b): Composition and performance of the three member models and the resulting committee. (c): Detailed performance comparison between the committee in (b) and the default single model.

(a)

Step Number	CPM	Features Automatically Selected			Classification Error, %	Remarks
1	3	2, 6, 8			24.4	Simplest model
2	1.8	2, 6, 8	7, 5, 1, 3		25.2	
3	0.5	2, 6, 8	7, 5, 1, 3	4	24.4	Best monolithic model encountered

(b)

Member Number	Feature Subset Assigned	CPM	Classification Error, %	Remarks	3-Member Committee	
					Output Combining Method	Classification Error, %
1	2, 1, 3	0.5	25.6	Best member	Simple averaging	22.4
2	6, 7, 4	0.5	31.2			
3	8, 5	1	31.6			

(c)

Model	Sensitivity, %	Specificity, %	Positive Predictive Value, %	Negative Predictive Value, %	Overall Classification Accuracy, %
Default Monolithic Model (All attributes, CPM = 1)	53.1	89	75	75.3	75.2
3- Member Committee in (b) above	51	94.2	84.5	75.5	77.6

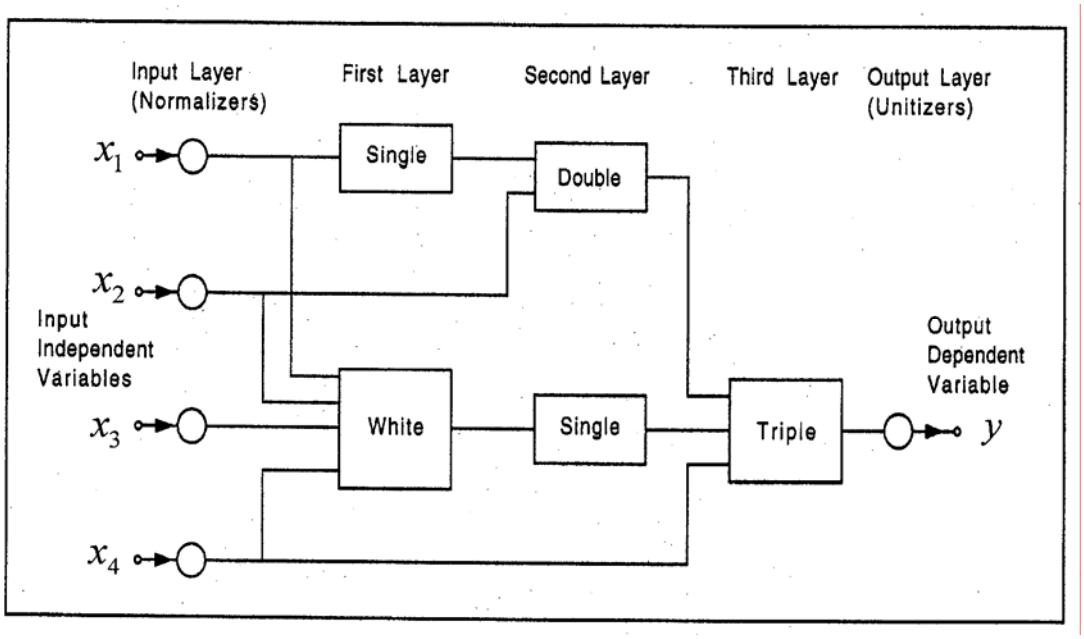


Fig. 1. AIM abductive network showing various types of functional elements.

