journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms

Akin Ozcift[a,*], *Arif Gulten*[b]

[a] *University of Gaziantep, Gaziantep Vocational School of Higher Education, Computer Programming Division, Gaziantep, Turkey*
[b] *Firat University, Engineering Faculty, Electrical-Electronics Department, Elazig, Turkey*

ARTICLE INFO

Article history:
Received 18 March 2010
Received in revised form
28 March 2011
Accepted 31 March 2011

Keywords:
Rotation forest
Ensemble learning
Classifier performance
Parkinson's
Diabetes
Cleveland heart
Computer aided diagnosis

ABSTRACT

Improving accuracies of machine learning algorithms is vital in designing high performance computer-aided diagnosis (CADx) systems. Researches have shown that a base classifier performance might be enhanced by ensemble classification strategies. In this study, we construct rotation forest (RF) ensemble classifiers of 30 machine learning algorithms to evaluate their classification performances using Parkinson's, diabetes and heart diseases from literature.

While making experiments, first the feature dimension of three datasets is reduced using correlation based feature selection (CFS) algorithm. Second, classification performances of 30 machine learning algorithms are calculated for three datasets. Third, 30 classifier ensembles are constructed based on RF algorithm to assess performances of respective classifiers with the same disease data. All the experiments are carried out with leave-one-out validation strategy and the performances of the 60 algorithms are evaluated using three metrics; classification accuracy (ACC), kappa error (KE) and area under the receiver operating characteristic (ROC) curve (AUC).

Base classifiers succeeded 72.15%, 77.52% and 84.43% average accuracies for diabetes, heart and Parkinson's datasets, respectively. As for RF classifier ensembles, they produced average accuracies of 74.47%, 80.49% and 87.13% for respective diseases.

RF, a newly proposed classifier ensemble algorithm, might be used to improve accuracy of miscellaneous machine learning algorithms to design advanced CADx systems.

1. Introduction

1.1. *Background*

Machine learning algorithms have been successfully applied to design CADx systems. These algorithms are first trained with diagnosed samples, i.e. with precedent diagnoses of medical experts. In the test phase, the algorithms are later used to assist the medical experts in making diagnosis of future samples [1]. In this aspect, success of an analysis strategy can be defined as the ability of algorithm to predict the correct status (normal or disease) of unseen data.

Performance of CADx systems might be enhanced with more accurate machine learning algorithms. Predictive ability of such analysis methods can be improved mainly with two

[*] *Corresponding author*. Tel.: +90 5055753627.
E-mail address: akinozcift@hotmail.com (A. Ozcift).

0169-2607/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2011.03.018

strategies: (i) application of feature selection methods on the dataset [2], (ii) construction of classifier ensembles [3].

Accuracy of classification strategies can be affected negatively with the use of too many features in the classification. This may lead to overfitting, in which noise or irrelevant features may decrease classification accuracy because of the finite size of the training samples [4]. In general, there are two widely used feature selection strategies: (i) filter approaches and (ii) wrappers. Wrapper methods find feature subsets based on the performance of a preselected classification algorithm on a training data set. In contrast, filters rely on properties of the features to select the best feature subset. While selecting a subset of features, both approaches utilize a search procedure such as individual ranking, forward search and backward search [5]. In this concept, CFS is a multivariate filter approach that can evaluate strength of features to return the most relevant variables [6]. CFS, in literature, is used in various medical diagnosis applications for feature selection purposes [7–10].

A powerful technique in machine learning to increase accuracy of conventional base classifiers is to construct classifier ensembles. An ensemble classifier consists of base classifiers that learn a target function by combining their prediction mutually [11]. Some of the ensemble learning approaches seen in the machine learning literature is composite classifier systems, mixture of experts, consensus aggregation, dynamic classifier selection, classifier fusion and committees of neural networks [12]. In machine learning literature, there are various CADx applications that use classifier ensembles (particularly RF algorithm) to improve accuracy of convenient classifiers [13–18].

Other than accuracy of the base classifiers, the performance of an ensemble algorithm is affected by diversity of the community of classifiers forming the ensemble. Diverse classifiers make different errors on different samples. Combination of such classifiers might lead to more accurate decisions [19].

1.2. *Introduction to the proposed system*

This study presents an evaluation study that can help to design CADx systems with increased performance. The strategy based on a two-step approach in constructing classifiers with enhanced accuracy. In the first step, feature dimensions of three benchmarking datasets are reduced by the use of CFS algorithm. In the second step, 30 base classifiers and corresponding RF classifier ensembles are used in diagnosis of Parkinson's, heart and diabetes diseases to evaluate the resultant accuracies of algorithms. All the experiments are validated with leave-one-out (10-fold cross validation) scheme.

2. Materials and methods

2.1. *Overview*

In this section, we explain our technique used for creating classifiers with improved accuracies. First, our feature selection strategy, i.e. CFS, is introduced. Following CFS explanation, RF ensemble classification scheme is explained with detail. Next, the datasets used to evaluate classifier performances are briefly introduced. Section 2 is ended with the explanation of evaluation metrics used through experiments.

2.2. *Variable selection with CFS algorithm*

In a classification problem, goodness of features from *correlation* point of view can be defined as follows: In general, a feature is evaluated as *good* if it is highly correlated to the class but inversely uncorrelated to any of the other features. In *feature selection* aspect, high correlation between the class and a feature means that the feature to be selected is predictive of the class compared to the unselected features [20]. In test theory, a composite test is designed to predict an external variable of interest. In case of feature selection, *features* are assumed to be individual tests that measure trait related to class (variable of interest) [21]. CFS uses a heuristics to select subset of features rather than ranking features individually. The heuristic assigns high scores to feature subsets that are highly correlated with the class and highly uncorrelated with each other. The scores for each subset of features are calculated with Eq. (1).

$$Merit_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{1}$$

In Eq. (1), $Merit_S$ is the *merit* of a feature subset S containing k features, $\overline{r_{cf}}$ the average feature-class correlation, and $\overline{r_{ff}}$ the average feature-feature inter-correlation [22].

