ELSEVIER

Decision Support Systems 36 (2004) 247–259

Decision Support Systems

www.elsevier.com/locate/dsw

Model selection for medical diagnosis decision support systems

Paul Mangiameli [a,*], David West [b], Rohit Rampal [c]

[a] College of Business Administration, University of Rhode Island, Kingston, RI 02881, USA
[b] Department of Decision Sciences, College of Business Administration, East Carolina University, Greenville, NC 27836, USA
[c] School of Business Administration, Portland State University, Portland, OR 97201, USA

Received 1 July 2002; accepted 30 July 2002

Abstract

In this paper, we examine the model section decision for a medical diagnostic decision support system (MDSS). Our purpose in doing this is to understand how model selection affects the accuracy of the decision support system. We explore two related research questions: (1) Do ensembles of models, acting as a single decision maker, perform more accurately than single models; and (2) How does model diversity affect the accuracy of the ensembles? Specifically, we compare 23 single models and bootstrap aggregating (i.e., bagging) models for their predictive abilities across five diverse medical data sets. We are able to reach important conclusions about our research objectives. Ensembles are more accurate than single models in their predictive ability. The best ensemble model achieves an error level significantly lower than the error of the best single model for four of the five medical applications analyzed. The magnitude of the error reduction ranges from 6.4% to 17.5%. Also, when designing an ensemble for an MDSS, the decision to diversify the model selection should be guided by the relationship between model instability and generalization error for the population of models under consideration.

Keywords: Model selection; Medical diagnosis; Neural networks; Bootstrap aggregating models; Diverse ensembles; Baseline ensembles; Bagging models

1. Introduction

The importance of making a correct medical diagnosis cannot be over-stressed. There are emotional, legal, and financial consequences if a patient is told they are ill when, in fact, they are not. The patient suffers extreme emotional distress; the physician may be legally liable for this distress, and, in this time of managed health care, costs for unnecessary medical procedures are incurred. Of far greater consequence is an improper diagnosis concluding the patient is disease-free when they are not. If proper treatment is withheld due to this misdiagnosis, the patient will suffer and possibly die unnecessarily. Any technology that can improve the ability to correctly diagnose human illness is a needed advance to humanity's well being. With the widespread use of electronic data capture and automation of medical records, medical diagnostic decision support systems (MDSS) have become a valuable aid in improving the accuracy of medical diagnosis [29,35]. Their purpose is to

* Corresponding author. Tel.: +1-401-874-4217; fax: +1-401-874-4312.
E-mail address: mangia@uri.edu (P. Mangiameli).

0167-9236/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0167-9236(02)00143-4

enhance, not replace, a physician's ability in the complex and highly intuitive process of medical diagnosis. The use of MDSS in medical diagnosis are predicted to increase 10-fold within the current decade [29].

Traditionally, MDSS are based on a best single model that learns certain physiological characteristics of a given disease and can then be used to diagnose patients who manifest these characteristics. The choice of model used in an MDSS ranges from simple parametric methods, through the non-parametric methods, to various feed forward neural networks. Unfortunately, there is no theory available to guide the selection of the best model. Most MDSS use a model that is most accurate among a limited set of models' relative performance in cross validation trials. In fact, recent research suggests that finding the single "best model" may be the wrong approach [10,11,47,49,50]. It is reported that combinations of single models, referred to as "aggregate predictors," built from perturbed versions of the learning sets may have significantly lower error than that of the best single predictor [11].

In order to avoid confusion, we will adopt the following terminology throughout the remainder of the manuscript. An individual classifier, such as a multilevel perceptron, Fisher's linear discriminant analysis, or kernel density, is referred to as a single model. When single models are combined into an ensemble (regardless of how the learning sets are perturbed) the term aggregate model is used. If the data in the learning sets are perturbed using a bootstrap method, we refer to the resulting model as a *bootstrap aggregating* or *bagging* model. If one model architecture is used to complete the bagging model then it is called a baseline bagging model. If multiple model architectures are used in the bagging model, we label it a diverse bagging model.

The purpose of this research is to investigate the choice of a model (be it single or aggregate) for medical diagnostic decision support systems. We compare the diagnostic accuracy of 23 single models, two diverse bagging models and 23 baseline bagging models, across five medical data sets. The single models studied are the parametric methods of linear discriminant analysis and logistic regression, the non-parametric k nearest neighbor and kernel density, and three categories of feed forward neural networks (multi-layer perceptrons, radial basis functions, and mixtures of experts). The two diverse bagging models are equally weighted (plurality) voting and unequally weighted (frequency) voting [10]. Each of the 23 baseline bagging models are ensembles comprising just one of each of the 23 single models. The criterion we use is the reduction in diagnostic error in terms of overall accuracy. We also report false positive results (diagnosing patients as ill when they are not) and false negative results (diagnosing patients as disease-free when they are ill).

In order to understand how model selection affects the accuracy of the medical decision support system we explore two related research questions: (1) Do ensembles of models, acting as a single decision maker, perform more accurately than single models; and (2) How does model diversity affect the accuracy of the ensembles?

In the next section of this paper, we briefly review the various models selected for this study based upon, in part, several recent MDSS implementations. The third section will then discuss our research methodology and experimental design that we will use to estimate diagnostic accuracy for each model. The fourth section presents our results. This paper concludes with a discussion of these results, and implications to guide the selection of models for medical diagnostic decision support systems.

2. Model selection

2.1. Single models

Many MDSS are implemented with logistic regression, a popular choice that regresses predictor variables on binary targets coded to represent the presence or absence of a disease. For example, logistic regression is used to predict or diagnose spondylarthropathy [17], acute myocardial infarction [20], coronary artery disease [21], liver metastases [24], gallstones [30], ulcers [33], mortality risk for reactive airway disease [37], and breast cancer [44]. Fisher linear discriminant analysis was used to diagnose coronary artery disease [14], acute myocardial infarction [20], and breast cancer [44].

