Guidelines to Select Machine Learning Scheme for Classification of Biomedical Datasets

Ajay Kumar Tanwani, Jamal Afridi, M. Zubair Shafiq, and Muddassar Farooq

Next Generation Intelligent Networks Research Center (nexGIN RC)
National University of Computer & Emerging Sciences (FAST-NU)
Islamabad, Pakistan
{ajay.tanwani,jamal.afridi,zubair.shafiq,muddassar.farooq}
@nexginrc.org

Abstract. Biomedical datasets pose a unique challenge to machine learning and data mining algorithms for classification because of their high dimensionality, multiple classes, noisy data and missing values. This paper provides a comprehensive evaluation of a set of diverse machine learning schemes on a number of biomedical datasets. To this end, we follow a four step evaluation methodology: (1) pre-processing the datasets to remove any redundancy, (2) classification of the datasets using six different machine learning algorithms; Naive Bayes (probabilistic), multi-layer perceptron (neural network), SMO (support vector machine), IBk (instance based learner), J48 (decision tree) and RIPPER (rule-based induction), (3) bagging and boosting each algorithm, and (4) combining the best version of each of the base classifiers to make a team of classifiers with stacking and voting techniques. Using this methodology, we have performed experiments on 31 different biomedical datasets. To the best of our knowledge, this is the first study in which such a diverse set of machine learning algorithms are evaluated on so many biomedical datasets. The important outcome of our extensive study is a set of promising guidelines which will help researchers in choosing the best classification scheme for a particular nature of biomedical dataset.

Keywords: Classification, Machine Learning, Biomedical Datasets.

1 Introduction

Recent advancements in the field of machine learning and data mining have enabled biomedical research to play a direct role in improving the general quality of health care. This fact is supported by a large number of applications developed in the field of *biomedical informatics* to provide solutions to a number of real-world problems. The modern research on mass spectrometry based proteomics, genome-wide association, DNA sequencing and microarrays is made possible by the efficient processing of high-dimensional biomedical data. The trend of keeping permanent medical records in the health management information systems is becoming a standard practice in many countries of the world. Moreover, modern medical equipments and diagnostic techniques generate heterogenous and voluminous data [1]. The ill-structured nature of the biomedical data, thus, require intelligent machine learning and data mining algorithms for automated analysis in order to make logical inferences from the stored raw data.

C. Pizzuti, M.D. Ritchie, and M. Giacobini (Eds.): EvoBIO 2009, LNCS 5483, pp. 128–139, 2009.

A diverse set of machine learning and data mining algorithms have been previously used to extract useful information from the biomedical data. These algorithms usually perform regression, clustering, visualization or classification of the biomedical data in order to assist the medical consultants in the decision making process[1]. The well-known machine learning and data mining based classification algorithms use probabilistic methods, rule-based learners, linear models such as neural networks and support vector machines, decision trees and instance-based learners. Further, a combination of different classification algorithms can result in improved classification accuracy [5]. The commonly used ensemble techniques are *bagging*, *boosting*, *voting* and *stacking*. The use of evolutionary algorithms in recent years is also gaining popularity for discovering knowledge in medical diagnoses [2]. However, their evaluation is beyond the scope of this paper.

Despite the great work and diversity in the existing machine learning schemes, no significant work is done so far to assist a researcher in selecting a suitable classification technique for a particular nature of biomedical dataset. In this paper, we provide a comprehensive empirical study on classification of 31 different biomedical datasets using a diverse set of machine learning schemes. We adopt a four step methodology to ensure a comprehensive evaluation of different machine learning schemes: (1) preprocessing the dataset using attribute selection, (2) providing the preprocessed features' set to six well-known classification algorithms, (3) bagging and boosting each of these classifiers, and (4) creating an ensemble of classifiers by using stacking and voting.

The main subject of this paper is to provide a systematic and unbiased evaluation of the existing machine learning schemes to resolve the uncertainties associated with the choice of classifier and the nature of biomedical data. We follow a question oriented research methodology to resolve a number of pertinent questions like: (1) Can the predictive results of classification be improved by diversity in machine learning schemes or is it largely a function of the dataset under consideration?, (2) What is the significance of the nature of biomedical dataset on classification accuracy?, (3) How various parameters of the dataset (instances, classes, missing values, number of attributes, type of attributes) affect the accuracy of classification?, (4) How the choice of a machine learning scheme affects the classification accuracy?, and (5) Which machine learning schemes are more useful and in what cases? The answers subsequently lead us to propose a number of guidelines that we believe will provide valuable support to researchers working on the classification of biomedical datasets.

Organization of the Paper. In the next section, we provide a brief review of the related work. In Section 3, we discuss the biomedical datasets used in our study. We present a review of our classification schemes in Section 4. In Section 5, we report the results of our experiments which are followed by the standard guidelines. Finally, we conclude the paper with an outlook to our future work in Section 6.