In our study, we used CFS algorithm from WEKA machine learning environment with *best-first search* strategy to reduce dimension of Parkinson's, heart and diabetes datasets. The results of variable selection with corresponding features of each dataset will be given in Section 3.

2.3. *Classifier ensembles with rotation forest*

Classifier ensembles are generally more accurate compared to a single base classifier. There are miscellaneous classifier ensemble models in machine learning literature such as boosting and bagging [23]. In bagging, diversity is provided building classifiers independent from one another using a randomized heuristic. Diversity in bagging is provided with further randomization yielding Random Forest ensemble model [24]. This ensemble model uses decision trees trained on bootstrap samples from the data set and it improves diversity with randomizing the feature choice at each node while constructing trees. Similar to Random Forest, RF is built with independent decision trees. However in RF each tree is trained with complete data set with a rotated feature space. As the algorithm builds classifiers use hyperplanes parallel to the feature axes and a small rotation of the axes lead to diverse trees [25]. More explicitly, the structure of the RF algorithm is given as follows:

Let X mean the training sample set and Y the class labels of dataset with F number of features. And If N denotes the number of training instances with n features, then X will become an N by n matrix. Assuming that class labels of Y is taken from set of classes $\{\omega_1, \ldots, \omega_c\}$ denoted by ω. The feature set of dataset is assumed to be partitioned into K subsets and the decision trees numbers of Rotation Forest algorithm is to be L with notation of $\{D_1, \ldots, D_L\}$. The data used in training of base classifier is created with a randomly split *K* feature set [26].

The training set for classifier D_i is handled in three steps:

(i) As a first step, *F* is divided into *K* feature sets randomly with each subset of $M = n/K$ number of features.

(ii) In this second step, let F_{ij} denote the *j*th subset of features to train classifier D_i and X_{ij} be the set of data for F_{ij}, being subset of features. A nonempty random subset is drawn from X_{ij} and then with bootstrap to form a new training set the %75 of this training data is selected as X'. A linear transformation is operated on X' to generate the coefficients of matrix C_{ij}. Each matrix X' has size of $M \times 1$ and the coefficients of this matrix are $a_{ij}^{(1)}, \ldots, a_{ij}^{(M_j)}$.

(iii) In this last step, having obtained the coefficients of matrix C_{ij}, a sparse rotation matrix R_i then formed as given in (2):

$$R_i = \begin{bmatrix} a_{i1}^{(1)}, \ldots, a_{i1}^{(M_1)} & \{0\} & \ldots & \{0\} \\ \{0\} & a_{i2}^{(1)}, \ldots, a_{i2}^{(M_2)} & \ldots & \{0\} \\ \ldots & \ldots & \ldots & \ldots \\ \{0\} & \{0\} & \ldots & a_{iK}^{(1)}, \ldots, a_{iK}^{(M_K)} \end{bmatrix} \quad (2)$$

Here at this point, the columns of R_i is rearranged with respect to the original feature set and the new rotation matrix is represented as R_i^a. The transformed training set for classifier D_i will become XR_i^a. By means of this approach, the classifiers are provided with parallel training [26].

While the classification phase is evaluated for a given instance *x*, let the probability of this instance being classified by classifier D_i to one of classes is denoted with $d_{ij}(xR_i^a)$. From this point, the confidence of a class is calculated by Eq. (3), and *x* is assigned to a class with the largest confidence calculated.

$$\mu_j(x) = \frac{1}{L}\sum_{i=1}^{L} d_{ij}(xR_i^a), \quad j = 1, \ldots, c \quad (3)$$

Rotation Forest algorithm applies Principal Component Analysis (PCA) transformation to each *K* subset to determine principal components that is expected to preserve variability of information in the data. By means of *K* axis rotations, the new features for base classifier are formed. Rotation approach in this method serves the ensemble with accuracy and diversity. In traditional Rotation Forest algorithm, decision trees are chosen for rotation task, because of their sensitivity to rotation of the feature axes. And hence the name 'forest' is inspired from this scheme and the more detailed explanation of algorithm is given in [26].

2.4. *Performance evaluation metrics*

Performance of any classifier needs to be tested with some metric, to assess the result and hence the quality of the algorithm. In our study, to evaluate the results of the experiments of 30 machine learning algorithms and corresponding RF classifier ensembles, we utilized three widely used metrics, i.e. classification accuracy (ACC), Kappa Error (KE) and Area under the Receiver Operating Characteristic Curve (AUC).

Most of the CADx problems deal with two class predictions to map data samples into one of the groups, i.e. benign or malignant. For such a two-class problem, the outcomes are labeled as positive (*p*) or negative (*n*). The possible outcomes with respect to this classification scheme is frequently defined in statistical learning as *true positive* (TP), *false positive* (FP), *true negative* (TN) and *false negative* (FN). These four outcomes are connected to each other with a table that is frequently called as *confusion matrix*. For a binary classification scheme, confusion matrix is used to derive most of the well known performance metrics such as sensitivity, specificity, accuracy, positive prediction value, F-measure, AUC and ROC curve. These metrics are evaluated using the confusion matrix outcomes, i.e. *TP*, *FP*, *TN* and *FN* predictive values [27].