Non-parametric models have also been used to diagnose or predict illness. The k nearest neighbor

was used in comparative studies to diagnose lower back disorders [9], to predict 30-day mortality and survival following acute myocardial infarction [20], and to separate cancerous and non-cancerous breast cancer tumor masses [44]. Kernel density was utilized to determine outcomes from a set of patients with severe head injury [38], and to differentiate malignant and benign cells taken from fine needle aspirates of breast tumors [46].

Neural networks have also been used in a great number of MDSS applications because of the belief that they have greater predictive power [42]. The traditional multilayer perceptron (MLP) was used to diagnose breast cancer [2,3,22,45,48], acute myocardial infarction [4–6,19,32], colorectal cancer [8], lower back disorders [9], drug/plasma concentration levels [39], hepatic cancer [23], sepsis [28], cytomegalovirus retinopathy [36], and ovarian cancer [45]. PAPNET, an MDSS based on an MLP, is now available for screening gynecologic cytology smears [26,27]. The radial basis function (RBF) neural network was used to diagnose lower back disorders [9], in a comparative study of acute pulmonary embolism [40], classify micro-calcifications in digital mammograms [41], and in control of blood transfusion costs for surgery [43]. In a comparison with other methods, RBF and the mixture of expert (MOE) neural networks were used to detect breast cancer [44].

2.2. Aggregate models

Based on the logic of Breiman's research [10,11], as well as that of Dietterich [15,16], we will also examine bootstrap aggregate models. Although not specifically examining MDSS applications, it was found that improved accuracy in classification and prediction is obtained if researchers used aggregate models employing perturbed versions of the learning set [1,10,11,15,16]. There are various methods of perturbing the learning set (e.g., Refs. [15,16]). We employ the bootstrap method, which is described in the next section. All our aggregate models are therefore *b*ootstrap *aggregat*ing models or *bagging* models. The logic behind the improved accuracy from aggregate models is based upon the instability of single model approaches. It is common for small changes in the learning set to create large changes in the results (i.e., accuracy) of a single model. Using an aggregate approach, the effects of these changes tend to be dampened (see Ref. [10], on the bias and variance properties of aggregate models).

Previous research of aggregate models largely focuses on ensembles of one particular method. For example [50], aggregates 30 multilayer-perceptron neural networks with varying numbers of hidden neurons to estimate polymer reactor quality. The author reports that the aggregate model is more robust than a single neural network model. Bootstrap aggregate models from classification and regression trees (CART) are tested on several benchmark data sets. The aggregate CART models achieve reductions in misclassification errors ranging from 6% to 77% [11]. Combinations of nearest neighbor classifiers are trained on a random subset of features; the aggregate model outperforms standard nearest neighbor variants [7]. Aggregate models of hybrid fuzzy logic neural networks are used to recognize swallow acceleration signals [13]. Bagging models of multilayer perceptron neural networks that have identical architecture and starting weights are explored in Ref. [31]. In Ref. [51], aggregate models composed of combinations of linear and quadratic discriminant analysis, logistic regression, and multiplayer perceptron neural networks classify brain spectra by magnetic resonance measurement. The authors report that the aggregate models are more accurate than any single model and that the performance of the single models varies widely, performing well on some data sets and poorly on others.

The literature on aggregate models reports encouraging results. What is not clear is whether the ensemble of models to be aggregated should be limited to just one or two architectural types or should contain a very diverse group of single models. Essentially, our second research question is barely addressed by the literature (see, e.g., Ref. [51]). We will therefore examine both diversified and baseline bagging models.

For those aggregate models with no diversity, we employ baseline bagging models. Each baseline model is an ensemble of a single model architecture. As we have 23 single models we therefore have 23 baseline bagging models. We discuss the design of these models in the next section.

We address model diversity using two aggregating methods. One method of aggregating is to use equally

weighted or plurality voting while the other aggregate method is unequally weighted or frequency voting. Before describing these two methods, it is necessary to understand the output of the single models. As explained in the next section, each model is trained on a training set, evaluated for accuracy (i.e., initially tested) on a validation set and then tested (i.e., the comparatively tested) on a common "hold-out" test set. The output of the single model is a value between zero and one for each data point. The closer to zero, the more likely that the data point belongs to group 0 (in this study, group 0 consists of disease-free patients). Conversely, the closer to one, the more likely that the data point belongs to group 1 (in this study, group 1 contains patients who are ill). For the valuation set data points, this value is often called the likelihood value. Given this output, each single model then classifies the data into the appropriate group (i.e., less than 0.5 is put in group 0 and greater than 0.5 is put in group 1).

The equally weighted or plurality voting aggregate method is relatively simple. The majority of single models that place a test set data point in a given group are the "winners." For example, if 12 of the 23 single models place a data point in group 1 while 11 place the same data point in group 0, then the equally weighted bagging model uses the plurality vote and will place the data point in group 1. This method is similar to that of Ref. [51]. It allows for the diversity of the ensemble to have the greatest impact on the aggregate model's predictive accuracy. We view this aggregating method as representing high model diversity.