2 Related Work

We now provide a brief overview of recent research done to analyze the accuracy of different machine learning schemes on various biomedical domains. In [3], the authors

[1] The scope of this paper is confined to the classical classification problem for prognosis.

study the impact of feature selection on the classification accuracy using an email and a drug discovery dataset. The authors in [4] present an empirical study of bagging and boosting techniques using neural networks and decision trees on 23 randomly chosen datasets. The results of their study suggest that bagging provides relatively better accuracy compared with each of the individual classifiers whereas boosting produces inconsistent results. The work in [5] evaluates the accuracy of different ensemble combinations of six classification algorithms (LDA, 1-NN, Decision Tree, Logistic Regression, Linear SVMs and MLP) on high-dimensional cancer proteomic datasets. In [6], the authors compare the performance of data mining schemes with the logistic and regression techniques on a clinical database of cancer patients. Their results show that pre-processing the data by attribute selection significantly improves the performance of a classifier while meta-learning is of little value. The study of machine learning methods on Atherosclerosis in [7] involves testing of different categories of machine learning schemes to predict future disorders and death causes. A comprehensive survey of biomedical applications utilizing machine learning schemes is done in [8].

The commonly observed methodology among medical researchers in various papers is to experiment on the dataset with only limited number of algorithms from the machine learning repository and select the one which gives relatively better results for their particular domain. The selection of machine learning algorithms for a particular domain appears to be inclined towards their own view of a particular scheme. Consequently, no guidelines are available to select the best classifier for a particular type of data. In our study, we provide a set of guidelines that will help a researcher in choosing an appropriate classifier based on a particular type of dataset.

3 Biomedical Datasets

Biomedical datasets are generally associated with high-dimensional features and multiple classes. The datasets obtained from clinical databases contain various systemic and human errors [9]. The noisy nature, sparseness and missing values hamper the classification accuracy of the machine learning schemes. These inconsistencies demand to treat the classification problem of biomedical datasets as a separate domain. To comprehensively evaluate the performance of various classification schemes on biomedical datasets, we have selected as many as 31 biomedical datasets publicly available from the UCI Machine Learning repository [10] and Center for Cancer Research [11]. Our selection criterion is to choose well-known datasets from a number of different biomedical domains. The summary of the datasets used in our study is shown in Table 1.

Our repository contains high-dimensional datasets (Ovarian 8-7-02 has a total of $15,154$ attributes), multi-class datasets (Thyroid0387 has a total of 32 classes followed by Cardiac Arrhythmia with 16 classes), imbalanced datasets (class distribution of Hyperthyroid, Cardiac Arrhythmia and Cleveland Heart is highly uneven), datasets with many instances (Protein Data contains $21,618$ instances), datasets with missing values (Hungarian Heart and Horse Colic contains up to 20 percent missing values) and datasets of DNA sequencing and mass spectrometry. We believe that the chosen datasets, therefore, encompass all important domains of biomedicine and bioinformatics.

Table 1. The summary of used datasets: The table shows the name of datasets in the alphabetical order; their year of donation; total number of instances; total classes; number of continuous, binary and nominal attributes; and the percentage of missing values in the attributes

Dataset	Year	Instances	Classes	Attributes			Missing
				Continuous	Binary	Nominal	Values (%)
Ann-Thyroid	1987	7200	3	6	15	0	0
Breast Cancer	1992	699	2	1	0	9	0.23
Breast Cancer Diagnostic	1995	569	2	31	0	0	0
Breast Cancer Prognostic	1995	198	2	33	0	0	0.06
Cardiac Arrhythmia	1998	452	16	272	7	0	0.32
Cleveland-Heart	1990	303	5	10	3	0	0.15
Contraceptive Method	1997	1473	3	2	3	4	0
Dermatology	1998	366	6	1	1	32	0.06
Echocardiogram	1989	132	2	8	2	2	4.67
E-Coli	1996	336	8	7	0	1	0
Haberman's Survival	1999	306	3	3	0	0	0
Hepatitis	1988	155	2	6	0	13	5.67
Horse Colic	1989	368	2	8	4	15	19.39
Hungarian Heart	1991	294	5	10	3	0	20.46
Hyper Thyroid	1989	3772	5	7	21	1	2.17
Hypo-Thyroid	1990	3163	2	7	18	0	6.74
Liver Disorders	1990	345	2	6	0	0	0
Lung Cancer	1992	32	3	0	0	56	0.28
Lymph Nodes	1988	148	4	3	9	6	0
Mammographic Masses	2007	961	2	1	0	4	3.37
New Thyroid	1992	215	3	5	0	0	0
Ovarian 8-7-02	2002	253	2	15154	0	0	0
Pima Indians Diabetes	1990	768	2	8	0	0	0
Post Operative Patient	1993	90	3	0	0	8	0.44
Promoters Genes Sequence	1990	106	2	0	0	58	0
Protein Data	-	21618	3	0	0	1	0
Sick	1989	2800	2	7	21	1	2.24
Statlog Heart	-	270	2	7	3	3	0
Switzerland Heart	1991	123	5	10	3	0	17.07
Thyroid0387	1992	9172	32	7	21	1	5.50
Splice-Junction Gene Sequence	1992	3190	3	0	0	61	0