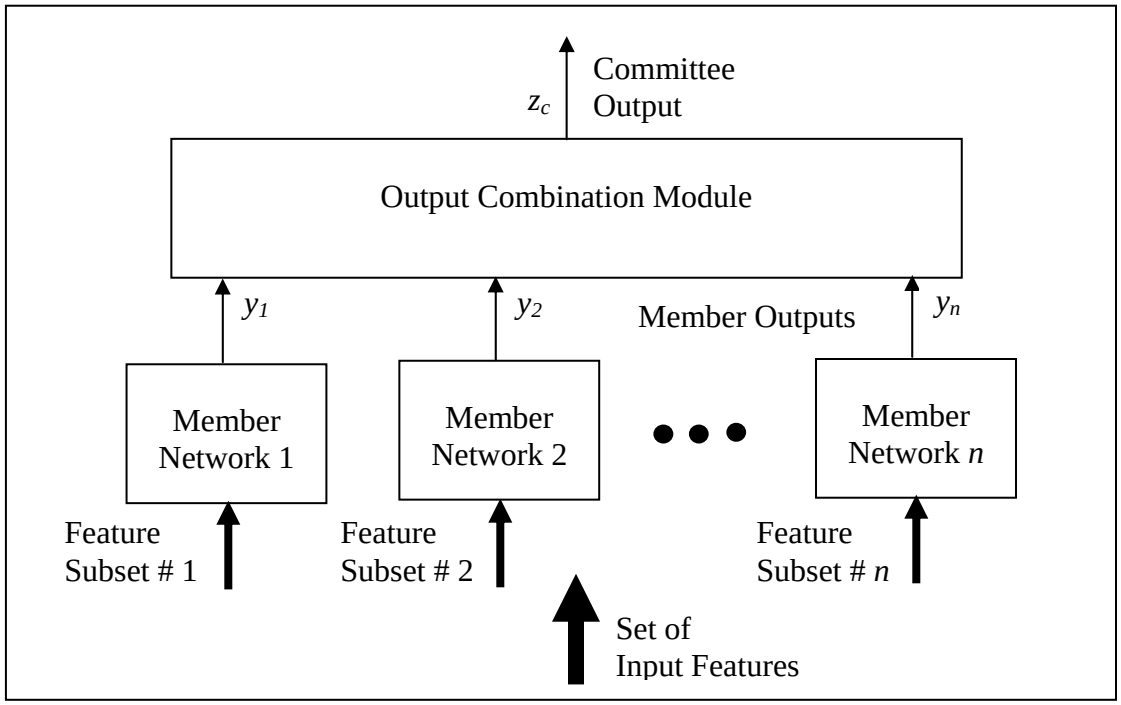


Fig. 2. Schematic of a network committee trained on different feature subsets.

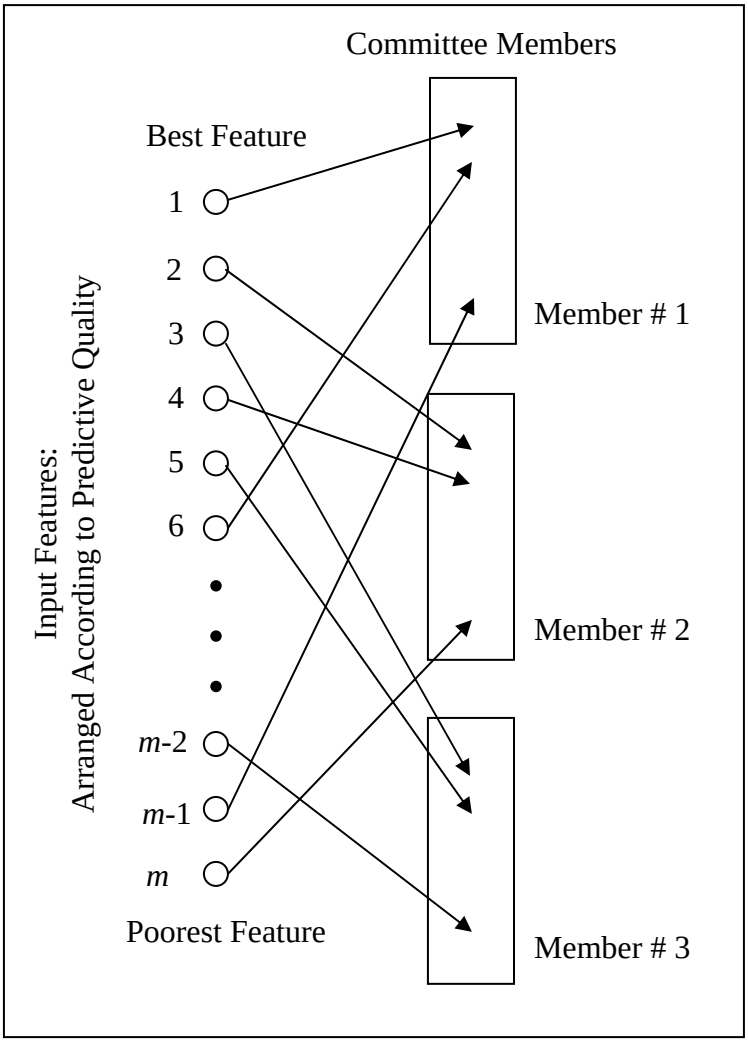


Fig. 3. Schematic showing the procedure for ensemble feature subset selection

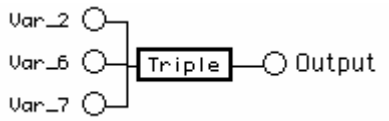
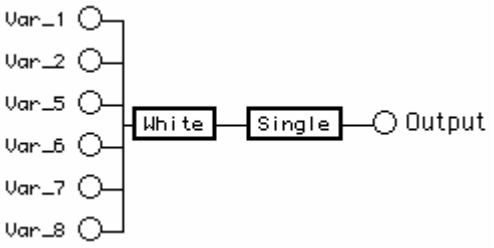
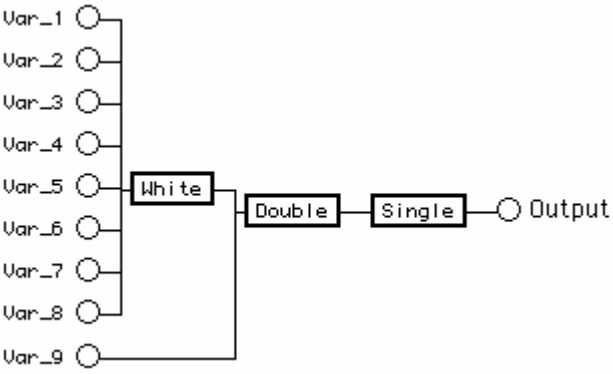
Step	CPM	Resulting Model Structure
1	2.5	
2	2	
3	1.5	

Fig. 4. 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.