The other diverse bagging model that we examine employs the unequally weighted or frequency voting method of aggregation. This method makes use of the accuracy of each single model's classification ability. Each of the 23 single models is evaluated for accuracy on a validation set. For each evaluation, one single model is the most accurate (as described in the next section there are 100 evaluations for each data set). We accumulate the percent of times each single model is the best over the 100 evaluations and refer to this as the frequency of the single model. For the unequally weighted or frequency voting bagging model, we multiply the likelihood value of the test set data point by the frequency of each single model. If the resulting value is below 0.5, the data point is placed within group 0 whereas if it is above 0.5 the data point is placed within group 1. A simple example will help illustrate this method. Three single models, *X*, *Y*, and *Z*, are evaluated individually on 100 validation sets. Model *X* was best 50% of the time (frequency of 0.5), model *Y* was best 30% of the time (frequency of 0.3) and model *Z* was best 20% of the time (frequency of 0.2). For a particular data point in the hold-out test set, model *X* had a likelihood value of 0.8, model *Y* had 0.2, and model *Z* had 0.9. The unequally weighted aggregate voting model would return a value of 0.64 for this data point ($[0.5 \times 0.8]+[0.3 \times 0.2]+[0.2 \times 0.9] = 0.64$). This data point would then be assigned to group 1. By using an unequal weight for each single model in the ensemble based upon that single model's predictive accuracy, the frequency voting aggregate method allows the more accurate single models to dominate the decision-making. We view this aggregating method as representing moderate model diversity.

In the next section, we describe our research methodology. Some of the points raised in the this section will be discussed in more detail.

3. Research methodology

Our research methodology is presented in three parts. The first part presents the five medical data sets that we examine in this study. The second describes the 23 single models that we examine. The third part presents our experimental design with further discussion of the aggregate models.

3.1. Data sets

The five medical data sets represent various illnesses. The diagnosis we seek to render is whether a given patient, based upon various physiological symptoms is ill with the disease in question or is disease-free. Specifically, the diseases we examine are heart disease, liver disease, lung cancer, breast cancer, and cellular cancer (actually cytological studies for metastasized breast tumors). These five data sets are relatively dissimilar and represent a fair test of the diagnostic accuracy of the various models. Please see Ref. [34], for additional description of the data sets.

The characteristics of these data sets vary along four dimensions: the number of data records, the number of variables, the percentage of categorical

variables, and balance. The number of examples contained in the data set ranges from a low of 194 records for the breast cancer data to a high of 683 for the cytology data. The number of variables defines the size of the input space and varies from 6 variables for the liver data to 32 variables for the breast cancer data set. The percent of categorical variables measures the relative mix between real and ordinal input variables. This ranges from 0% for several data sets with only real variables to 100% for the cytology data set, which contains only ordinal variables. Balance measures the ratio of the examples in the largest classification group to examples in the smallest group. This ranges from a relatively balanced value of 1.13 for the lung cancer data to a very unbalanced 3.21 for the breast cancer data. The properties of each of the five data sets used in this research are summarized in Table 1.

Andras Janosi, MD, compiled the heart disease data set at the Hungarian Institute of Cardiology. Ten variables define personal information, pain type and location, blood pressure, sugar, cholesterol, etc. The objective is to classify each instance to one of two groups that represent the diagnosis of the angiographic disease status. The liver disorder data was collected by BUPA Medical Research and includes data from blood tests used to determine if liver damage is present. The cytology data consists of records of breast cytology used in breast cancer diagnosis research at the University of Wisconsin [25,46]. The breast cancer data set represents follow-up data for breast cancer cases and includes only those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis [25,46]. The lung cancer data predicts 6-month survival rate for patients with primary cancer of the lung [18].

Table 1
Data set characteristics

Data set	Number of examples	Number of variables	Percent of categorical variables	Balance
Heart disease	261	10	50	1.66
Liver disease	345	6	0	1.38
Breast cancer	194	32	6	3.21
Cytology	683	9	100	1.86
Lung cancer	200	13	69	1.13

3.2. Single model description

The 23 single models include neural networks, parametric models, and nonparametric models. Three different neural network architectures are used: multi layer perceptrons (MLP), mixture of experts (MOE), and radial basis functions (RBF). The parametric models include Fishers linear discriminant analysis (LDA), and logistic regression (LR) while the nonparametric models included are k nearest neighbor (kNN), and kernel density estimation (KD).

Several key design decisions involving the topology and the learning process are required to define the neural network models. Topology decisions establish the network architecture and include the number of hidden layers and number of neurons in each layer (input, hidden, and output). The number of neurons in the input layer of the neural models is simply the number of variables in the data set. For the neural output layer, exclusive coding is used with an output neuron dedicated to the two medical conditions, ill (group 0) or disease-free (group 1). The hidden layer is more difficult to define. A relatively large hidden layer creates a more flexible diagnostic model. The diagnostic error for such a model will tend to have a low bias component with a large variance caused by the tendency of the model to over-fit the training data. A relatively small hidden layer results in a model with a higher error bias and a lower variance. The design of the hidden layer, therefore, involves a tradeoff between error components. In this research, we use five different hidden layer designs from small to large for each neural network model. This allows us to incorporate a range of architectures from simple to more complex models. The parameters for the nonparametric nearest neighbors and kernel density models are established by trial and error. We found that 5, 7, and 9 of the "nearest neighbors" was a good range to examine. Kernel densities of 0.1, 0.5, and 1.0 performed well in our preliminary trials. We therefore had 23 single models (five each of the three neural network architectures, three each of the two nonparametric models and one each of the parametric models) to examine.

For the neural networks, diagnostic accuracy is also dependent on the dynamics of the network learning process. The most accurate diagnosis results typically do not coincide with network error conver-

gence for the training data [12,50]. We conducted initial experiments with training lengths varying from 30,000 to 300,000 iterations. Based on these experiments, the following training guidelines are used in this research. Network weights are updated after each learning epoch, defined as one cycle through the training data set. The network learning parameters (the learning rate and learning momentum) are decreased every 10 learning epochs, and training is terminated after 50 learning epochs. It is our experience that setting relatively low values for the network learning rate and momentum increases decision accuracy. We therefore use a learning rate of 0.3 and a momentum of 0.4.

3.3. Experimental design

A controlled experimental comparison requires a data partitioning strategy. To assess the accuracy of the diagnostic models, the data set must be split into three partitions: a training set, a validation set for establishing generalization ability, and a final independent test set to measure model accuracy.