(i) *ACC* is a widely used metric to determine class discrimination ability of classifiers, and it is calculated with Eq. (4).

$$ACC = \frac{(TP + TN)}{(P + N)} \quad (4)$$

ACC is one of primary metrics in evaluating classifier performances and it is defined as the percentage of test samples that are correctly classified by the algorithm.

(ii) AUC, being a valued classification performance measure, is widely used to measure classifier performances with relevant acceptance. AUC value is calculated from the area under the ROC curve. ROC curves are usually plotted using *true positives rate* versus *false positives rate*, as the discrimination threshold of classification algorithm is varied. In this aspect, since a ROC curve compares the classifiers' performance across the entire range of class distributions and error costs; an AUC value is accepted to be a good measure for comparative performances of classification algorithms. In terms of *TP*, *FP*, *TN* and *FN* predictive values, AUC is calculated using Eq. (5).

$$AUC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \quad (5)$$

Since, it is relatively difficult to compare performances of 60 algorithms (30 base classifiers and 30 classifier ensembles) for each dataset in terms of ROC curves; we prefer *AUC* values to *ROC* curves for the sake of convenience.

(iii) *KE* or *Cohen's Kappa Statistics* value is used to support the comparative performances of classifiers. Performance comparisons depending only on accuracy based metrics might produce misleading results. The cost of error must also be taken into account in such assessments. KE, in this aspect, is a good measure to inspect classifications that may be due to chance. In general, KE takes values between (−1,1). As the Kappa value calculated for classifiers approaches to '1', then the performance of the classifier is assumed to be more realistic rather than by chance. Therefore, in the performance analysis of classifiers, KE is a recommended metric to consider for evaluation purposes [28] and it is calculated with Eq. (6).

$$KE = \frac{p_0 - p_c}{1 - p_c} \quad (6)$$

Table 1 – The features of diabetes dataset.

ID	Feature	ID	Feature
1	No. of times pregnant	5	2-h serum insulin (μU/ml)
2	Plasma glucose concentration	6	Body mass index (kg/m^2)
3	Diastolic blood pressure (mm Hg)	7	Diabetes pedigree function
4	Triceps skin fold thickness (mm)	8	Age

Table 2 – The features of Cleveland heart dataset.

ID	Feature	ID	Feature
1	Age	8	Maximum heart rate achieved
2	Sex	9	Exercise induced angina
3	Chest pain type (four values)	10	Old peak = ST depression induced by exercise relative to rest
4	Resting blood pressure	11	The slope of the peak exercise ST segment
5	Serum cholesterol in mg/dl	12	Number of major vessels (0–3) colored by fluoroscopy
6	Fasting blood sugar >120 mg/dl	13	Thal: 3 = normal; 6 = fixed defect and 7 = reversible defect
7	Resting electrocardiographic results (values 0, 1 and 2)		

In Eq. (6), p_0 demonstrates total agreement probability and p_c agreement probability due to chance.

As a testing method, we used leave-one-out strategy of 10-fold cross validation. In this scheme, the classification accuracy is computed 10 times, each time leaving out one of the sub-samples from the computations and using that sub-sample as a test sample for cross-validation. In this structure, each sub-sample is used 9 times in the learning sample and just once as the test sample [29].

3. The benchmarking data with the application of CFS algorithm

We utilized three medical datasets, i.e. diabetes, heart and Parkinson's, from UCI machine learning repository for benchmarking purposes.

The diabetes dataset contains 768 data samples and each sample is defined with 8 features of Table 1. In the dataset, there are two classes as negative to diabetes and positive to diabetes. The two classes involve 500 and 268 samples, respectively.

With the application of CFS algorithm to diabetes dataset, the features with IDs of {2,6,7,8} are retained while the others eliminated.

The Cleveland heart disease dataset contains 303 data samples and each sample is defined with 13 features of Table 2. In the dataset, there are two classes determined with vessel narrowing measure. If the vessel is narrowed less than 50%, then the status is normal (165 samples in this class) and if the vessel diameter is lessened greater than 50% the status is angiographic (138 samples in disease).

As for Cleveland hearth disease, we used CFS algorithm to select the most important features of heart dataset and the we obtained feature IDs of {3,7,8,9,10,12,13}.

The last benchmarking dataset of this study is a neurological disorder, i.e. Parkinson's disease [30]. The dataset contains 22 features and it is composed of a range of voice measurements from 31 people, 23 with Parkinson's disease. Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals. The feature structure of Parkinson's dataset is given in Table 3.

Subsequent to CFS algorithm application to Parkinson's dataset, we obtained features of {1,2,3,6,13,15,18,20,21,22} as the most relevant variables.

4. Machine learning algorithms and their abbreviations used in the study

In order to evaluate the performance of widely used machine learning algorithms with their corresponding RF classifier ensembles, we selected 30 algorithms from Weka data mining software. While selecting the algorithms, we attempted to keep diversity of algorithms. For the ease of evaluation in all of the figures and tables, we make use of ID num-

Table 3 – The features of Parkinson's dataset.

ID	Feature	ID	Feature
1	MDVP Fo (Hz): average vocal fundamental frequency	12	Shimmer APQ5: variation in amplitude
2	MDVP Fhi (Hz): maximum vocal fundamental frequency	13	MDVP APQ: variation in amplitude
3	MDVP Flo (Hz): minimum vocal fundamental frequency	14	Shimmer DDA: variation in amplitude
4	MDVP Jitter (%): variation in fundamental frequency	15	NHR: ratio of noise to tonal components in the voice
5	MDVP Jitter (Abs): variation in fundamental frequency	16	HNR: ratio of noise to tonal components in the voice
6	MDVP RAP: variation in fundamental frequency	17	RPDE: nonlinear dynamical complexity measures
7	MDVP PPQ: variation in fundamental frequency	18	D2: nonlinear dynamical complexity measure
8	Jitter DDP: variation in fundamental frequency	19	DFA: Signal fractal scaling exponent
9	MDVP Shimmer: variation in amplitude	20	spread1: fundamental frequency variation
10	MDVP Shimmer (dB): variation in amplitude	21	spread2: fundamental frequency variation
11	Shimmer APQ3: variation in amplitude	22	PPE: fundamental frequency variation

Table 4 – The classifiers and their respective abbreviations.

