Abductive Network Committees for Improved Classification of Medical Data

R. E. Abdel-Aal
Center for Applied Physical Sciences, Research Institute
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

Objectives: To introduce abductive network classifier committees as an ensemble method for improving classification accuracy in medical diagnosis. While neural networks allow many ways to introduce enough diversity among member models to improve performance when forming a committee, the self-organizing, automatic-stopping nature, and learning approach used by abductive networks are not very conducive for this purpose. We explore ways of overcoming this limitation and demonstrate improved classification on three standard medical datasets.

Methods: Two standard 2-class medical datasets (Pima Indians Diabetes and Heart Disease) and a 6-class dataset (Dermatology) were used to investigate ways of training abductive networks with adequate independence, as well as methods of combining their outputs to form a network that improves performance beyond that of single models.

Results: Two- or three-member committees of models trained on completely or partially different subsets of training data and using simple output combination methods achieve improvements between 2 and 5 percentage points in the classification accuracy over the best single model developed using the full training set.

Conclusions: Varying model complexity alone gives abductive network models that are too correlated to ensure enough diversity for forming a useful committee. Diversity achieved through training member networks on independent subsets of the training data outweighs limitations of the smaller training set for each, resulting in net gain in committee performance. As such models train faster and can be trained in parallel, this can also speed up classifier development.

Keywords:

Abductive networks, Neural networks, Ensemble methods, Network Committee, Committee of experts, Classification accuracy, Medical diagnosis, Diabetes, Heart disease, Dermatology.

1. Introduction

Data mining and machine learning techniques for classification, association detection, sequential/temporal pattern recognition, and clustering/segmentation offer new and effective approaches to handle the data overload problem in medical informatics. They automatically discover patterns in medical data to provide support for the decision-making process in many health care areas, including screening, diagnosis, prognosis, monitoring, therapy, survival analysis, and hospital management. Tools used for performing such functions include: Bayesian and nearest-neighbor classifiers, rule induction methods, decision trees, genetic algorithms, fuzzy logic, and artificial neural networks (ANNs). The latter tool has been proposed for various medical applications, including diagnostic systems, biochemical analysis, and image analysis. In spite of the wide use of ANNs as a modeling tool, their opaque (black box) nature has limited their acceptance in medicine [1]. Learned knowledge is concealed in a maze of connections and weights, making it difficult to provide justifications and explanations often sought by physicians [2]. Techniques to improve the comprehensibility of neural networks, such as rule extraction [3], have been used in medicine [4]. Other limitations with neural networks include the difficulty in determining optimum network topology and training parameters. There are many choices to be made in determining numerous critical design parameters with little guidance available [5], and designers often resort to trial and error approaches which can be tedious and time consuming. Such design parameters include the number and size of the hidden layers, the type of neuron transfer functions for the various layers, the learning rate, momentum coefficient, and stopping criteria to avoid over-fitting and ensure adequate generalization with new data.

Compared to neural networks, the alternative approach of self-organizing abductive or polynomial networks based on the group method of data handling (GMDH) algorithm offers the

advantages of more automated and faster model development requiring little or no user intervention [6]. The method offers automatic selection of relevant input variables, automatic configuration of model structures, and faster convergence without the problem of getting stuck in local minima [7]. With the resulting model represented as a hierarchy of polynomial expressions, derived analytical relationships can provide better insight into the modeled phenomena and allow comparison with previously used empirical models. The technique automatically avoids over-fitting through using a criterion for penalizing complexity [8], without requiring a dedicated validation dataset; thus leaving more training data for use in actual training. Medical applications of GMDH-based techniques include modeling obesity [9], analysis of school health surveys [10], drug detection from EEG measurements [11], medical image recognition [12], and screening for delayed gastric emptying [13]. A simplified abductive network model for the waist-to-hip ratio [9] automatically selects only two out 13 input variables, giving a manageable analytical relationship and allowing greater insight into the data. A neural network model would provide no information as to which of the 13 inputs are most influential or how they affect the model output.