A sampling plan that follows the general spirit of earlier works [11,12,47,50] is employed in this paper. First, the data set is randomly divided into a ''hold-out'' test set, which contains approximately 10% of data, and a learning set consisting of the remaining observations. The hold-out test set is used to test the accuracy of all single and bagging models. The learning set is further partitioned into 10 mutually exclusive data sets. One partition is used as a validation set and the other nine as a training set. This process is repeated 10 times with each of the partitions taking its turn as the validation set. The best single model is then identified as the model with the lowest error across the 10 validation sets. An estimate of its accuracy on future unseen patients is measured using the hold-out test set. The process described in this paragraph is typical of the methodology used by practitioners developing an MDSS based on a best single model. The merits of cross validation are that the model is trained with a large proportion of the available data (90% in this case), and that all of the data is used to test the models.

After all 23 single models have been trained and tested, training sets for the aggregate models are constructed from bootstrap samples of the previously created learning set. The purpose of the bootstrap samples is to create different training sets for each of the 23 models thereby increasing the independence of the prediction errors and generating more accurate bagging models. Twenty-three bootstrap samples are constructed from the learning set and each of the single models in the ensemble is trained on one unique bootstrap sample. All of these single models are combined into one ensemble to create the diverse bagging models. The aggregating methods (plurality and frequency voting) have already been discussed in the previous section.

The baseline bagging models are developed similarly to the diverse models. After the each single model is trained and tested, we employ bootstrap perturbations of the training set and, in the case of the neural network models, random initial starting conditions to create 23 different variants of the same single model. All 23 bootstrap variants are put into an ensemble and are aggregated using plurality voting. As there are 23 single models, we have 23 baseline bagging models as well.

Please note that all the aggregate models have the same number in the ensemble, 23. The use of an odd number of models is important as this prevents tie votes. As with the single models, the bootstrap aggregate or bagging models use the same hold-out test set.

The random division of the data into the learning set and test set as well as the 10-fold cross validation and bootstrap sampling described in the previous paragraphs are repeated 100 times for each of the five data sets. The overall error, as well as the false negatives and false positive errors, reported in the next section, are the averages over these 100 iterations.

4. Results

The identification of the best single model is summarized by data set in Table 2. The importance of beginning with a broad search for model accuracy is evident. Table 2 demonstrates that no single model, in fact not one group of models of similar architecture consistently outperforms the other models across the five medical data sets. The best single model for the Cytology data set is almost exclusively one of the radial basis function neural network architectures. For these networks, accuracy is

Table 2
Best single models by data set

	Model	Data set				
		Cytology	Heart disease	Liver disease	Breast cancer	Lung disease
1	MLPa	0.02	0.01	0	0.05	0.14
2	MLPb	0.02	0	0	0.03	0.02
3	MLPc	0	0	0	0.08	0.09
4	MLPd	0	0	0	0.07	0.05
5	MLPe	0	0.01	0	0.06	0.04
6	MOEa	0.02	0.02	0	0.01	0.08
7	MOEb	0.01	0.04	0	0.03	0.02
8	MOEc	0.02	0.06	0	0.02	0.05
9	MOEd	0.03	0.05	0	0.06	0
10	MOEe	0.03	0.07	0	0.07	0.04
11	RBFa	0.03	0.18	0	0.05	0.01
12	RBFb	0.07	0.02	0	0.05	0.01
13	RBFc	0.12	0.01	0	0.08	0.01
14	RBFd	0.32	0.03	0	0.18	0.01
15	RBFe	0.29	0.01	0.08	0.14	0
16	LDA	0	0.09	0	0	0.16
17	LR	0	0.04	0.03	0	0.27
18	KNNa	0	0	0.05	0	0
19	KNNb	0.02	0.03	0.13	0.01	0
20	KNNc	0	0.05	0.14	0.01	0
21	KDa	0	0	0	0	0
22	KDb	0	0	0.27	0	0
23	KDc	0	0.28	0.3	0	0

The numbers in the cells represent the percent of times that a particular model performed the best for the given data set.

	Best single model characteristics by data sets				
	Larger RBF models	Varied mix	All nonparametric	Reasonably uniform mixture of ANNs	Parametric models plus simpler ANNs
Percent stable	0.02	0.49	0.92	0.02	0.43
Number of models above 0%	13	17	7	17	15

seen to increase with network complexity. A very different pattern arises in the liver disease results. With only one exception, RBF, the neural network models are never the most accurate. Accuracy in this data set is dominated by the nonparametric k nearest neighbor and kernel density models. The results for the Heart disease data and Lung disease data are similar. The nonparametric and parametric models are the most accurate for 50% of the validation tests with the remaining best single models scattered across the MOE and RBF networks. The most accurate models for lung cancer consist of the parametric methods with smaller representations of the simpler neural network architectures. The nonparametric methods are never the most accurate for this data set. Lastly, the breast cancer results include a reasonably uniform mix of neural network models. Parametric and nonparametric methods are seldom the most accurate for the breast cancer data. Failure to begin with a broad array of models may result in a sacrifice of accuracy in the MDSS application.

In the following discussion, we compare the error rates of the bagging models to the results of the over-

Table 3
Overall diagnostic error

Data set	Best single model	Plurality voting bagging model	Percentage improvement plurality voting bagging model to best single model	Frequency voting bagging model	Percentage improvement frequency voting bagging model to best single model	Best baseline bagging model	Percentage improvement best baseline bagging model to best single model
Cytology	0.0292	0.0290	0.6849	0.0257***	11.9863	0.0285	2.3973
Heart	0.1971	0.1764***	10.5023	0.1840**	6.6464	0.1712***	13.1405
Liver	0.3370	0.3449	− 2.3442	0.3154**	6.4095	0.3001**	10.9496
Breast cancer	0.2262	0.1885***	16.6667	0.1865***	17.5508	0.2095*	7.3829
Lung	0.3013	0.3050	− 1.2280	0.2935	2.5888	0.3051	− 1.2612

* Difference significant at the 0.05 level.
** Difference significant at the 0.01 level.
*** Difference significant at the 0.001 level.

all single most accurate model for 100 iterations for each of the five data sets investigated. The overall best single model is identified as the single model having the lowest average diagnostic error on the 100 runs of the 10-fold cross validation data. Its error on the hold-out test set is then compared to each bagging model's error on the same hold-out test set.