No.	Algorithm	No.	Algorithm	No.	Algorithm
1	Bayesian Logistic Regression	11	Fuzzy Lattice Reasoning	21	Decision Table
2	BayesNet	12	HiperPipes	22	Best-first Decision Tree
3	Naive Bayes	13	Voting Feature Intervals	23	Decision Stump
4	Logistic	14	Conjunctive Rule	24	Functional Tree Learner
5	Multi Layer Perceptron	15	JRIP	25	J48
6	RBF Network	16	Nnge	26	Alternating Decision Tree
7	Simple Logistic	17	OneR	27	Logistic Model Trees
8	SMO	18	PART Decision Learner	28	Random Tree
9	KSTAR	19	Ripple-Down Rule learner	29	Fast Decision Tree Learner
10	Locally Weighed Learner	20	ZeroR	30	Simple Chart

ber of the algorithms as a replacement for their names. The ID numbers and respective name of the algorithms are given in Table 4. We used default parameters of classifiers, while carrying out the experiments. Though, Weka supports some other feature transformation techniques for RF algorithm, we used the default method, i.e. Principal Component Analysis, in all of the classifier ensembles. Furthermore, another parameter that might affect the classification performance is the number of classifiers used in the ensemble. We kept the classifier number as 10 through all experiments and we discussed the effect of this point in Section 5.

5. Experimental results

In this section, the results of the experiments for diabetes, Cleveland heart and Parkinson's datasets are given in Tables 5–7, respectively. In the tables, '*e*' means RF classifier ensemble corresponding to base classifier measures. 'Diff' means 'Difference' while 'AVG' stands for 'average'.

As Table 5 is examined with three metrics (ACC, KE and AUC) simultaneously, 24 out of 30 base classifiers' performance is seen to be improved by the use of corresponding RF classifier ensemble models. Moreover, RF classifier ensem-

Table 5 – Diabetes disease classification results.

Alg. ID	ACC (%)	*e*ACC (%)	Diff (%)	KE	*e*KE	AUC	*e*AUC
1	64.19	77.73	13.54	0.42	0.48	0.494	0.738
2	75.52	75.91	0.39	0.44	0.46	0.802	0.828
3	77.47	75.39	−2.08	0.48	0.44	0.829	0.809
4	77.73	77.73	0	0.48	0.48	0.825	0.825
5	75.52	75.91	0.39	0.45	0.46	0.809	0.829
6	76.56	76.95	0.39	0.46	0.46	0.824	0.826
7	77.21	77.34	0.13	0.47	0.47	0.825	0.825
8	76.82	76.69	−0.13	0.45	0.45	0.711	0.727
9	69.92	70.83	0.91	0.31	0.34	0.764	0.762
10	73.43	75.65	2.22	0.41	0.44	0.798	0.813
11	72.43	72.43	0	0.39	0.39	0.706	0.715
12	65.23	65.88	0.65	0.43	0.44	0.506	0.584
13	43.35	53.64	10.29	0.21	0.25	0.550	0.626
14	71.09	75.91	4.82	0.34	0.44	0.697	0.798
15	75.52	76.82	1.3	0.43	0.45	0.709	0.792
16	71.87	75.00	3.13	0.37	0.42	0.684	0.793
17	72.78	74.08	1.3	0.36	0.39	0.67	0.798
18	72.26	76.56	4.3	0.39	0.45	0.787	0.820
19	72.26	75.91	3.65	0.33	0.41	0.655	0.790
20	65.10	65.10	0	0.45	0.45	0.497	0.497
21	73.95	77.47	3.52	0.41	0.48	0.791	0.835
22	73.43	75.65	2.22	0.39	0.43	0.716	0.814
23	71.87	75.78	3.91	0.37	0.44	0.684	0.803
24	76.69	77.34	0.65	0.46	0.47	0.768	0.789
25	74.86	76.82	1.96	0.42	0.46	0.791	0.814
26	74.21	77.09	2.88	0.44	0.46	0.806	0.837
27	77.21	77.73	0.52	0.47	0.48	0.825	0.827
28	66.92	72.52	5.6	0.28	0.36	0.642	0.735
29	74.73	75.91	1.18	0.42	0.47	0.778	0.807
30	74.60	76.56	1.96	0.41	0.47	0.758	0.789
AVG	72.16	74.48	2.32	0.40	0.43	0.723	0.774

Table 6 – Cleveland hearth disease classification results.