The importance of classification accuracy for medical diagnosis cannot be over-emphasized. In screening applications, for example, a high percentage of false negatives increases the risk of real patients not receiving the thorough investigation they need. On the other hand, a larger portion of false positives causes unnecessary inconvenience 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. [14-20]. With this approach, a number of classifiers are used simultaneously and their outputs combined to produce the final predicted committee output, see Fig. 1. The output combination module in Fig. 1 often performs simple functions on the outputs of individual members, such as majority voting or weighted averaging,

without involving the input vectors of attributes [21]. Alternatively, a gating network may use the input vectors to determine the optimum weighting factors for each case to be classified [22]. In the stacked generalization approach, the combiner takes the form of another higher-level network trained on the outputs of individual members to generate the committee classification output [23]. 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. 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 [24]. It can be shown [25], 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 [26] 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. In the 'committee of experts' approach, members are developed using different machine learning techniques that adopt different ways to build decision boundaries for the classification problem at hand, such as neural networks, nearest neighbor classifiers, classification and regression trees (CART), etc. This allows training adequate individual models on the full training dataset available while ensuring a good degree of diversity among them. However, in many situations a committee is restricted to use only one machine learning technology. Neural networks allow great diversity in the available architectures (multi-layer perceptron (MLP), radial basis function (RBF), etc.), learning algorithms (back propagation, simulated annealing, etc.), and in the

parameters that can be varied during training (e.g. network topology, neuron transfer functions, initial random weights, learning rate, momentum, stopping criteria, etc.). This allows many possibilities for constructing individual committee members that are reasonably independent using the same training data.

Although neural network committees have been reported for many applications, there appears to be no mention of GMDH-based abductive (or polynomial) network committees in the literature. 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 diversity in the models that can be synthesized using the same training data. This paper investigates ways of obtaining diverse committee members and demonstrates that committees made up of abductive networks trained on independent subsets of the available training data can give better classification performance than individual committee members and single models that utilize the full training dataset. In effect, this may lead to better utilization of a given training dataset through splitting it into *n* subsets and forming an n-member committee from models trained on those subsets. Individual member networks are expected to train faster than the single model. They can also train in parallel, thus reducing the overall training time. Section 2 gives a brief introduction of the GMDH algorithm and the abductive network modeling tool used, and describes the approach adopted for constructing abductive network committees through training individual members and combining their outputs. Section 3 gives a brief outline of the three medical datasets used in the investigation. Section 4 presents classification results obtained using abductive network committees that employ various methods of combining outputs from individual members. The main objective here has been to demonstrate the effectiveness of abductive networks in improving performance over single models in spite of the limited scope

for injecting diversity. Therefore, no special efforts were taken to optimize the performance of individual models or compare results with committees using other machine learning 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) [27] 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 [28]:

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

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

[8]. 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 [8] 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* [8]. 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) [27], see Fig. 2. 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 Abductive Network Committees

2.2.1 Training individual members

As described in Section 2.1 above, abductive networks adopt a radically different learning approach from that used to build neural networks. With neural networks, the user can choose the size and number of hidden layers, the type of neuron transfer function for the various layers, the values of the initial random weights, the training parameters, and the stopping time or criteria. While this makes reaching an optimum design difficult, it opens the scope for greater diversity

among the resulting models. On the other hand, while self-organizing and self-stopping features of abductive networks simplify model development and reduce user intervention, they are not very conducive in the way of increasing diversity and independence of models generated from a given training dataset. Only a few parameters can be controlled by the user when developing AIM abductive networks, and their default values are often used. The most effective of these is the *CPM* parameter, which directly influences the complexity of the resulting model. We have investigated the effect of changing the *CPM* value on the correlation between the resulting models. As the results given in Section 4.1 indicate, errors by such models turn out to be highly correlated, leading to poor diversity among the committee members and therefore poor gain in classification performance. One alternative is to use different training sets, each of size *NT* cases, to train the individual members. Splitting the full training set available into *n* mutually exclusive subsets to train an n-member committee ensures the greatest possible model independence, but this may reduce the quality of individual models, particularly when the total data available for training is limited. However, since no separate validation set is required to stop training to avoid overfitting (as is the case with neural networks), this reduces the severity of this problem with abductive networks as more data is made available for actual training. Results in Section 4 indicate that the diversity obtained by training on different subsets of data can outweigh the limitation of smaller training sets for the individual models, leading to a net gain in committee performance. Resampling in a form similar to cross validation partitioning may also help strike a balance between accuracy and independence of committee members. The bagging [29] and boosting [30] ensembling techniques use bootstrap resampling.

2.2.2 Combining individual network outputs

We have adopted cooperation schemes in the form of simple combination rules to generate the committee output from the outputs of individual members, see Fig. 1. Consider an n-member, 2-

class classification committee. The negative class is represented by 0 and the positive class is represented by 1. Let y_i and z_c be the continuous outputs from member *i* and from the committee combiner, respectively. Unless otherwise specified, corresponding categorical (classification) outputs are derived from these values by simple rounding (thresholding at 0.5). Following is a brief account of the various output combination methods used:

a. Simple majority vote of categorical members outputs:

The categorical committee output from the combiner is obtained directly by a simple majority vote among the categorical outputs of individual members. For this purpose *n* should preferably be an odd number, with a minimum value of 3.