Table 3 summarizes the overall diagnostic error for each data set. The plurality voting bagging model had a statistically significant reduction in overall error relative to the best single model for two of the five data sets. The percentage improvement is 10.5% for heart disease and 16.67% for breast cancer. Frequency voting, the other diverse bagging model, had statistically significant reductions in diagnostic error for four of the five data sets relative to the best single model. Its diagnostic improvement ranged from 6.41% for liver disease to 17.55% for breast cancer. The last two columns in Table 3 report the results of the most accurate "Best Baseline Bagging Model." Please recall that there are 23 baseline bagging models for each data set. Depicted in this table is the result of the model that achieved the minimum diagnostic error of all 23 models. The best baseline bagging model had statistically lower diagnostic error on three of the five data sets. The magnitude of the error reduction ranged from 7.38% for breast cancer to 13.14% for heart disease. No single best model had statistically lower diagnostic error than any aggregate model for any data set.

Based upon these results we can reach a definite conclusion regarding our first research question. Bootstrap aggregate models are more accurate than single models identified in cross validation trials.

The second research question involving the role of model diversity is a more complicated issue. Table 4 presents the average diagnostic error of all 23 baseline bagging models as well as the maximum and minimum error. Note from Table 4 that the best baseline model varied for each data set and that there was not a strong correlation between the most accurate baseline bagging model and the more accurate single models (see Table 2). From a strictly statistical view, the average error is an expected value obtained when one randomly selects a baseline model from the set of 23 models investigated. Of course, in the real world no one actually selects a baseline bagging model without some a priori guidance. This guidance often tends to gravitate to a researcher's "favorite model" such as neural networks,

Table 4
Overall diagnostic errors of baseline bagging models

Data set	Best baseline bagging model	Minimum	Average	Maximum
Cytology	Radial Basis Function—c	0.0285	0.0386	0.0712
Heart	Kernel Density—d	0.1712	0.1901	0.2209
Liver	Logistic Regression	0.3001	0.3567	0.4072
Breast cancer	Radial Basis Function—c	0.2095	0.2759	0.3858
Lung	Mixture of Experts—b	0.3051	0.3324	0.4102

etc. It can be argued that choosing a favorite model is similar to random selection and that a method of model selection based on a favorite model will, in the long run, be close to the average value. For those few aggregate models having limited diversity with ensemble membership constrained to three or four models, we anticipate that the long run results will also be close to the average. Only those researchers who systematically explore a wide range of models including parametric, nonparametric, and neural network models can expect to get baseline errors approaching the minimums we report in Table 4.

A diverse group of baseline bagging models does not guarantee that the minimum baseline model has a smaller diagnostic error than the diverse aggregate models. From our results it appears that the baseline model may be a better choice only when there is a single model with low error and high instability. This concept is best explained by inspecting a plot of generalization error versus model instability. Fig. 1 will be used to illustrate our point using just the breast cancer and the liver data sets. Model instability is the degree to which the bootstrap training perturbations effect model decisions and is measured by the range of outputs obtained from 500 iterations of the baseline bagging models for each model and for each data set [12,31]. The diverse frequency voting bagging model has the lowest error for data sets with a large positive slope between model instability and generalization error such as the breast cancer data shown in Fig. 1 with a regression slope of 2.50 ($p = 0.000$) (although not illustrated, the Cytology data with a slope of 1.56 ($p = 0.002$) would also fall into this category). For data sets exhibiting a positive relationship between instability and error, there is a subset of models to the lower left portion of the scatter plot that are high potential candidates for an effective bagging model. The frequency voting methods assigns significant weights to these models and excludes models from the upper right portion of the plot that would adversely affect the error of the bagging ensemble. The single baseline bagging model is most effective for data exhibiting a strong negative correlation between model instability and generalization error such as the liver data plotted in Fig. 1 with a regression slope of -1.75 ($p = 0.003$). In these situations, an effective bagging model can be formed from a single model with the ideal properties of low generalization error and high model instability. Such models are located to the lower right of the liver data plot in Fig. 1. Increasing the diversity of a baseline bagging model in these situations involves adding additional models that sacrifice both accuracy and instability. The plurality vote bagging model is most effective for data sets where the magnitude of the slope

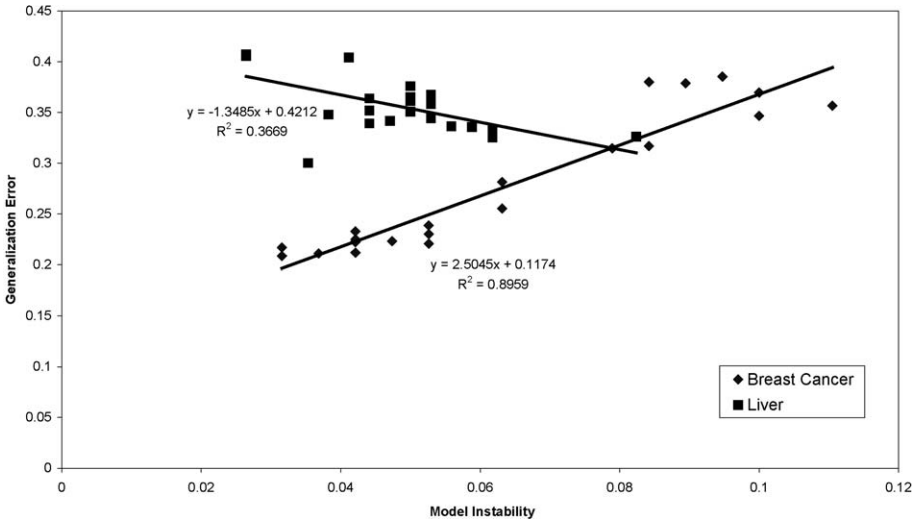

Fig. 1. Generalization error versus model instability.