4 A Review of Classification Schemes

We adopt a four step evaluation methodology to ensure an unbiased evaluation of different machine learning schemes: (1) preprocessing the dataset using attribute selection to remove redundant and useless features, (2) providing the preprocessed features' set to six well-known classification algorithms, (3) bagging and boosting each of these classifiers to analyze their merits in improving the accuracy, and (4) finally creating a team of classifiers by combining the the best version (individual, bagged and boosted) of each of the six base classifiers using stacking and voting in order to further enhance the accuracy. We use the standard implementations of these schemes in Wakaito Environment for Knowledge Acquisition (WEKA) [12].

4.1 Data Pre-processing

The attribute selection technique [13] is used as a pre-processing filter to remove the redundant or useless features in the dataset. We use Best First algorithm for the attribute selection that performs greedy hill climbing with a backtracking search method [12].

4.2 Base Classifiers

Naive Bayes. Naive Bayes (NB) utilizes a probabilistic method for classification by multiplying the individual probabilities of every attribute-value pair [14]. This simple algorithm assumes independence among the attributes and even then provides excellent classification results.

Neural Networks using Multi Layer Perceptron. The Multi Layer Perceptron (MLP) consists of input layer (attributes), output layer (classes) and hidden layer(s) that are interconnected through various neurons. The back propagation algorithm tends to optimize the weights of these connections through training instances of the dataset [15]. We have used default parameters for MLP in WEKA. The number of epochs is equal to 500, the learning rate is 0.3 and the momentum of updating weights is 0.2.

Support Vector Machines using Sequential Minimal Optimization. The Support Vector Machine (SVM) algorithm builds a hyperplane to separate different instances into their respective classes [18]. A pairwise classification scheme is used to do multi-class classification. We use Sequential Minimal Optimization (SMO) which is a fast and an efficient version of SVM implemented in WEKA.

Instance Based Learner. The Instance Based Classifier (IBk) is the simplest among the algorithms used in our study [16]. The classification is done on the basis of a majority vote of k neighboring instances. We have used k=5 while taking default values of WEKA for rest of the parameters. The window size is zero that allows maximum number of instances in the training pool without replacements.

Decision Tree. The decision tree (J48) is an implementation of C4.5 in WEKA. The tree comprises of nodes (attributes) at every stage that are structured with the help of training examples [17].

Inductive Rule Learner. Repeated Incremental Pruning to Produce Error Reduction (RIPPER) is a propositional rule learner that defines a rule based detection model and seeks to improve it iteratively by using different heuristic techniques [19]. The constructed rule set is then used to classify new instances. We have implemented this rule based system in WEKA using JRIP with default parameters.

4.3 Resampling Based Ensembles

Bagging. Bagging combines the multiple models generated by training a single algorithm on random sub-samples of a given dataset [20]. Unbiased voting is used during the fusion process.

Boosting. Boosting, in contrast to bagging, uses weighted voting to generate more misclassified instances in its successive models [21].

4.4 Meta-learning Based Ensembles

Stacking. Stacking combines the outputs of two or more base-level classifiers by training them with a meta-learner [22]. In all of our experiments, we use Naive Bayes as

a standard meta-learner while the ensemble comprises of the best version (individual, bagging and boosting) of each of the six different base classifiers.

Voting. Voting is a meta-learning technique that uses different combinations of probability estimates of base classifiers for classification [23]. The selection criteria for choosing the six base learners for voting is same as that of stacking. We have implemented voting in WEKA using an average of the probabilities as the combination rule.

5 Experiments, Results and Guidelines

We now report the results of our experiments that we have done to analyze the classification accuracy of 20 different types of machine learning algorithms on 31 different biomedical datasets. We use Area Under an ROC (Receiver Operating Characteristic) Curve (AUC) ($0 \leq$ AUC ≤ 1) metric to quantify the classification accuracy of an algorithm. AUC is known to be a 'more complete' performance metric as compared to other traditionally used metrics [24], [25]. The ROC curves are generated by varying the threshold on output class probability. AUC = 100% represents the best accuracy while AUC = 0% represents the worst accuracy. The results in Table 2 show the mean AUCs of the machine learning algorithms used in our comparative study on biomedical datasets. Some of the experiments could not be completed even after running for several days and are indicated by blank spaces in our results. We now present our analysis and important insights on the basis of the results obtained from these experiments. Our primary motivation is to to investigate the factors that can potentially affect the classification accuracy of a particular machine learning scheme. The main variables that determine the classification accuracy are categorized by: (1) the nature of a dataset, (2) the pre-processing filter, and (3) the choice of a classification scheme.