b. Simple averaging of continuous members outputs:

In this basic ensemble method, the continuous committee output is obtained by simple averaging of individual outputs using the relationship [21]:

$$z_c = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{6}$$

c. Weighted averaging of continuous members outputs using static certainty measures:

In method b above, outputs from all members are assumed to be of equal accuracy. In practice, some outputs may have greater certainty than others, and individual outputs may be weighted to reflect this fact [21]. The committee output is the weighted sum of the outputs of all members:

$$z_c = \sum_{i=1}^{n} \alpha_i y_i \,, \tag{7}$$

$$\text{where } \sum_{i=1}^{n} \alpha_i = 1. \tag{8}$$

As a static measure, the certainty c_i of the output from member (*i*) can be expressed as the inverse of the variance of the error (σ_i^2) by that member over its training set [31]:

$$c_i = \frac{1}{\sigma_i^2}, \quad (9)$$

and the weight α_i is then determined by:

$$\alpha_i = \frac{c_i}{\sum_{j=1}^{n} c_j}, \quad (10)$$

which satisfies the condition on the weights in equation (8).

d. Weighted averaging of continuous members outputs using dynamic certainty measures:

Here the certainty of the output from each member is defined for each input vector classified, and is determined by the closeness of the continuous output to one of the target classification outputs. The output y_i is first limited to the region {0,1} and the certainty of y_i is given by [21]:

$$c(y_i) = \begin{cases} y_i & if \ y_i \geq 0.5 \\ 1 - y_i & otherwise \end{cases} \quad (11)$$

The weights are determined using equation (10) for each case being classified with $c(y_i)$ replacing c_i, and then the continuous committee output is obtained using equation (7).

3. Material

Three standard medical classification datasets from the UCI Machine Learning Repository [32] were used for this study. Following is a brief description of each dataset:

3.1 The Pima Indians Diabetes Dataset

This dataset [33] consists of 768 records of female patients at least 21 years old of Pima Indian heritage. There are eight numerical attributes representing physiological measurements and medical test results, including: number of pregnancies, plasma glucose concentration in an oral glucose tolerance test, diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2-hour serum insulin (μU/ml), body mass index, diabetes pedigree function, and age (years). A

binary-valued class variable indicates whether the patient shows signs of diabetes according to World Health Organization criteria (1) or not (0). The percentage of positives in the whole dataset is approximately 34.9%. 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 [34].

3.2 The Heart Disease Dataset

This dataset [35] is based on data from the Cleveland Clinic Foundation and consists of 270 records, each having 13 attributes (a subset of an original set of 75 attributes). The attributes include age, sex, chest pain type (4 values), resting blood pressure, serum cholesterol (mg/dl), a binary variable indicating if fasting blood sugar exceeds 120 mg/dl or not, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise induced angina, oldpeak = ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy, thal: 3 = normal; 6 = fixed defect; 7 = reversable defect. A binary-valued class variable indicates the presence (1) or absence (0) of heart disease. The percentage of positives in the whole dataset is approximately 44.4%. 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 [34].

3.3 The Dermatology Dataset

This multiple-class dataset [36] has been used for the differential diagnosis of Erythemato-Squamous diseases. It consists of 366 records, each having 34 attributes including age, family history, 10 other clinical attributes, and 22 histopathological features determined by the analysis

of skin samples under the microscope. Each attribute other than age and family history was given a degree in the range 0 to 3, where 0 indicates the feature being absent, 3 indicates the largest amount possible, and 1, 2 indicate intermediate values. The class variable is an integer code ranging from 1 to 6 that indicates one of the following six possible diseases: psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. Eight records in the original dataset had the age attribute missing, and these were excluded leaving 358 records for use in this study.

4. Results

4.1 The Pima Indians Diabetes Data

The 768-case diabetes dataset was randomly split into a training set of 669 cases and an evaluation set of 99 cases, with the percentages of positives in each being 34.1% and 40.4%, respectively. Classification performance reported here may be somewhat degraded by the larger percentage of positives in the evaluation set as compared to the training set. Our first experiment was to construct a 3-member committee from three models, each trained on the full training set but having different levels of model complexity (different *CPM* values of 0.5, 1, and 2, respectively). The top row of Table 1 gives the individual classification accuracies as well as the average value for the three models over the evaluation set. Pair-wise scatter plots for the errors in the continuous outputs of the three models over the evaluation set are shown in the left hand side of Fig. 3. The plots show that errors by the three models are highly correlated and that different values for the *CPM* parameter do not give enough diversity and independence in models trained on the same data. Table 1 gives the root mean square (RMS) value of the three correlation coefficients as a measure of the average correlation between the three models. The table shows also the percentage of cases in the evaluation set for which all three models err together

(unanimous error). When this occurs, there is no way a committee can make the correct decision. The more correlated (less independent) the models are, the larger this percentage will be and the poorer the classification performance of the committee. As a simple committee example, the table gives the classification accuracy for a committee classifier that adopts simple majority vote (combination method a in Section 2.2.2).