Table 5
Decomposition of significant differences in false positive diagnostic error

Data set	Best single model	Plurality voting bagging model	Percentage improvement plurality voting bagging model to best single model	Frequency voting bagging model	Percentage improvement frequency voting bagging model to best single model	Best baseline bagging model	Percentage improvement best baseline bagging model to best single model
Cytology	0.0295	NS	NS	0.0281	4.628	NS	NS
Heart disease	0.1365	0.1097	19.651	0.1196	12.369	0.176	− 28.94
Liver disease	0.2339	NS	NS	0.2168	7.316	0.247	− 5.60
Breast cancer	0.0804	0.0378	52.983	0.0464	42.334	0.040	50.25

NS—difference in overall error was not significant.

between model instability and error is relatively small such as the heart data set with a slope of 1.01 (p = 0.001). Under these conditions, it may be reasonable to include all models in the bagging ensemble and to weight their individual decisions equally to determine an aggregate decision.

For medical diagnostic applications, the costs of misclassification errors are not equal and it is informative to decompose the overall error into group specific errors. Table 5 contains the average false positive errors for the best single model and the bagging models for those data sets with significant differences in overall error. The plurality voting bagging model achieves a false positive error improvement, as compared to the best single model, for the heart disease (19.65%) and breast cancer (52.98%) data sets. The frequency voting bagging model improves accuracy by reducing false positive diagnostic errors ranging from 4.63% for the cytology to 42.33% for the breast cancer data sets. Note that baseline bagging model reduces the false positive error rate for only the breast cancer data set. It increases the false positive error rate for both the heart and liver disease data sets.

Table 6 reports the corresponding average false negative errors. The plurality voting bagging model has a minor improvement in false negative diagnostic error, as compared to the best single model, for the breast cancer data set (2.41%) and for the heart disease data set (3.63%). The frequency voting bagging model reduces false negative diagnostic errors for the cytology (27.44%) and breast cancer (7.56%) data sets and with more modest reductions for the heart (2.54%) and

Table 6
Decomposition of significant differences in false negative diagnostic error

Data set	Best single model	Plurality voting bagging model	Percentage improvement plurality voting bagging model to best single model	Frequency voting bagging model	Percentage improvement frequency voting bagging model to best single model	Best baseline bagging model	Percentage improvement best baseline bagging model to best single model
Cytology	0.0275	NS	NS	0.0200	27.440	NS	NS
Heart disease	0.3070	0.2958	3.632	0.2992	2.538	0.159	48.21
Liver disease	0.4746	NS	NS	0.4466	5.892	0.337	28.99
Breast cancer	0.7299	0.7123	2.408	0.6747	7.561	0.737	− 0.97

NS—difference in overall error was not significant.

liver disease (5.89%) data sets. The baseline bagging model achieves dramatic reductions of false negative errors for the heart (48.21%) and liver disease (28.99%) data sets.

5. Concluding discussion

In this paper, we examine the model section decision for a medical diagnostic decision support system. We examine 23 single models, two diverse bootstrap aggregate or bagging models, and 23 baseline bagging models for their predictive accuracy across five diverse medical data sets. Our purpose in doing this is to understand how model selection affects the accuracy of the decision support system. We raised and addressed two related research questions: (1) Do bagging models perform more accurately than single models; and (2) How does model diversity affect the accuracy of the aggregate models? We find that bootstrap aggregate models are more accurate than single models in their predictive ability. Also, when designing an aggregate model for an MDSS, model diversity is critical to selecting the most accurate model. On one level, it is impossible to know, a priori, which baseline bagging model to select. Without a diverse selection from which to choose, the long run error would average above the single best model. This does not imply that a diverse aggregate model is more accurate than a baseline bagging model however. The decision to diversify the model selection should be guided by the relationship between model instability and generalization error for the population of models under consideration. When the slope of the plot of instability versus error is negative, a baseline bagging model may be highly accurate. In cases with large positive slopes the frequency voting aggregate model is expected to be superior because of its ability to screen out models with high error terms. If the plot demonstrates a small positive slope, then the plurality voting aggregate model achieves very accurate results.

Our research is directed at model selection issues to find the most accurate starting point for an MDSS implementation. For any specific application, the economics dictate that one type of misclassification error may be significantly more costly than the other. The idea of intentionally biasing the medical decision support system to minimize the total costs of misclassification is an important aspect of the implementation phase. While we have decomposed the overall error into false positive and false negative rates assuming equal misclassification costs (i.e., a threshold of 0.5), the reader should appreciate that a tradeoff exists between these component errors. There are several methods that can be employed to bias the MDSS ensemble decision. The practitioner can adjust the threshold to values other than 0.5 in order to bias the decisions of the individual ensemble members. A second alternative is to require a more stringent vote total than the majority vote used in this research. For example, a disease-free decision could require the consensus of all ensemble members. Another possibility is to explicitly include prior probability and misclassification costs in the classification objective function. This is feasible for some models like linear discriminant analysis and neural networks but is not possible for others such as logistic regression or k nearest neighbor. It is also possible to intentionally bias the training examples using a stratified sampling strategy.