Alg. ID	ACC (%)	eACC (%)	Diff (%)	KE	eKE	AUC	eAUC
1	64.97	64.97	0	0.39	0.39	0.602	0.612
2	83.16	82.83	−0.33	0.65	0.63	0.900	0.875
3	84.48	82.17	−2.31	0.68	0.63	0.897	0.887
4	82.83	82.50	−0.33	0.65	0.64	0.902	0.906
5	82.50	82.50	0	0.64	0.64	0.866	0.895
6	83.49	84.48	0.99	0.66	0.68	0.887	0.895
7	82.83	82.83	0	0.65	0.65	0.899	0.908
8	83.49	83.82	0.33	0.66	0.67	0.830	0.879
9	78.54	78.87	0.33	0.56	0.57	0.881	0.865
10	71.61	81.51	9.9	0.43	0.62	0.847	0.900
11	79.63	79.63	0	0.57	0.57	0.889	0.893
12	55.77	76.56	20.79	0.32	0.51	0.516	0.796
13	81.18	77.55	−3.63	0.62	0.54	0.867	0.816
14	71.94	82.17	10.23	0.43	0.64	0.715	0.890
15	80.19	82.83	2.64	0.59	0.65	0.798	0.886
16	80.85	83.49	2.64	0.61	0.66	0.807	0.886
17	71.61	83.82	12.21	0.43	0.67	0.716	0.892
18	81.18	82.17	0.99	0.62	0.64	0.845	0.898
19	77.22	82.83	5.61	0.53	0.65	0.768	0.893
20	54.45	54.45	0	0.30	0.30	0.487	0.487
21	82.83	82.83	0	0.64	0.64	0.875	0.879
22	78.54	82.50	3.96	0.56	0.64	0.810	0.897
23	71.61	80.85	9.24	0.43	0.61	0.680	0.891
24	81.51	84.15	2.64	0.62	0.67	0.843	0.900
25	77.22	82.17	4.95	0.53	0.64	0.795	0.895
26	81.51	82.50	0.99	0.62	0.63	0.879	0.882
27	82.83	82.83	0	0.65	0.65	0.899	0.907
28	78.54	80.52	1.98	0.57	0.60	0.791	0.875
29	79.20	82.17	2.97	0.57	0.64	0.837	0.895
30	79.86	82.17	2.31	0.59	0.63	0.835	0.894
AVG	77.52	80.49	2.97	0.56	0.61	0.805	0.862

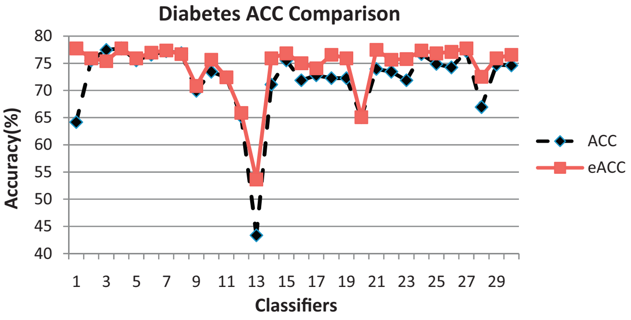

Fig. 1 – Accuracy comparison of classifiers and corresponding RF ensembles for diabetes dataset.

ble strategy has no impact on the performances of classifiers with IDs of {4,11,20}. On the other hand, classifiers of {3,8} have performance loss in terms of three metrics. Table 5 shows that the average accuracy of 30 base classifiers is 72.16% without RF scheme. However, if RF classifier ensemble strategy is used, the average accuracy is increased by 2.32%. This increase is supported with increase in KE and AUC values. In this manner, average KE increases from 0.40 to 0.43 and average AUC increases from 0.723 to 0.774. The variation of ACC with corresponding RF ensemble measures is shown in Fig. 1.

It is visually obvious from Fig. 1 that RF classifier ensembles increase percentage accuracies of base classifiers significantly.

For Cleveland heart database, the inspection of Table 6 shows that 19 of 30 classifiers have higher performance in ACC, KE and AUC values. Furthermore, RF ensemble classifiers with IDs {1,5,7,11,20,21,27} have no change while classifiers of {2,3,4,13} have loss in performance metrics. It is obviously seen that, the average accuracy of base classifiers is increased from 77.52% to 80.49% with the application of RF classifier ensemble algorithm. The overall increase in average accuracy is 2.97% in total. The increase in average ACC is supported with the raise in KE and AUC values. More significantly, KE value is increased from 0.56 to 0.61 and AUC is increased from 0.805 to 0.862. The variation of classifiers accuracy for Cleveland heart dataset is shown in Fig. 2.

Fig. 2 demonstrates success of RF classifier ensembles in terms of accuracy increase of base classifiers.

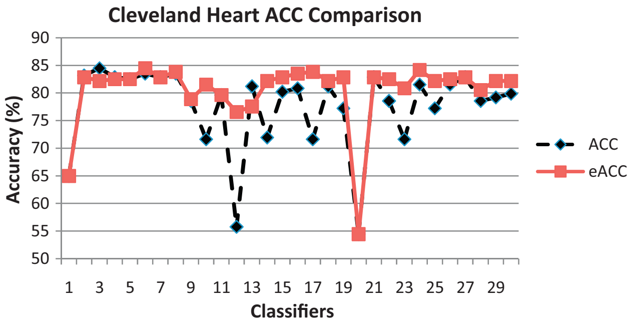

Fig. 2 – Accuracy comparison of classifiers and corresponding RF ensembles for Cleveland heart dataset.

Table 7 – Parkinson's disease classification results.