The second experiment involved randomly splitting the training set into 3 equal subsets (each containing 223 cases). Using the resampling approach, we synthesized 3 models, each trained using two of the three subsets (446 cases) and leaving out the third subset in a manner similar to cross-validation partitioning. In this way, each model differs from each of the other two models in 50% of its training data. All three models were trained using the default value of *CPM* = 1. The middle row of Table 1 lists the corresponding results. The greater independence among the models introduced by the difference in their training data has led to a reduction in both the average correlation coefficient and the percentage of unanimous errors. In the third experiment, the models were trained on the three mutually exclusive subsets of 223 cases each, thus ensuring 100% independence in their training data, and the results are shown in the bottom row of Table 1. This has led to a significant drop in the average correlation coefficient and a reduction in the percentage of unanimous errors. Scatter plots for the model errors in this case are shown at the right hand side of Fig. 3. It is noted that in all three experiments the majority vote network did not achieve any gain in classification accuracy over the best member model. However, results in experiment 3 are more promising, as the committee achieves a larger performance gain over the average member (a gain of 3.1 percent points). This corresponds to the largest diversity among the three committee members as evidenced by the lowest average correlation and the smallest

percentage of unanimous errors. Other methods described below for combining the model outputs achieve a net gain in classification accuracy beyond that of the best committee member.

We have investigated the use of the four methods described in Section 2.2.2 above for combining the three model outputs in experiment 3, and the results are shown in Table 2. All output combination methods except majority vote (method a) lead to a gain in classification accuracy compared to that of the best member model, given in Table 1 as 76.8%. It is noted that all four methods give superior performance to all single models utilizing the full training set of 669 cases at different model complexities (Experiment 1 in Table 1). This indicates that committees formed using abductive network models trained on mutually exclusive subsets of a given training set may outperform a single model trained on the full training set. Diversity in decision making by the committee members compensates for the poorer performance expected from individual members due to the smaller training dataset. As member models can train in parallel and are expected to train faster because of the smaller individual training sets, this approach may also achieve a reduction in the overall training time required compared to a single model. Combination methods b and c give identical performance because the three member models have nearly identical error variances on the training dataset, and weighting by method c roughly amounts to simple averaging. Best committee performance is achieved using output combination method d which gives a classification accuracy of 78.8%, a 5.1 percentage points improvement over the best single model. Table 3 gives a more detailed comparison of the performance of these two classifiers, showing an improvement of 8.7 percentage points in the positive predictive value for the committee classifier.

4.2 The Heart Disease Data

The 270-case heart disease dataset was randomly split into a training set of 220 cases and an evaluation set of 50 cases, with the percentages of positives in each being 42.7% and 52%, respectively. Classification performance reported here may be somewhat degraded by the larger percentage of positives in the evaluation set as compared to the training set. Three models were developed on the full training set using different *CPM* values of 1, 2, and 0.5. The best model has a classification accuracy of 76% on the evaluation set and the average RMS value for the Pearson correlation coefficient for the evaluation errors by the three models is 0.84. The full training set was then split randomly into two subsets, each of 110 cases, and a model was trained on each subset with *CPM* = 1. The best of the two models scored a classification accuracy of 78%, and the correlation coefficient between the errors of the two models was 0.60. Table 4 compares the classification accuracy of the 3-member committee and the 2-member committee using various methods of output combination. The table indicates the superior performance of the 2-member committee over the 3-member committee due to the greater independence between the constituent members. Best performance is obtained using the method of simple averaging and the method of weighting with a dynamic certainty measure, where the accuracy of the 2-member committee exceeds that of the best single model trained on the full training set.