While we feel the medical data used in this research is representative of diagnostic decision support applications, the reader is cautioned that the conclusions are based on five specific medical domains. More research would be useful to establish whether these results generalize to other medical domains and to areas beyond health care such as bankruptcy prediction and credit scoring. We also acknowledge that there is potential to increase the accuracy of aggregate model with more sophisticated combining rules as well as to the level and degree of diversity of the ensemble of models.

Acknowledgement

The authors thankfully acknowledge the extremely helpful comments, suggestions and guidance of the anonymous referee.

References

[1] K. Ali, M. Pazzani, Error reduction through learning multiple descriptions, Machine Learning 24 (3) (1996).
[2] J.A. Baker, P.J. Kornguth, J.Y. Lo, M.E. Williford, C.E. Floyd,

Breast cancer: prediction with artificial neural network based on bi-rads standardized lexicon, Radiology 196 (1995) 817–822.
[3] J.A. Baker, P.J. Kornguth, J.Y. Lo, C.E. Floyd, Artificial neural network: improving the quality of breast biopsy recommendations, Radiology 198 (1996) 131–135.
[4] W.G. Baxt, Use of an artificial neural network for data analysis in clinical decision-making: the diagnosis of acute coronary occlusion, Neural Computation 2 (1990) 480–489.
[5] W.G. Baxt, Use of an artificial neural network for the diagnosis of myocardial infarction, Annals of Internal Medicine 115 (1991) 843–848.
[6] W.G. Baxt, A neural network trained to identify the presence of myocardial infarction bases: some decisions on clinical associations that differ from accepted clinical teaching, Medical Decision Making 14 (1994) 217–222.
[7] S.D. Bay, Nearest neighbor classification from multiple feature subsets, Intelligent Data Analysis 3 (1999) 191–209.
[8] L. Bottaci, P.J. Drew, J.E. Hartley, M.B. Hadfieldet, Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions, The Lancet 350 (1997) 469–472.
[9] D.G. Bounds, P.J. Lloyd, B.G. Mathew, A comparison of neural network and other pattern recognition approaches to the diagnosis of low back disorders, Neural Networks 3 (1990) 583–591.
[10] L. Breiman, Stacked regressions, Machine Learning 24 (1995) 49–64.
[11] L. Breiman, Bagging predictors, Machine Learning 26 (1996) 123–140.
[12] P. Cunningham, J. Carney, S. Jacob, Stability problems with artificial neural networks and the ensemble solution, Artificial Intelligence in Medicine 20 (2000) 217–225.
[13] A. Das, N.P. Reddy, J. Narayanan, Hybrid fuzzy logic committee neural networks for recognition of swallow acceleration signals, Computer Methods and Programs in Biomedicine 64 (2) (Feb. 2001) 87–99.
[14] R. Detrano, A. Janosi, W. Steinbrun, M. Pfisterer, J. Schmid, S. Sandhu, K.H. Guppy, S. Lee, V. Froelicher, International application of a new probability algorithm for the diagnosis of coronary artery disease, The American Journal of Cardiology 64 (1989) 304–310.
[15] T.G. Dietterich, Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, Machine Learning 40 (2) (August 2000) 139–157.
[16] T.G. Dietterich, Ensemble methods in machine learning, in: J. Kittler, F. Roli (Eds.), First International Workshop on Multiple Classifier Systems, Springer Verlag, New York, 2000, pp. 1–15.
[17] M. Dougados, S. vander Linden, R. Juhlin, B. Huitfeldt, B. Amor, A. Calin, A. Cats, B. Dijkmans, I. Olivieri, G. Pasero, E. Veys, H. Zeidler, The European Spondylarthropathy Study Group preliminary criteria for the classification of spondylarthropathy, Arthritis and Rheumatism 34 (10) (1991) 1218–1230.
[18] R. Feinstein, Multivariable Analysis: An Introduction, Yale Univ. Press, New Haven, 1996.
[19] J. Fricker, Artificial neural networks improve diagnosis of acute myocardial infarction, The Lancet 350 (1997) 935.
[20] E. Gilpin, R. Olshen, H. Henning, J. Ross Jr., Risk prediction after myocardial infarction, Cardiology 70 (1983) 73–84.
[21] B.L. Hubbard, R.J. Gibbons, A.C. Lapeyre, A.R. Zinsmeister, L.P. Clements, Identification of severe coronary artery disease using simple clinical parameters, Archives of Internal Medicine 152 (1992) 309–312.
[22] D. Josefson, Computers beat doctors in interpreting ECGs, British Medical Journal 315 (1997) 764–765.
[23] P.S. Maclin, J. Dempsey, How to improve a neural network for early detection of hepatic cancer, Cancer Letters 77 (1994) 95–101.
[24] R.W. Makuch, P.S. Rosenberg, Identifying prognostic factors in binary outcome data: an application using liver function tests and age to predict liver metastases, Statistics in Medicine 7 (1988) 843–856.
[25] O.L. Mangasarian, W.N. Street, W.H. Wolberg, Breast cancer diagnosis and prognosis via linear programming, Operations Research 43 (4) (July–August 1995) 570–577.
[26] L.J. Mango, Computer-assisted cervical cancer screening using neural networks, Cancer Letters 77 (1994) 155–162.
[27] L.J. Mango, Reducing false negatives in clinical practice: the role of neural network technology, American Journal of Obstetrics and Gynecology 175 (4) (1996) 1114–1119.
[28] R.P. Marble, J.C. Healy, A neural network approach to the diagnosis of morbidity outcomes in trauma care, Artificial Intelligence in Medicine 15 (1999) 299–307.
[29] R.A. Miller, Medical diagnostic decision support systems—past, present, and future: a threaded bibliography and brief commentary, Journal of the American Medical Informatics Association 1 (1) (1994) 8–27.
[30] H. Nomura, S. Kashiwagi, J. Hayashi, W. Kajiyama, H. Ikematsu, A. Noguchi, S. Tani, M. Goto, Prevalence of gallstone disease in a general population of Okinawa, Japan, American Journal of Epidemiology 128 (3) (1988) 598–605.
[31] B. Parmanto, P.W. Munro, H.R. Doyle, Reducing variance of committee prediction with resampling techniques, Connection Science 8 (3–4) (1996) 405–425.
[32] C. Rosenberg, J. Erel, H. Atlan, Neural network that learns to interpret myocardial planar thallium scintigrams, Neural Computation 5 (1993) 492–502.
[33] T.T. Schubert, S.D. Bologna, Y. Nensey, A. Schubert, E.J. Mascha, C.K. Ma, Ulcer risk factors: interactions between *Heliocobacter pylori* infection, nonsteroidal use and age, The American Journal of Medicine 94 (1993) 416–418.
[34] R.S. Sexton, R.E. Dorsey, Reliable classification using neural networks: a genetic algorithm and backpropagation comparison, Decision Support Systems 30 (1) (Dec. 15, 2000) 11–22.
[35] O.R.L. Sheng, Editorial: decision support for healthcare in a new information age, Decision Support Systems 30 (2) (Dec. 27, 2000) 101–103.
[36] D. Sheppard, D. McPhee, C. Darke, B. Shrethra, R. Moore, A. Jurewits, A. Gray, Predicting cytomegalovirus disease after renal transplantation: an artificial neural network approach, International Journal of Medical Informatics 54 (1) (1999) 55–71.