Alg. ID	ACC (%)	*e*ACC (%)	Diff (%)	KE	*e*KE	AUC	*e*AUC
1	75.4	86.2	10.8	0.41	0.56	0.648	0.845
2	82.6	85.1	2.5	0.53	0.58	0.827	0.847
3	77.4	75.9	−1.5	0.49	0.41	0.788	0.768
4	84.1	83.1	−1.0	0.56	0.53	0.837	0.828
5	89.7	90.3	0.6	0.73	0.74	0.898	0.903
6	87.7	87.7	0	0.64	0.64	0.873	0.872
7	85.1	86.2	1.1	0.56	0.59	0.843	0.854
8	87.2	87.7	0.5	0.58	0.61	0.856	0.863
9	91.8	93.8	2.0	0.78	0.85	0.917	0.94
10	84.1	88.2	4.1	0.55	0.63	0.838	0.871
11	82.6	85.6	3.0	0.53	0.61	0.827	0.855
12	84.1	86.6	2.5	0.45	0.52	0.824	0.857
13	73.3	78.5	5.2	0.42	0.51	0.751	0.797
14	80.5	81.0	0.5	0.36	0.32	0.777	0.766
15	88.2	89.7	1.5	0.65	0.71	0.877	0.894
16	88.2	88.8	0.6	0.65	0.66	0.877	0.879
17	86.2	86.7	0.5	0.58	0.57	0.852	0.849
18	81.1	92.3	11.2	0.46	0.79	0.805	0.922
19	87.2	89.7	2.5	0.65	0.66	0.872	0.881
20	75.4	75.4	0	0.37	0.37	0.648	0.648
21	88.7	92.3	3.6	0.70	0.78	0.888	0.922
22	89.2	90.8	1.6	0.69	0.73	0.888	0.904
23	83.1	85.6	2.5	0.52	0.55	0.826	0.844
24	83.6	87.7	4.1	0.52	0.66	0.829	0.876
25	85.6	89.7	4.1	0.62	0.72	0.857	0.896
26	88.7	92.3	3.6	0.71	0.79	0.888	0.922
27	86.2	88.2	2.0	0.61	0.68	0.856	0.881
28	84.6	91.3	6.7	0.57	0.71	0.844	0.893
29	82.6	85.7	3.1	0.54	0.61	0.827	0.855
30	88.7	91.8	3.1	0.68	0.78	0.856	0.892
AVG	84.4	87.1	2.7	0.57	0.63	0.833	0.860

For the last benchmark dataset, i.e. Parkinson's, the results of experiments are given in Table 7. While inspecting Table 7 for ACC values, it is clearly seen that 26 classifiers from 30 has an increased performance. More explicitly, average ACC value is seen to increase from 84.4% to 87.1% with RF ensemble application. Quantitatively, this is an average increase of 2.7% in overall accuracy. This performance increase is supported additionally with KE and AUC values. In more clear terms, average KE value is increased from 0.57 to 0.63 and average AUC value is increased from 0.833 to 0.860. On the other hand, it is further observed from Table 7 that the classifiers of {6,20} have no change in performance while classifiers of {3,4} have decrease in performance metrics. The overall accuracy performance of classifiers in Parkinson's dataset is shown in Fig. 3.

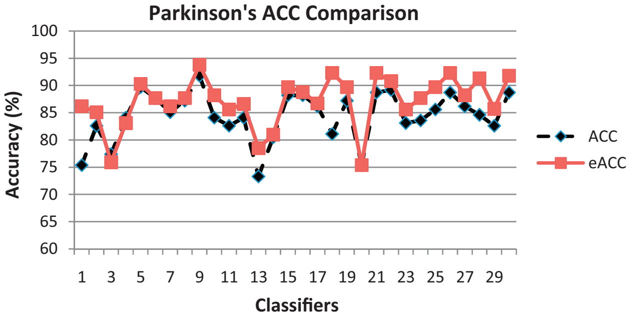

Fig. 3 – Accuracy comparison of classifiers and corresponding RF ensembles for Parkinson's dataset.

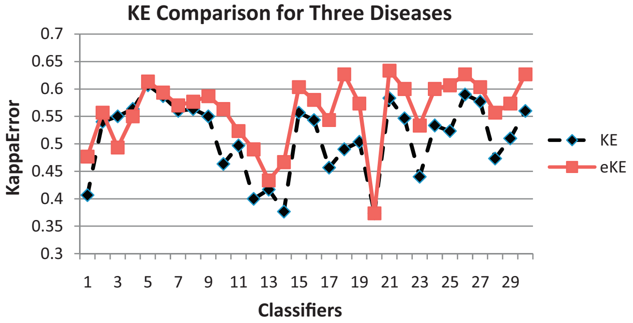

Fig. 4 – KE values of classifiers and corresponding RF ensembles for three disease datasets.

In Fig. 3, it can be seen that RF classifier ensembles increases the classification performance of classifiers in Parkinson's diagnosis.

Though the increase in classifier accuracies gives a concise idea about performance of the RF classifier ensemble strategy, we also need to observe changes in KE and AUC metrics to support accuracy evaluation. In order to evaluate changes in KE value, we will make use of a compact figure. Fig. 4 demonstrates a comparative change in KE and *e*KE by taking averages of three values of each disease.

In Fig. 4, it is easy to see increases in KE values for most of the classifiers. This overall increase also support the success of the RF classifier ensemble strategy in KE metric.

Table 8 – Algorithms with increased classification performance for all datasets.

No.	Name	No.	Name
1	Locally Weighed Learner	7	J48
2	JRIP	8	Alternating Decision Tree
3	PART Decision Learner	9	Random Tree
4	Ripple-Down Rule learner	10	Fast Decision Tree Learner
5	Best-first Decision Tree	11	Simple Chart
6	Decision Stump	12	

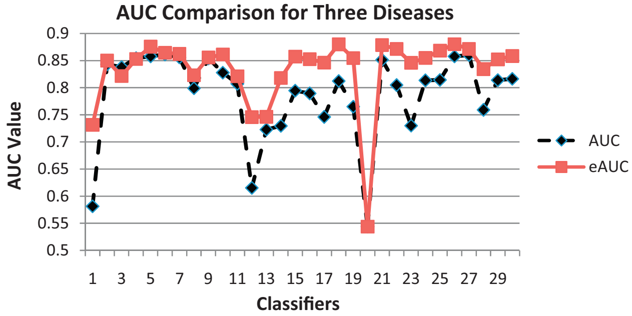

Fig. 5 – AUC values of classifiers and corresponding RF ensembles for three disease datasets.