4.3 The Dermatology Data

The 358-case dermatology dataset was randomly split into a training set of 258 cases and an evaluation set of 100 cases. Three models were developed using the full training set with *CPM* = 1, 0.5, and 0.2. The training set was also randomly split into two halves, each having 129 cases. Table 5 shows classification performance for the various single models and committees formed using a number of output combination methods. In all cases, raw model outputs were first limited to the region {0.5,6.5} prior to combination or derivation of class categorical outputs through

simple rounding. Committee C1 consists of 3 models trained on the full training set with different *CPM* values. Committee C2 consists of models M4 and M5, developed on the different subsets of 129 cases each, and M2 developed on the full training set of 258 cases, all using *CPM* = 0.5. This committee achieves an improvement of two percentage points compared to the best single model that uses the full training data. It is noted that this network adopts some form of resampling for the training data, as models M4 and M5 differ from M2 in only 50% of the data. Results show that the best combination method is to take the committee class output as the class categorical output having the simple majority among the members. If no such majority exists (i.e. all categorical outputs of the three members are different), then the class output is obtained by averaging all outputs followed by simple rounding. Table 6 gives the confusion matrix for committee C2 on the 100-case evaluation set. It shows that classes 1 and 4 are classified with 100% sensitivity and that classes 1 and 6 have 100% positive predictive values.

5. Conclusions

Abductive networks can overcome the opacity and incomprehensibility limitations of neural networks that hamper their wide acceptability in medical diagnosis. This is achieved through automatic selection of only significant inputs and providing analytical model relationships. Other advantages include the more automated model synthesis through self-organization and self-stopping. However, when forming abductive network committees such advantages, coupled with the regression-based approach to learning, tend to limit the diversity that can be achieved among resulting models, thus degrading committee performance. We explored ways to overcome this limitation and demonstrated gain in classification accuracy on standard binary and multi-class medical datasets. Two- or three-member committees of models trained on completely or partially different datasets using simple output combination methods achieve improvements between 2

and 5 percentage points in the classification accuracy over the best single model developed using the full training set. For binary classification, best results were obtained with member outputs being combined by simple averaging or through weighting using a dynamic certainty measure. Future work would investigate the effect of other ensembling techniques that use bootstrap sampling, such as bagging and boosting.

Acknowledgement

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

References

[1] Kononenko I. Machine learning for medical diagnosis: History, state of the art and perspective. Artificial Intelligence in Medicine 2001;23:89-109.

[2] Brause RW. Medical analysis and diagnosis by neural networks. 2nd International Symposium on Medical Data Analysis, ISMDA, Madrid, Spain, 2001.

[3] Andrews R, Diederich J, Tickle AB. A survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-Based Systems 1995;8:373-89.

[4] Setiono R. Extracting rules from pruned networks for breast cancer diagnosis. Artificial Intelligence in Medicine 1996;8:37-51.

[5] Hippert HS, Pedreira CE, Souza R.C. Neural networks for short-term load forecasting: A review and Evaluation. IEEE Transactions on Power Systems 2001;16:44–55.

[6] Montgomery GJ, Drake KC. Abductive networks, Proc. of the SPIE Conf. on the Applications of Artificial Neural Networks, Orlando, Florida, 1990, pp. 56-64.

[7] Alves da Silva AP, Rodrigues UP, Rocha Reis AJ, Moulin LS. NeuroDem - a neural network based short term demand forecaster**.** Presented at the IEEE Power Tech. Conf., Porto, Portugal, 2001.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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, 1181-86.

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

[14] Peters BO, Pfurtscheller G, Flyvbjerg H. Automatic differentiation of multichannel EEG signals**.** IEEE Transactions on Biomedical Engineering 2001;48:111–16.

[15] Reddy NP, Rothschild BM. Hybrid fuzzy logic committee neural networks for classification in medical decision support systems**.** Proceedings of the 24th IEEE Conference on Engineering in Medicine and Biology, 2002, 30–31.

[16] Lippmann RP, Shahian DM. Coronary artery bypass risk prediction using neural networks. Annals of Thoracic Surgery 1997;63:1635-43.

[17] Gopinath, P, Reddy, NP. 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. 179-82.

[18] Reddy NP, Das A, Simcox D. Hybrid fuzzy-neural committee networks for recognition of swallow acceleration signals. Proceedings of the 20th IEEE Conference on Engineering in Medicine and Biology, 1998, 1375–76.

[19] Sharkey AJC, Sharkey NE, Cross SS. Adapting an ensemble approach for the diagnosis of breast cancer. Proceedings of the International Conference on Artificial Neural Networks, 1998, 281-286.

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

[21] Jimenez D. Dynamically weighted ensemble neural networks for classification. IEEE World Congress on Computational Intelligence. 1998, 753-56.

[22] Su M, Basu M. Gating improves neural network performance**.** IEEE International Joint Conference on Neural Networks, 2001, 2159–64.

[23] Wolpert DH. Stacked generalization. Neural Networks 1992;5:241-60.

[24] Swann A, Allinson N. Fast committee learning: Preliminary results. Electronics Letters 1998;34:1408-10.

[25] Kim S-J, Zhang B-T. Combining locally trained neural networks by introducing a reject class**.** IEEE International Joint Conference on Neural Networks,1999, 4043-47.