[37] W.H. Tierney, M.D. Murray, D.L. Gaskins, X.H. Zhou, Using computer-based medical records to predict mortality risk for inner-city patients with reactive airways disease, Journal of the Medical Informatics Association 4 (1997) 313–321.
[38] D.M. Titterington, G.D. Murray, L.S. Murray, D.J. Spiegelhalter, A.M. Skene, J.D.F. Habbema, G.J. Gelpke, Comparison of discrimination techniques applied to a complex data set of head injured patients, Journal of the Royal Statistical Society 144 (2) (1981) 145–175.
[39] K.M. Tolle, H. Chen, H. Chow, Estimating drug/plasma concentration levels by applying neural networks to pharmacokinetic data sets, Decision Support Systems 30 (2) (Dec. 27, 2000) 139–151.
[40] G.D. Tourassi, C.E. Floyd, H.D. Sostman, R.E. Coleman, Acute pulmonary embolism: artificial neural network approach for diagnosis, Radiology 189 (1993) 555–558.
[41] O. Tsujji, M.T. Freedman, S.K. Mun, Classification of microcalcifications in digital mammograms using trend-oriented radial basis function neural network, Pattern Recognition 32 (1999) 891–903.
[42] J.V. Tu, M.C. Weinstein, B.J. McNeil, C.D. Naylor, and The Steering Committee of the Cardiac Care Network of Ontario, Predicting mortality after coronary artery bypass surgery: what do artificial neural networks learn? Medical Decision Making 18 (2) (1998) 229–235.
[43] S. Walczak, J.E. Scharf, Reducing surgical patient cost through the use of an artificial neural network to predict transfusion requirements, Decision Support Systems 30 (2) (Dec. 27, 2000) 125–138.
[44] D. West, V. West, Model selection for a medical diagnostic decision support system: a breast cancer detection case, Artificial Intelligence in Medicine 20 (3) (2000) 183–204.
[45] P. Wilding, M.A. Morgan, A.E. Grygotis, M.A. Shoffner, E.F. Rosato, Application of back propagation neural networks to diagnosis of breast and ovarian cancer, Cancer Letters 77 (1994) 145–153.
[46] W.H. Wolberg, W.N. Street, D.M. Heisey, O.L. Mangasarian, Computerized breast cancer diagnosis and prognosis from fine needle aspirates, Archives of Surgery 130 (1995) 511–516.
[47] D. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259.
[48] Y. Wu, M.L. Giger, K. Doi, C.J. Vyborny, R.A. Schmidt, C.E. Metz, Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer, Radiology 187 (1993) 81–87.
[49] J. Zhang, Developing robust non-linear models through bootstrap aggregated neural networks, Neurocomputing 25 (1999) 93–113.
[50] J. Zhang, Inferential estimation of polymer quality using bootstrap aggregated neural networks, Neural Networks 12 (1999) 927–938.
[51] P.A. Zhilkin, R.L. Somorjai, Application of several methods of classification fusion to magnetic resonance spectra, Connection Science 8 (1996) 427–442.

Paul M. Mangiameli is a Professor in the College of Business Administration of the University of Rhode Island where he is the Area Coordinator of Management Information Systems. Dr. Mangiameli's research interests include decision support systems, artificial intelligence in health care, and quality management. He has published in journals such as *Decision Sciences*, *European Journal of Operational Research*, *Omega, Computers and Operations Research*, *Integrated Manufacturing Systems*, *International Journal of Production Research*, *Journal of Operations Management*, and *International Journal of Quality and Reliability Management*.

Rohit Rampal is an Assistant Professor of Management Information Systems at the School of Business Administration, Portland State University, OR. He received his PhD from Oklahoma State University. He has previously worked at the College of Business Administration, University of Rhode Island. His areas of research include telecommunications, information systems in manufacturing, virtual enterprises, DSS, and neural networks. He has published in the *International Journal of Production Research*, *Annals of Cases in Information Technology*, and the *Encyclopedia of Library and Information Science*.

David West is an Associate Professor of Decision Sciences at East Carolina University in Greenville, North Carolina where he teaches operations management and management science. Dr. West received his PhD in Business Administration from the University of Rhode Island. His research interests include the application of neural network technology to such areas as classification decisions, manufacturing process control, and group clustering. He has published in the *European Journal of Operations Research, Computers and Operations Research*, *Decision Support Systems*, *Journal of Quality Technology* and *Omega—The International Journal of Management Science*.