Rotation Forest: a new classifier ensemble method

Paper Review

Miguel A. Veganzones

Grupo de Inteligencia Computacional
Universidad del País Vasco

2012-01-27

Outline

Outline

Paper

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 28, NO. 10, OCTOBER 2006 1619

Rotation Forest: A New Classifier Ensemble Method

Juan J. Rodríguez, *Member*, *IEEE Computer Society*,
Ludmila I. Kuncheva, *Member*, *IEEE*, and Carlos J. Alonso

Motivation

- Two approaches for constructing classifier ensembles:
  - *Bagging*: takes bootstrap samples of objects and trains a classifier on each sample. **Random Forest**.
  - *Boosting*: combine weak classifiers so a new classifier is trained on data which have been 'hard' for the previous ensembled methods. **AdaBoost**.

Motivation (II)

- On average AdaBoost is the best method.
  - For large ensemble sizes differences dissapear.
  - Quest: consistently good ensemble strategy for small ensemble sizes?
- The sucess of AdaBoost has been explained by its large diversity boosting the ensemble performance.
  - Accuracy-diversity dilemma: it seems that classifiers cannot be both very accurate and have very diverse outputs.

Proposal

- New classifier ensemble method:
  - Based on feature extraction (PCA) and decision trees (J48).
  - Achieving both, accuracy and diversity.
- Compared to Bagging, AdaBoost and Random Forest.
- Using 33 benchmark datasets from UCI repository.

Outline

Outline

Idea

- To create the training data:
  1. The feature set is randomly split into K subsets.
  2. PCA is applied to each subset.
  3. All principal components are retained to preserve the variability information in the data.
- Thus, K axis rotations take place to form the new features for a base classifier.
  - Encourage simultaneously individual accuracy and diversity within the ensemble.
- Decision trees were choosen because they are sensitive to rotation of the feature axes.

Training phase

Given

- X: the objects in the training data set (an $N \times n$ matrix)
- Y: the labels of the training set (an $N \times 1$ matrix)
- L: the number of classifiers in the ensemble
- K: the number of subsets
- $\{\omega_1, \dots, \omega_c\}$: the set of class labels

For $i = 1 \dots L$

- Prepare the rotation matrix R_i^a:
  - Split $\mathbf{F}$ (the feature set) into K subsets: $\mathbf{F}_{i,j}$ (for $j = 1 \dots K$)
  - For $j = 1 \dots K$
    - Let $X_{i,j}$ be the data set X for the features in $\mathbf{F}_{i,j}$
    - Eliminate from $X_{i,j}$ a random subset of classes
    - Select a bootstrap sample from $X_{i,j}$ of size 75% of the number of objects in $X_{i,j}$. Denote the new set by $X'_{i,j}$
    - Apply PCA on $X'_{i,j}$ to obtain the coefficients in a matrix $C_{i,j}$
  - Arrange the $C_{i,j}$, for $j = 1 \dots K$ in a rotation matrix R_i as in equation (1)
  - Construct R_i^a by rearranging the the columns of R_i so as to match the order of features in $\mathbf{F}$.
- Build classifier D_i using (XR_i^a, Y) as the training set

Rotation matrix

$$R_i = \begin{bmatrix} \mathbf{a}_{i,1}^{(1)}, \mathbf{a}_{i,1}^{(2)}, \ldots, \mathbf{a}_{i,1}^{(M_1)}, & [0] & \cdots & [0] \\ [0] & \mathbf{a}_{i,2}^{(1)}, \mathbf{a}_{i,2}^{(2)}, \ldots, \mathbf{a}_{i,2}^{(M_2)}, & \cdots & [0] \\ \vdots & \vdots & \ddots & \vdots \\ [0] & [0] & \cdots & \mathbf{a}_{i,K}^{(1)}, \mathbf{a}_{i,K}^{(2)}, \ldots, \mathbf{a}_{i,K}^{(M_K)} \end{bmatrix}$$

- $\mathbf{a}_{i,j} \in \mathbb{R}^M$, where $M = n/K$.
- Dimensionality: $n \times \sum_j M_j$.
  - $M_j \leq M$ (some eigenvalues could be zero).
- Columns must be rearranged so that they correspond to the original features.

Classification phase

- For a given $\mathbf{x}$, let $d_{i,j}(\mathbf{x}R_i^a)$ be the probability assigned by the classifier D_i to the hypothesis that $\mathbf{x}$ comes from class ω_j. Calculate the confidence for each class, ω_j, by the average combination method:

$$\mu_j(\mathbf{x}) = \frac{1}{L}\sum_{i=1}^{L} d_{i,j}(\mathbf{x}R_i^a), \quad j = 1, \ldots, c.$$

- Assign $\mathbf{x}$ to the class with the largest confidence.