[26] Krogh J, Vedelsby A. Neural network ensembles, cross validation, and active learning. Proceedings of Neural Information Processing Systems, NIPS'94, 1995, 231-38.

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

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

[29] Breiman L. Bagging predictors. Machine Learning 1996;24:123-40.

[30] Freund Y, Schapire R. Experiments with a new boosting algorithm. Proceedings of the 13th International Conference on Machine Learning, 1996, 148-56.

[31] Guo J-J, Luh, PB. Market clearing price prediction using a committee machine with adaptive weighting coefficients. IEEE Power Engineering Society Winter Meeting, 2002, 77–82.

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

[33] Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings of 12th Symposium on Computer Applications in Medical Care (Greenes RA, ed.), IEEE Computer Society Press, 1988, 261-265.

[34] Opitz DW, Maclin RF. An empirical evaluation of bagging and boosting for artificial neural networks. International Conference on Neural Networks, 1997, 1400–05.

[35] 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;64:304-310.

[36] Guvenir HA, Demiroz G, Ilter N. Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals. Artificial Intelligence in Medicine 1998;13:147-65.

Table 1. Summary of results for the three experiments performed on the diabetes dataset.

NT is the training dataset size for individual models.

Experiment	Approach to Member Training	*NT*	Average (RMS) of the 3 Error Correlation Coefficients	Members Classification Accuracies, %				% Unanimous Errors	Majority Vote Committee Classification Accuracy, %
				Member 1	Member 2	Member 3	Average		
1	Full training set, Different *CPM*s (*CPM* = 0.5, 1, 2)	669	0.96	73.7	72.7	71.7	72.7	23.2	73.7
2	Re-sampling (3-fold cross validation partitioning) (Same *CPM* of 1)	446	0.90	71.7	73.7	72.7	72.7	18.2	71.7
3	Split training set (Same *CPM* of 1)	223	0.80	74.7	69.7	76.8	73.7	16.2	76.8

Table 2. Classification performance of the committee in experiment 3 (Table 1) using four methods of combining members outputs.

Outputs Combination Method		Classification Accuracy, %
a	Majority Vote	76.8
b	Simple averaging of raw outputs	77.8
c	Weighted averaging using static certainty measure based on error variance on training sets	77.8
d	Weighted averaging using dynamic certainty measure	78.8

Table 3. Comparison of classification performance for the committee in experiment 3 (Table 1) and the best single model that uses the full training dataset.

Model	Sensitivity, %	Specificity, %	Positive Predictive Value, %	Negative Predictive Value, %	Overall Classification Accuracy, %
Single model (*NT* = 669, *CPM* = 0.5) (Table 1)	57.5	84.7	71.9	74.6	73.7
Committee of Experiment 3 in Table 1, using combination method d in Table 2	62.5	89.8	80.6	77.9	78.8

Table 4. Comparison of classification performance of two committees for the heart disease dataset using four methods of combining members outputs.

Outputs Combination Method		Classification Accuracy, %	
		3-Member Committee, Same training set (*NT*=220) Different *CPM*s (*CPM* = 0.5, 1, 2)	2-Member Committee, Different training sets (*NT*=110) Same *CPM* of 1
a	Majority Vote	76	-
b	Simple averaging of raw outputs	78	80
c	Weighted averaging using static certainty measure based on error variance on training sets	76	76
d	Weighted averaging using dynamic certainty measure	76	80
Best single model among committee members		76	78

Table 5. Summary of results for experiments performed on the dermatology dataset.

Model/Committee	For Single Models:		For Committees: Output Combination Method	Classification Accuracy, %	Remarks
	NT	*CPM*			
Model M1	258	1	-	84	
Model M2	258	0.5	-	91	
Model M3	258	0.2	-	87	
Committee C1: {M1,M2,M3}	-		Simple averaging of raw outputs after limiting to {0.5,6.5}	90	
	-		Majority of categorical outputs	91	Average categorical outputs if no majority exists
Model M4	129	0.5	-	82	Two independent training sets through splitting the full training set
Model M5	129	0.5	-	85	
Committee C2: {M4,M5,M2}	-		Simple averaging of raw outputs after limiting to {0.5,6.5}	91	
	-		Majority of categorical outputs	93	Average categorical outputs if no majority exists

Table 6. Confusion matrix showing best performance of committee C2 (Table 5) on the evaluation set of the dermatology data.