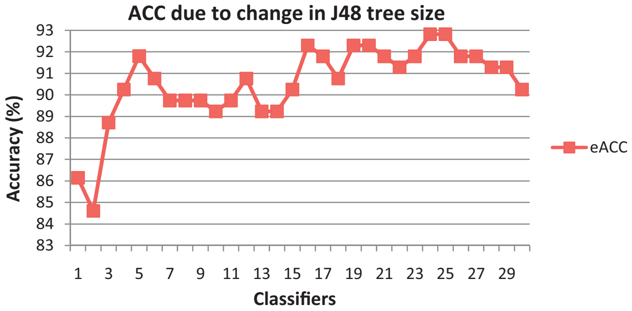

Fig. 6 – ACC change dependent on the number of classifiers in the ensemble.

In machine learning, ROC curve is a widely used evaluator to compare classifier algorithms. In our study, 30 base classifiers with their 30 corresponding ensembles require 60 ROC drawings for comparison of each disease. In order to simplify this comparison, we will develop a figure using averages of AUC values of ROC curves. Fig. 5 demonstrates averages of AUC values corresponding to 30 classifiers with their RF ensembles for three diseases.

Fig. 5 shows that the performances of most base classifiers are increased with RF ensemble scheme significantly.

6. Conclusion and remarks

Machine learning applications, particularly CADx systems, needs classifiers with enhanced accuracies. Such applications, in general, require a two-step approach: (i) a relevant feature selection algorithm to find the most powerful features and (ii) a high accuracy classifier to obtain the highest classification performance.

In our study, we did not evaluate the effect of feature selection algorithm on classifier performances. Instead, we used a simple CFS algorithm to decrease the feature size of the algorithms. The high number of the algorithms requires a more specific evaluation study to test the effect of the feature selection on classification accuracy. However, as an example the accuracies of diabetes, Cleveland heart and Parkinson's diseases without feature selection (with all features used in RF-J48 ensemble algorithm) are 76.17%, 82.17% and 89.23%, respectively. On the other hand, after CFS algorithm applied to each dataset, the classification accuracies are obtained as 76.82%, 82.17% and 89.7% in respective order. The comparison of results shows the effect of feature selection on the classification accuracies for this classification scheme.

In general, obtaining a high accurate classifier requires an assessment study to find the highest performance algorithm for particular dataset. In this concept, classifier ensemble approaches are important strategies to increase classifier performances further. Experiments demonstrate that RF, as a newly proposed ensemble algorithm, proves itself to be efficient to increase classifier accuracies significantly. For three benchmarking datasets, i.e. diabetes, Cleveland heart and Parkinson's, the average performance of the 30 base classifiers is increased by 2.32%, 2.97% and 2.7% with the use of RF ensemble algorithm. Furthermore, rise in averages of AUC and KE values support the performance increase in classifiers.

We analyzed the response of 30 base classifiers to RF classifier ensemble application in classification of three diseases. While Tables 5–7 are examined for the classifiers that improved classification accuracies more than 1% for all datasets, we obtain Table 8.

From Table 8, it is seen that 11 classifiers are grouped in three categories, i.e. a lazy learner {1}, three rule based learners {2,3,4} and seven {5,6,7,8,9,10,11} decision tree learners. From this distribution, it might be recommended to give priority to use either rule based learners or decision tree learners to enhance CADx system performance.

An issue with ensemble classification approaches is that number of base classifiers used in community might affect the resultant ACC. It would be impossible to analyze this change for all classifiers of this study. However, we visually demonstrate the effect of changing number of classifiers on resultant ACC in Fig. 6. We selected J48 for RF ensemble algorithm and changed number of J48 in the community from 1 to 30. In order to test the performance of the RF ensemble, we applied the algorithm to Parkinson's disease with 10-fold cross validation.

Fig. 6 shows interestingly that increasing the number of J48 classifiers in the ensemble might increase the accuracy. Therefore, CADx applications based on ensemble strategies might require determining optimum number of classifiers in the community.

Another issue that might be evaluated is the class structure of the datasets. More explicitly, the three datasets used in this study have binary class labels. The study might be extended to multi-class datasets to assess classification performance of RF classifier ensemble approach.

There is no universal predictor that guarantees high classification performances for all data analysis studies. In this aspect, obtaining high accurate classifiers for a particular data requires some evaluation of various algorithms. Having obtained an optimum algorithm for a classification task, it is further possible to increase classifier performance with the utilization of ensemble learning approaches. This study shows that RF classifier algorithm might be used to enhance classification performance of base algorithms. In CADx literature, it is valuable to improve performance of a system even with minor percentages. Therefore, for a given classification task, it is rational to use ensemble algorithms (particularly RF approach) to obtain enhanced classification accuracies.

REFERENCES

[1] L. Ming, Z. Zhi-Hua, Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples, IEEE Trans. Syst. Man Cybern. A: Syst. Hum. 37 (2007) 1088–1098.

[2] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

[3] C. Demir, E. Alpaydin, Cost-conscious classifier ensembles, Pattern Recogn. Lett. 26 (2005) 2206–2214.

[4] M.C. Lee, L. Boroczky, K. Sungur-Stasik, A.D. Cann, A.C. Borczuk, S.M. Kawut, C.A. Powell, A two-step approach for feature selection and classifier ensemble construction in computer-aided diagnosis, in: 21st IEEE International Symposium on Computer-Based Medical Systems, 2008 (CBMS ’08), 2008, pp. 548–553.