5.1 How Does the Nature of a Dataset Affect the Classification Accuracy?

The classification accuracy of a given algorithm is largely dependent on the nature of dataset rather than the algorithm itself. The main characteristics of a dataset are its attributes, classes and number of instances. We answer the following pertinent questions to systematically study the nature of a dataset.

Role of Attributes. The attributes of a dataset vary in terms of their quality, number and type (continuous, binary or nominal). The quality of information that attributes can provide is an important factor that determines the *classification potential* of a dataset. The quality of information can be quantified using well-known parameters like information gain, entropy, gain ratio etc. We use information gain in our study. The results of our experiments demonstrate that the classification accuracy is directly proportional to the information gain of a dataset. The low information gain of datasets like Protein Data, Liver Disorders and Haberman's Survival etc is mainly responsible for relatively poor classification accuracy of all algorithms on them. For example, the Protein Data dataset has only one attribute with an information gain of just 0.0647 which results in the best mean AUC value of only 63.27% among all the applied machine learning schemes.

Table 2. Mean AUCs for biomedical datasets using a diverse set of machine learning algorithms. The bold entries in every row represent the best results. The blank spaces are used to show the missing results. The acronym (Bag) is used for *bagging* while (Bos) represents *boosting*.

Dataset	Individual Classifiers																		Team of Classifiers	
	NB			MLP			SMO			IB*k*			J48			JRIP				
	NB	NB (Bos)	NB (Bag)	MLP	MLP (Bos)	MLP (Bag)	SMO	SMO (Bos)	SMO (Bag)	IB*k*	IB*k* (Bos)	IB*k* (Bag)	J48	J48 (Bos)	J48 (Bag)	JRIP	JRIP (Bos)	JRIP (Bag)	Stacking	Voting
Ann-Thyroid	96.52	86.24	96.31	99.08	91.29	99.13	62.21	86.81	65.01	94.26	90.11	95.98	99.12	99.46	**99.59**	98.99	99.44	99.55	99.28	99.49
Breast Cancer	98.53	98.35	98.80	98.65	97.56	99.00	96.29	96.88	96.96	99.11	97.95	**99.16**	95.85	98.74	98.74	96.51	98.26	98.69	97.62	99.19
Breast Cancer Diagnostic	98.30	98.20	98.70	98.30	97.50	**99.20**	95.60	98.30	96.70	98.70	96.90	98.80	94.20	98.90	98.40	94.60	99.00	98.50	98.20	99.10
Breast Cancer Prognostic	73.20	70.20	**73.70**	70.60	64.30	72.90	50.00	66.60	49.70	68.00	64.10	67.40	53.80	69.00	66.20	59.80	65.70	69.50	70.10	73.40
Cardiac Arrhythmia	66.36	70.98	66.75	70.19	72.89	77.38	65.52	53.57	67.15	73.91	69.86	72.56	66.13	84.92	75.40	64.36	83.27	73.45	66.22	**94.78**
Cleveland-Heart	76.40	60.18	76.30	72.10	75.64	74.00	69.88	66.66	70.90	71.34	61.22	71.80	57.22	73.20	72.36	55.66	57.52	66.34	**79.02**	77.26
Contraceptive Method	70.57	65.50	70.20	72.70	68.23	**74.00**	63.40	62.70	65.17	68.93	65.50	69.60	71.20	68.47	71.07	64.17	64.17	68.30	73.17	73.10
Dermatology	**99.75**	99.20	99.70	99.15	98.82	99.63	98.98	99.73	99.62	99.43	97.58	99.50	97.35	98.98	99.07	97.30	99.33	99.23	-	99.82
Echocardiogram	82.75	83.20	83.05	81.70	82.75	79.85	79.70	80.90	79.65	80.20	77.80	81.35	81.45	81.65	82.60	78.05	83.30	79.60	80.45	**83.50**
E-Coli	74.38	75.84	-	82.29	82.29	-	85.99	86.34	-	87.18	77.93	-	77.90	83.22	-	76.89	**91.93**	-	-	-
Haberman's Survival	**69.70**	66.80	68.00	68.60	58.50	68.30	50.00	65.50	52.40	63.50	56.90	65.90	53.10	63.40	66.80	60.40	65.00	64.90	67.30	67.50