		Predicted						
	Class	1	2	3	4	5	6	Total
True	1	34	0	0	0	0	0	34
	2	0	12	2	1	0	0	15
	3	0	1	15	0	0	0	16
	4	0	0	0	10	0	0	10
	5	0	0	0	1	17	0	18
	6	0	0	0	1	1	5	7
	Total	34	13	17	13	18	5	100

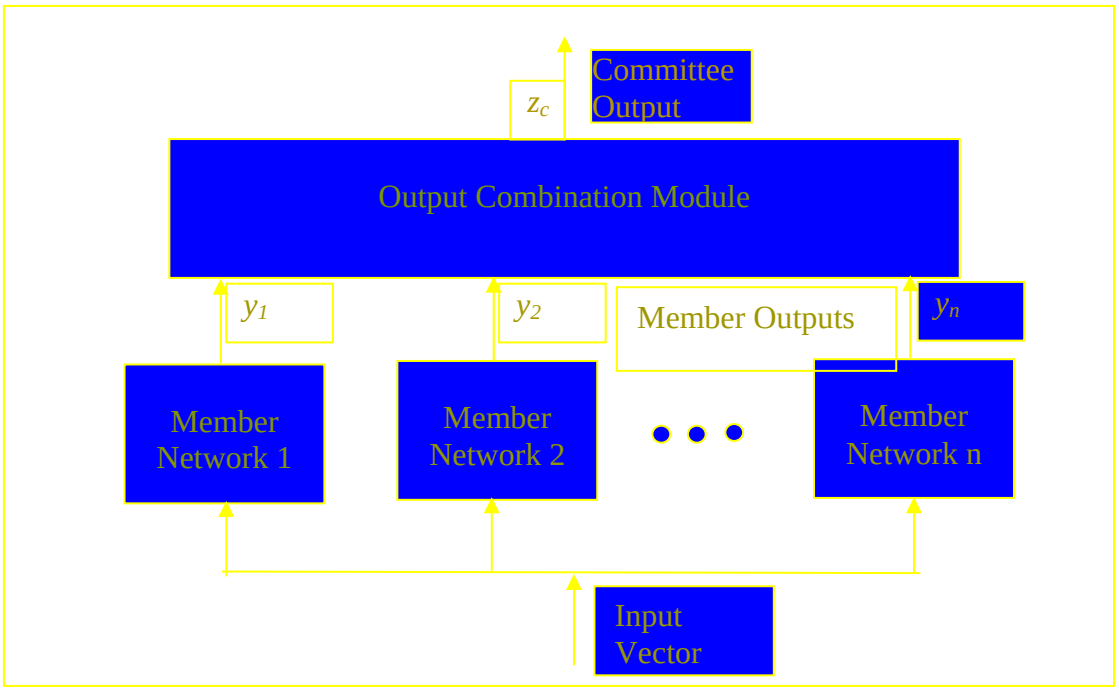

Fig. 1. Schematic of a network committee.

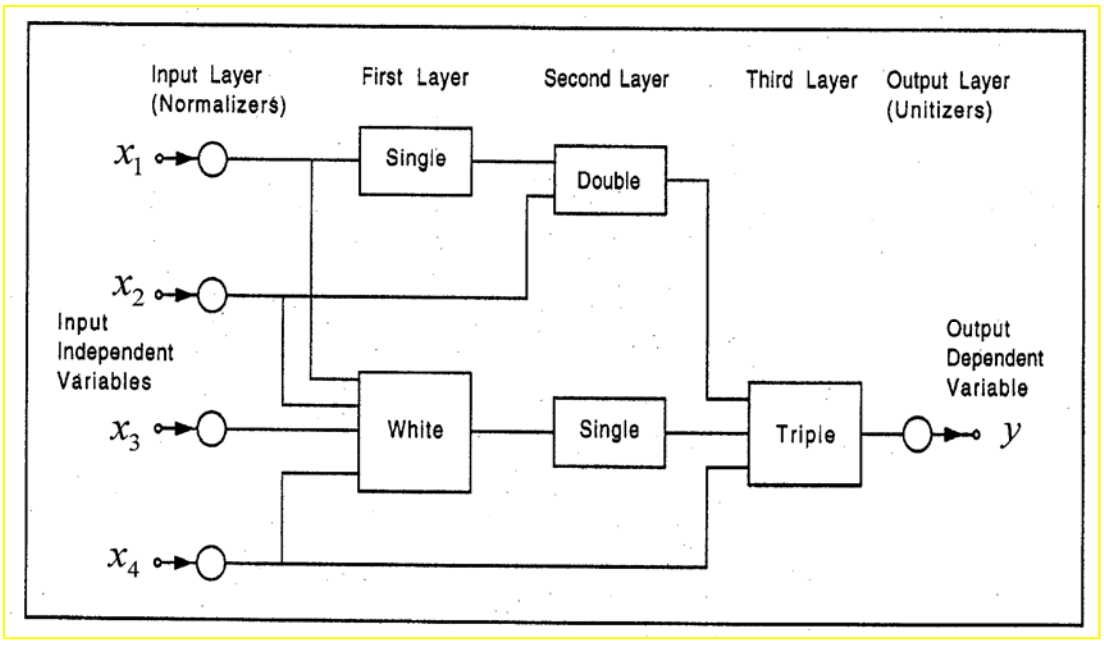

Fig. 2. AIM abductive network showing various types of functional elements.