[5] K. Michalak, H. Kwasnicka, Correlation-based feature selection strategy in neural classification, Sixth International Conference on Intelligent Systems Design and Applications, 2006 (ISDA ’06), 2006, pp. 741–746.

[6] A. Mendiburu, J. Miguel-Alonso, J.A. Lozano, M. Ostra, C. Ubide, Parallel and multi-objective EDAs to create multivariate calibration models for quantitative chemical applications, in: Proceedings of the 2005 International Conference on Parallel Processing Workshops, IEEE Computer Society, 2005, pp. 596–603.

[7] R.E Abdel-Aal, GMDH-based feature ranking and selection for improved classification of medical data, J. Biomed. Inform. 38 (2005) 456–468.

[8] I. Skrypnyk, Comparison of feature selection strategies for hearing impairments diagnostics, in: Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems (CBMS’02), IEEE Computer Society, 2002, p. 231.

[9] A.G. Karegowda, M.A. Jayaram, Cascading GA; CFS for feature subset selection in medical data mining, in: IEEE International Advance Computing Conference, 2009 (IACC 2009), 2009, pp. 1428–1431.

[10] Y. Cheng-San, C. Li-Yeh, K. Chao-Hsuan, Y. Cheng-Hong, A hybrid approach for selecting gene subsets using gene expression data, in: IEEE Conference on Soft Computing in Industrial Applications, 2008 (SMCia ’08), 2008, pp. 159–164.

[11] R. Duangsoithong, T. Windeatt, Relevance and redundancy analysis for ensemble classifiers, in: Proceedings of the 6th International Conference on Machine Learning and Data Mining in Pattern Recognition, Springer-Verlag, Leipzig, Germany, 2009, pp. 206–220.

[12] R. Polikar, Ensemble based systems in decision making, IEEE Circuits Syst. Mag. 6 (2006) 21–45.

[13] K. Nigar Sen, Y. Nese, Y. Gunes, Ensemble classifiers for medical diagnosis of knee osteoarthritis using gait data, in: 5th International Conference on Machine Learning and Applications, 2006 (ICMLA ’06), 2006, pp. 225–230.

[14] H. Zhonghui, C. Yunze, L. Ye, X. Xioaming, Support vector machine based ensemble classifier, in: Proceedings of the American Control Conference, 2005, vol. 742, 2005, pp. 745–749.

[15] M.A. Mazurowski, J.M. Zurada, An adaptive incremental approach to constructing ensemble classifiers: application in an information-theoretic computer-aided decision system for detection of masses in mammograms, Med. Phys. 36 (2009) 2976–2984.

[16] K.-J. Kim, S.-B. Cho, Ensemble classifiers based on correlation analysis for DNA microarray classification, Neurocomputing 70 (2006) 187–199.

[17] J.-H. Eom, S.-C. Kim, B.-T. Zhang, AptaCDSS-E: a classifier ensemble-based clinical decision support system for cardiovascular disease level prediction, Expert Syst. Appl. 34 (2008) 2465–2479.

[18] K.-H. Liu, D.-S. Huang, Cancer classification using rotation forest, Comput. Biol. Med. 38 (2008) 601–610.

[19] A.H.R. Ko, R. Sabourin, A. de Souza Britt, Combining diversity and classification accuracy for ensemble selection in random subspaces, in: International Joint Conference on Neural Networks, 2006 (IJCNN ’06), 2006, pp. 2144–2151.

[20] L. Yu, H. Liu, Feature selection for high-dimensional data: a fast correlation-based filter solution, in: Proceedings of ICML, 2003, pp. 856–863.

[21] M. Hall, Correlation-based Feature Selection for Machine Learning. PhD thesis, Department of Computer Science, University of Waikato, New Zealand, 1999, pp. 51–74.

[22] M.A. Hall, G. Holmes, Benchmarking attribute selection techniques for discrete class data mining, IEEE Trans. Knowl. Data Eng. 15 (2003) 1437–1447.

[23] R. Polikar, A. Topalis, D. Parikh, D. Green, J. Frymiare, J. Kounios, C.M. Clark, An ensemble based data fusion approach for early diagnosis of Alzheimer’s disease, Inf. Fusion 9 (2008) 83–95.

[24] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[25] L.I. Kuncheva, J.J. Rodriguez, An experimental study on rotation forest ensembles, in: Proceedings of the 7th International Conference on Multiple Classifier Systems, Springer-Verlag, Prague, Czech Republic, 2007, pp. 459–468.

[26] J.J. Rodriguez, L.I. Kuncheva, C.J. Alonso, Rotation forest: a new classifier ensemble method, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1619–1630.

[27] I.H. Witten, H. Ian, Data mining: practical machine learning tools and techniques, in: Morgan Kaufmann Series in Data Management Systems, 2005, pp. 153–168.

[28] A. Ben-David, Comparison of classification accuracy using Cohen’s Weighted Kappa, Expert Syst. Appl. 34 (2008) 825–832.

[29] C. Loy, W. Lai, C. Lim, Dimensionality reduction of protein mass spectrometry data using random projection, in: I. King, J. Wang, L. Chan, D. Wang (Eds.), Neural Information Processing, Springer, Berlin/Heidelberg, 2006, pp. 776–785.

[30] M.A. Little, P.E. McSharry, E.J. Hunter, J. Spielman, L.O. Ramig, Suitability of dysphonia measurements for telemonitoring of Parkinson’s disease, IEEE Trans. Biomed. Eng. 56 (2009) 1015–1022.