Hepatitis	**89.10**	83.74	89.05	85.77	81.87	86.51	67.47	87.37	79.41	81.96	72.05	84.68	65.60	82.16	83.22	67.20	67.20	78.76	88.35	88.64
Horse Colic	86.30	84.10	86.40	85.10	84.30	87.30	81.40	86.40	82.40	86.00	82.60	86.50	86.80	85.90	88.30	83.50	87.20	88.50	89.20	**89.90**
Hungarian Heart	89.90	87.20	90.20	87.60	85.80	89.80	78.40	88.40	81.00	79.30	64.20	84.60	78.90	87.90	87.10	75.00	88.00	86.80	89.95	**90.70**
Hyper Thyroid	92.28	93.30	92.46	90.80	86.70	**95.98**	62.20	92.68	55.96	88.84	70.84	83.12	78.58	94.80	85.26	69.42	93.94	77.70	81.42	96.64
Hypo-Thyroid	97.58	96.91	97.58	98.55	94.20	98.84	64.14	97.82	68.06	88.18	87.46	89.39	95.45	98.14	97.20	95.05	97.29	95.47	98.14	**98.90**
Liver Disorders	57.90	58.70	57.00	59.70	59.00	**61.00**	50.00	58.60	50.00	56.50	56.70	57.50	55.60	55.80	58.50	56.00	59.80	57.20	59.70	60.50
Lung Cancer	95.65	82.37	95.17	90.82	90.82	90.82	73.43	88.89	81.88	93.72	91.54	90.82	72.95	81.64	85.51	74.88	89.13	76.33	86.36	**96.23**
Lymph Nodes	87.40	71.89	88.30	92.70	83.01	95.56	92.70	82.04	92.99	93.94	75.21	**96.25**	74.85	87.15	93.09	71.96	92.70	90.64	71.23	96.23
Mammographic Masses	89.60	86.40	89.50	87.00	85.70	89.60	79.60	85.00	84.70	85.00	82.00	86.30	85.80	86.80	88.80	84.70	89.00	88.60	89.70	**89.80**
New Thyroid	99.55	99.67	99.63	99.70	98.80	99.67	89.37	99.43	93.63	97.93	97.20	97.90	91.00	99.50	98.00	91.87	97.47	99.27	99.03	**99.77**
Ovarian 8-7-02	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	97.10	97.12	98.77	94.80	97.52	98.77	**100**	-
Pima Indians Diabetes	**82.90**	79.60	82.80	80.50	79.20	82.50	71.50	75.60	74.10	77.90	69.50	79.40	79.10	77.90	80.90	72.00	78.10	77.70	82.45	82.80
Post Operative Patient	31.53	45.50	31.96	45.20	**55.50**	44.07	48.93	39.90	47.20	34.50	42.43	36.23	33.17	36.67	32.70	33.53	33.20	32.60	29.40	44.40
Promoters Genes Sequence	93.40	95.39	**97.90**	**97.90**	**97.90**	97.80	95.30	95.28	95.20	96.60	94.87	97.20	87.70	97.69	93.10	83.00	96.01	97.70	95.90	-
Protein Data	63.17	54.47	63.27	61.83	56.70	63.10	54.00	54.47	55.23	63.17	54.47	**63.27**	50.00	50.00	50.00	50.00	50.00	50.00	-	62.80
Sick	93.09	88.55	93.21	**94.88**	90.83	94.38	49.98	91.42	54.54	89.40	86.55	86.82	90.05	92.48	91.22	89.20	92.57	91.26	92.83	94.57
Statlog Heart	89.15	86.10	89.38	84.17	87.54	87.83	84.42	84.62	86.55	85.36	81.49	86.18	80.16	85.08	89.08	78.88	85.02	88.60	88.49	**90.37**
Switzerland Heart	47.44	47.44	47.88	53.72	53.72	53.96	53.94	53.94	48.04	46.76	47.14	47.42	47.30	47.30	47.56	41.64	41.64	47.66	**59.44**	49.12
Thyroid0387	**89.60**	-	-	79.86	-	-	73.87	-	-	75.91	-	-	85.21	-	-	75.09	-	-	-	-
Splice-Junction Gene Sequence	**98.78**	96.49	**98.78**	98.01	96.03	98.56	95.94	96.94	97.49	96.85	93.02	96.99	93.98	98.03	98.02	95.48	97.75	97.94	98.25	98.77
Mean	83.34	80.57	83.52	83.62	81.90	84.85	73.25	80.57	74.88	81.35	76.79	81.78	76.32	82.03	81.95	75.1	80.98	80.68	82.63	**85.08**

Table 3. Classification differences with and without attribute selection as pre-processor. Bold entries in every row represent the best accuracy.

Dataset	With Pre-Processing			Without Pre-Processing		
	Total Attributes	Net Information Gain	Best Mean AUC	Total Attributes	Net Information Gain	Best Mean AUC
Liver Disorder	1	0.051	61.00	6	0.057	**75.40**
Haberman's Survival	1	0.072	69.70	3	0.072	**71.20**

Guideline 1: Use information gain to quantify the quality of attributes in order to determine the classification potential of a dataset.