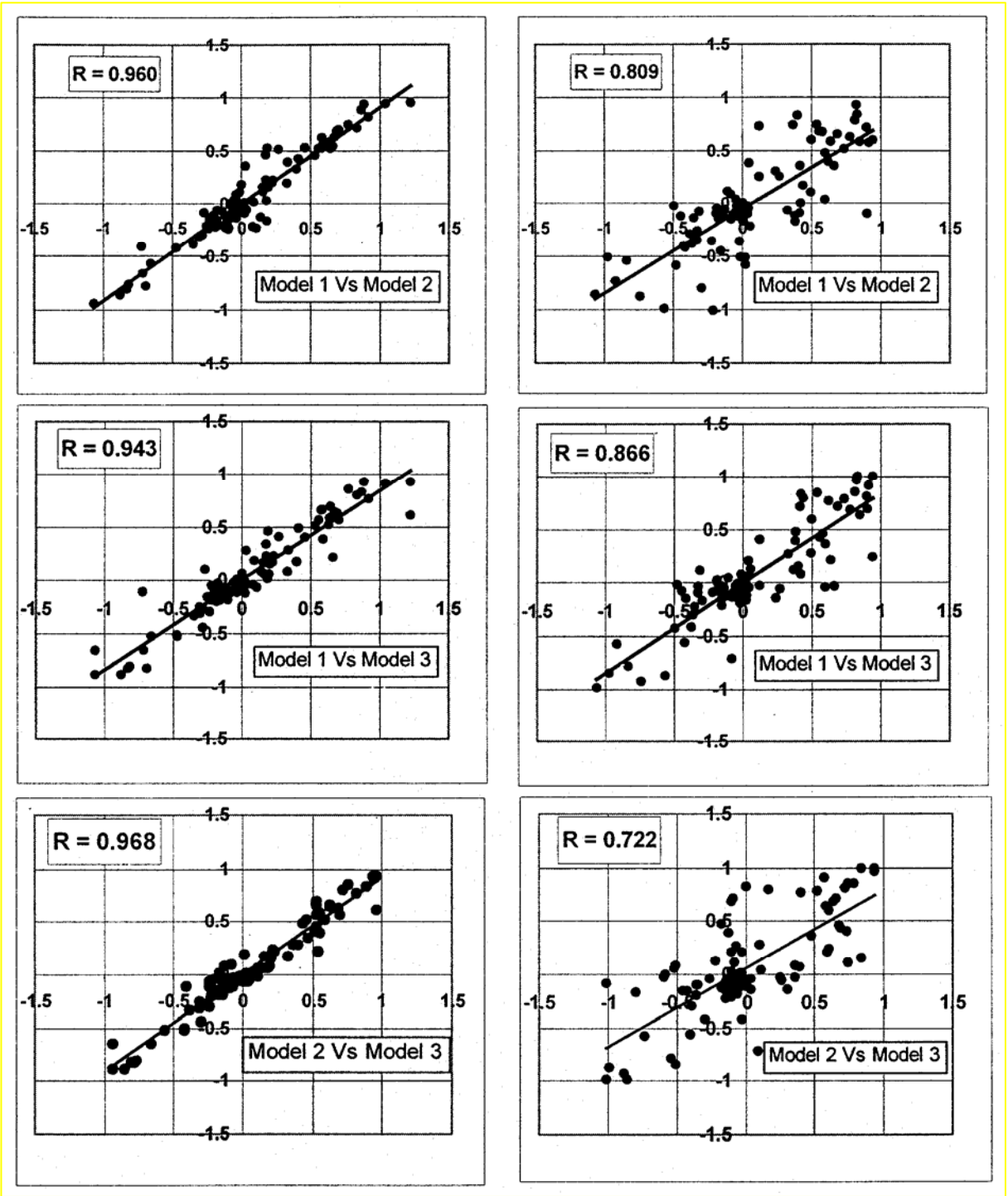

Fig. 3. Pair-wise scatter plots for the raw prediction errors on the evaluation dataset of the diabetes data by individual models used to form the committee in experiment 1 (left) and the committee in experiment 3 (right) in Table 1.