Role of output Classes. The multiple output classes lead to imbalanced datasets when the class distribution is not even. Our experiments reveal that the multiclass imbalanced datasets pose a significant challenge in terms of the classification accuracy. For example, the class distributions of Cardiac Arrhythmia (16 classes) and Cleveland Heart (5 classes) datasets are highly imbalanced in favor of one class that logically results in their relatively low mean AUC values. However, the accuracy significantly improves if we deploy a team of classifiers. For example, in case of Cardiac Arrhythmia dataset, the classification accuracy improves from best mean AUC value of 84.92% obtained with all the individual classifiers to 94.78% when the meta-learning technique of voting is used. In comparison, for Cleveland Heart dataset, the best mean AUC value increases from 76.4% to 79.02% when stacking is used.

Guideline 2: Use a team of classifiers for multi-class imbalanced datasets.

Role of Instances. The number of instances, however, have little role on the classification accuracy of algorithms. It is the quality of instances quantified with the help of information gain, which determines the classification potential of a dataset. For example, the Lung Cancer dataset has only 32 instances compared to $21,618$ instances of Protein Data dataset. However, the best mean AUC for the former is 95.95% while for the later it is just 63.27%. The information gain for both the datasets are respectively 1.521 and 0.0647. This proves our thesis that the large AUC for Lung Cancer dataset even with small instances is due to the large information gain of its attributes.

Guideline 3: Do not contemplate on the classification potential of a dataset on the basis of its number of instances only.

5.2 When to Use the Pre-processing Filter?

The attribute selection is used as a pre-processer to remove the redundant and useless attributes in a dataset. The pre-processing filter in most of the cases improves the classification accuracy of datasets with the exception of few ones. Therefore, it is important to identify when to use a pre-processing filter. Our study again suggests that the decision should be based on the information gain of attributes. If the net information gain of a dataset is small or the number of attributes become too less after pre-processing, then the pre-processing filter should not be used. In Table 3, we report the results with and without pre-processing filter on two of the datasets (Liver Disorder and Haberman's Survival) which are relatively challenging for classification. The results prove our hypothesis that we should not use a pre-processing filter if it further degrades the quality of attributes.

Table 4. Mean AUCs and standard deviations of six base classifiers over all datasets used in this study. Bold entries in every row represent the best accuracy.

Classification Scheme	NB	MLP	SMO	IB*k*	J48	JRIP	Mean
Individual	83.34 ± 16.99	**83.62 ± 15.19**	73.25 ± 17.15	81.35 ± 16.54	76.32 ± 17.68	75.1 ± 17.47	78.83 ± 16.84
Bagging	83.52 ± 16.05	**84.85 ± 14.57**	74.88 ± 16.89	81.78 ± 16.32	81.95 ± 17.18	80.67± 18.73	81.28 ± 16.62
Boosting	80.57 ± 22.88	81.90 ± 21.67	80.57 ± 22.26	76.79 ± 22.05	**82.03 ± 22.83**	80.98 ± 22.95	80.48± 22.44
Mean	82.48 ± 18.64	**83.46 ± 17.14**	76.23 ± 18.77	79.97 ± 18.31	80.10 ± 19.23	78.92 ± 19.72	80.20± 18.63

Guideline 4: Do not use attribute selection as a pre-processor filter on the datasets if : (1) they have low quality information attributes, or (2) the remaining attributes after the preprocessing are too less to be of any value.

5.3 How Does the Machine Learning Scheme Affect the Classification Accuracy?

In this section, we analyze the effect of different machine learning algorithms on classification accuracy of a dataset.

Resampling based Ensembles vs Individual Classifiers? Resampling Based Ensemble techniques are preferable over individual classifiers because the final classification is done by training the algorithm on different regions of the sample space. As a result, these ensembles reduce the over fitting bias of an algorithm. Table 4 provides the net mean AUC's of six base classifiers with resampling based ensembles over all the datasets. It is clear that combining multiple resampling methods for classifier enhancement (such as bagging or boosting) are generally more effective than the individual classifier. Moreover, it is only 30 out of 174 times (17.24%) when a single classifier produced better accuracies than the respective bagging and boosting models of the classifiers. Our results show that the overall mean AUC of resampling based ensembles is 80.86% compared to that of 78.83% for individual classifiers.
Guideline 5: Use resampling based classifier enhancement techniques (bagging and boosting) over individual classifiers.

When is Bagging particularly useful? Bagging neutralizes the instability of algorithms by using unbiased voting procedure for combining multiple samples [12]. This explains the reason behind the better average AUCs of bagging for all the individual classifiers. We can see in Table 4 that average AUC for bagging is 81.28% compared with 80.48% of boosting and 78.83% of the individual classifiers. Moreover, our results show that 130 out of 179 times (72.63 %) bagging has improved the accuracy of individual classifiers. These insights support our argument bagging is a particularly useful technique for classifier enhancement.
Guideline 6: Use bagging as classifier enhancement to improve the classification accuracy of the individual algorithms.

When is Boosting particularly useful? The biased voting and weighted selection of instances in boosting often gives inconsistent results compared with those of bagging or individual classification schemes. The reason is that boosting over fits on noisy datasets

[4]. The unpredictable behavior of boosting often leads to significantly low AUCs for unstable algorithms. For example, boosted IB*k* in Hepatitis dataset decreases the mean AUC value of individual IB*k* from 81.96% to 72.05%. Similarly, for the Hyperthyroid dataset, the mean AUC value of boosted IB*k* is 70.84% compared with AUC value of 88.84% for the individual classifier. In comparison, boosting significantly improves the AUC values for stable algorithms. For example, boosted SMO in Sick dataset increases the AUC of SMO from 49.98% to 91.42%; boosted JRIP in Cardiac Arrhythmia dataset increases the AUC of JRIP from 64.36% to 83.27%; and boosted J48 in Hyperthyroid dataset increases the mean AUC value of J48 from 78.58% to 94.80%. We can see in Table 2 that the improvements due to boosting on SMO, JRIP and J48 are scalable to other datasets as well. Boosting is particulary suited for SMO because its average AUC values are 80.57% compared with 74.88% of bagged SMO and simple SMO of 73.25%.
***Guideline 7: Use boosting on stable algorithms like SMO, JRIP, and J48 and do not use it on unstable algorithms like MLP and IB*k.**

Bagging Naive Bayes vs Individual Naive Bayes? Naive Bayes results are excellent for datasets like Haberman, Hepatitis, Ovarian 8-7-02, Pima Indian Diabetes and Splice Junction Gene Sequencing. The classification accuracy of bagging and simple Naive Bayes is in general better than the boosted Naive Bayes. Therefore, it becomes relevant to have a guideline when to enhance Naive Bayes with bagging? The problem can be analyzed by dividing the *significant attributes* (the attributes after the attribute selection phase) in two groups: (1) continuous and multinominal attributes having more than *n* values, and (2) multinominal attributes having less than *n* values. If the net information gain of the first group is greater than that of the second group, then use bagging Naive Bayes. This conjecture works well with *n = 4*. For example, the significant attributes in Hyperthyroid dataset comprise of 3 continuous and 2 binary valued attributes and the information gain distribution is: (1) *total information gain of multinominal attributes with less than 4 values = 0.0335*, and (2) *total information gain of other remaining attributes = 0.14*. The results in Table 2 show that the classification accuracy increases from 92.28% to 92.46% in favor of bagging Naive Bayes compared to the individual Naive Bayes. In a similar way, the information gain distribution of Cleveland Heart dataset after attribute selection is: (1) *total information gain of multinominal attributes with less than 4 values = 0.847*, and (2) *total information gain of other remaining attributes = 0.347*. It can be seen in Table 2 that individual Naive Bayes proved to be better in this case with a mean AUC of 76.4% compared to 76.3% obtained from bagging Naive Bayes. The datasets like Breast Cancer Prognostic, Breast Cancer Diagnostic, Lung Cancer, Contraceptive Method etc. all support this conjecture.
Guideline 8: Use bagged version of Naive Bayes instead of individual one only if after attribute selection, the net information gain of continuous and multinominal attributes with more than* n *values (*n = 4*) is greater than the information gain of multinominal attributes with less than* n *values.

Meta-Learning Based Ensembles - Voting vs Stacking? The criterion that we use to select the base classifiers for making a good team of classifiers is based on both their diversity and accuracy. We choose the best version among individual classifier, their bagged and boosted version for each of the six different individual classifiers, and

use Naive Bayes as a meta-learner to produce a meta-learning ensemble. Our experiments demonstrate that stacking in general do not improve the classification accuracy of medical datasets. The mean AUC values of stacking are comparable with those of other techniques. On the other hand, the classification accuracy of voting is much better than those of all other classification techniques with an overall average AUC value of 85.08%.

Guideline 9: Use voting instead of stacking for meta-learning ensembles to achieve better AUC values.

Which classification algorithm is the best? We choose the best classification algorithm on two parameters: (1) overall classification accuracy, and (2) variance in accuracy that determines the stability and consistency of an algorithm. We can see from Table 4 that *Bagging MLP not only gives on the average the best overall classification accuracy with an AUC value of* 84.85% *but also the least standard deviation of* 14.57.

Guideline 10: Use bagging MLP for classification if the nature of a biomedical dataset is unknown.

6 Conclusion

In this paper, we have presented a comprehensive empirical study of a diverse set of machine learning algorithms on a large number of biomedical datasets. The diversity is added by using resampling based ensemble methods of bagging and boosting and meta-learning techniques of stacking and voting. We conclude that the nature of a given dataset plays an important role on the classification accuracy of algorithms; therefore, it is imperative to choose an appropriate algorithm for a particular dataset. We have identified some general characteristics of a dataset that can be useful in selecting the most suitable algorithm as per the nature of underlying dataset. We have also evaluated the performance of various machine learning schemes under different scenarios to study the effect of diversity on the classification results. The results of our experiments show that voting in general is the most powerful technique among the compared machine learning schemes. On the basis of our study, we have been able to formulate 10 generic guidelines that can help researchers of biomedical classification community to select an appropriate classifier for their particular problem. In future, we would like to devise a metaheuristic framework that can recommend the most suitable classifier for the dataset by analyzing the patterns in the dataset.

Acknowledgements. The authors of this paper are supported, in part, by the National ICT R&D Fund, Ministry of Information Technology, Government of Pakistan. The information, data, comments, and views detailed herein may not necessarily reflect the endorsements of views of the National ICT R&D Fund.

References

1. Wasan, S., Bhatnagar, V., Kaur, H.: The impact of data mining techniques on medical diagnostics. Data Science Journal 5, 119–126 (2006)
2. Pena-Reyes, C.A., Sipper, M.: Evolutionary computation in medicine: an overview. Journal of Artificial Intelligence in Medicine 19(1), 1–23 (2000)

3. Janecek, A.G.K., Gansterer, W.N., Demel, M.A., Ecker, G.F.: On the relationship between feature selection and classification accuracy. Journal of Machine Learning and Research 4, 90–105 (2008)
4. Opitz, D., Maclin, R.: Popular ensemble methods: an empirical study. Journal of Artificial Intelligence Research 11, 169–198 (1999)
5. Assareh, A., Moradi, M.H., Volkert, L.G.: A hybrid random subspace classifier fusion approach for protein mass spectra classification. In: Marchiori, E., Moore, J.H. (eds.) EvoBIO 2008. LNCS, vol. 4973, pp. 1–11. Springer, Heidelberg (2008)
6. Hayward, J., Alvarez, S., Ruiz, C., Sullivan, M., Tseng, J., Whalen, G.: Knowledge discovery in clinical performance of cancer patients. In: IEEE International Conference on Bioinformatics and Biomedicine, USA, pp. 51–58 (2008)
7. Serrano, J.I., Tomeckova, M., Zvarova, J.: Machine learning methods for knowledge discovery in medical data on Atherosclerosis. European Journal for Biomedical Informatics 2(1), 6–33 (2006)
8. Kononenko, I.: Machine learning for medical diagnosis: History, state of the art and perspective. Artificial Intelligence in Medicine 23(1), 89–109 (1995)
9. Lavrac, N.: Selected techniques for data mining in medicine. Artificial Intelligence in Medicine 16, 3–23 (1999)
10. UCI repository of machine learning databases, University of California-Irvine, Department of Information and Computer Science, `www.ics.uci.edu/~mlearn/MLRepository.html`
11. Ovarian cancer studies, center for cancer research, National Cancer Institute, USA, `http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp`
12. Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
13. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning and Research 3, 1157–1182 (2003)
14. Rish, I.: An empirical study of the naive Bayes classifier. In: IJCAI Workshop on Empirical Methods in Artifical Intelligence, pp. 41–46 (2001)
15. Haykin, S.: Neural networks: a comprehensive foundation, 2nd edn. Pearson Education, London (1998)
16. Aha, D.W., Kibler, D., Albert, M.K.: Instance based learning algorithms. Machine Learning 6(1), 37–66 (1991)
17. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann, San Francisco (1993)
18. Vapnik, V.N.: Statistical learning theory. Wiley Interscience, USA (1998)
19. Cohen, W.W.: Fast effective rule induction. In: Proceedings of Twelfth International Conference on Machine Learning, USA, pp. 115–123 (1995)
20. Breiman, L.: Bagging Predictors. Machine Learning 24(2), 123–140 (1996)
21. Freund, Y., Schapire, R.: Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on Machine Learning, Italy, pp. 148–156 (1996)
22. Ting, K.M., Witten, I.H.: Stacked generalization: when does it work. In: Proceedings of the Fifteenth IJCAI, pp. 866–871. Morgan Kaufmann, San Francisco (1997)
23. Abe, H., Yamaguchi, T.: Constructive meta-learning with machine learning method repository. In: Orchard, B., Yang, C., Ali, M. (eds.) IEA/AIE 2004. LNCS, vol. 3029, pp. 502–511. Springer, Heidelberg (2004)
24. Fawcett, T.: ROC graphs: notes and practical considerations for researchers, TR HPL-2003-4, HP Labs, USA (2004)
25. Walter, S.D.: The partial area under the summary ROC curve. Statistics in Medicine 24(13), 2025–2040 (2005)