Outline

PCA

- PCA is not particularly suitable for feature extraction in classification because it does not include discriminatory information in calculating the optimal rotation of the axes.
  - Problems are related to dimensionality reduction.
- In the proposed algorithm authors keep all the components so the discriminatory information will be preserved.
  - Keeping all the components does not mean that the classification will be easier in the new space of extracted features.
- **Even if the rotation does not contribute much to finding good discriminatory directions, it is valuable here as a divesifying heuristic.**

Diversity

- The intended diversity will come from the difference in the possible feature subsets:
  - There are in total $T = \frac{n!}{K!(M!)^K}$ different partitions of the feature set into K subsets of size M, each given raise to a classifier.
  - If the ensemble consists of L classifiers, assuming each partition is equally probable, the probability that all classifiers will be different is $P = \frac{T!}{(T-L)!T^L}$.

Example

The chance to have all different classifiers in an ensemble of $L = 50$ classifiers for $K = 3$ and $n = 9$ is less than 0.01.

- There is a need for an extra randomization of the ensemble.

Extra randomization

- Applying PCA to:
  - A bootstrap sample from X.
  - A random subset of X.
  - A random selection of classes.

Outline

Outline

Datasets

- 33 datasets from UCI repository.

Data set	Classes	Objects	Discrete features	Continuous features
anneal	6	898	32	6
audiology	24	226	69	0
autos	7	205	10	16
balance-scale	3	625	0	4
breast-cancer	2	286	10	0
cleveland-14-heart	5	307	7	6
credit-rating	2	690	9	6
german-credit	2	1000	13	7
glass	7	214	0	9
heart-statlog	2	270	0	13
hepatitis	2	155	13	6
horse-colic	2	368	16	7
hungarian-14-heart	5	294	7	6
hypothyroid	4	3772	22	7
ionosphere	2	351	0	34
iris	3	150	0	4
labor	2	57	8	8
letter	26	20000	0	16
lymphography	4	148	15	3
pendigits	10	10992	0	16
pima-diabetes	2	768	0	8

Data set	Classes	Objects	Discrete features	Continuous features
primary-tumor	22	239	17	0
segment	7	2310	0	19
sonar	2	208	0	60
soybean	19	683	35	0
splice	3	3190	60	0
vehicle	4	846	0	18
vote	2	435	16	0
vowel-c	11	990	2	10
vowel-n	11	990	0	10
waveform	3	5000	0	40
wisconsin-breast	2	699	0	9
zoo	7	101	16	2

Algorithms

- Compare Rotation Forest with Bagging, AdaBoost and Random Forest.
  - In all ensemble methods decision trees were used as the base classifier.
- The decision tree construction method was J48 (a reimplementation of C4.5).
  - Except for the Random Forest method.
- All implementations are from Weka.

Algorithms settings

- As PCA is defined for numerical features, discrete features were converted to numeric ones for Rotation Forest. **Important!**
  - Each categorical feature was replaced by s binary features, where s is the number of possible categories of the feature.
- The parameters of Bagging, AdaBoost and Random Forest were kept at their default values.
- For Random Forest the number of features to select from at each node is set at $\log_2(n)+1$.
- For Rotation Forest the number of features in each subset was fixed to $M=3$.
  - If n did not divide by 3, the remainder subset was completed by features randomly selected from the rest of the feature set.

Pruning

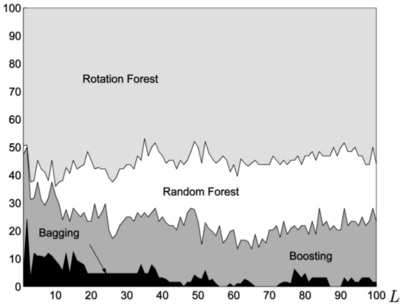
- The decision tree classifier, J48, uses an error-based pruning algorithm.
  - Confidence value to be used when pruning the tree is set the default of 25 percent.
- Thus, two versions of each algoritm, with pruning or without pruning.
  - This standar implementation was not suitable for Random Forest, so there is only unpruned Random Forest.

Ensemble size

- The ensemble size L can be regarded as an hyperparameter of the ensemble method.
  - It can be tuned through cross-validation.
- L can also be though of as an indicator of the operating complexity of the ensemble.
  - Then we can choose the most accurate ensemble of a fixed complexity.
- As we are interested in ensembles of a small (fixed) size, we decided to train all the ensemble methods with the same $L = 10$.

Ensemble size (II)

- Percentage graph for ensembles of unpruned decision trees using one 10-fold cross validation.
- The x-axis is the ensemble size L. The y-axis shows the percent of the datasets in which the method has been the one with the lowest error.



Validation measures

- For each dataset and ensemble method, 15 10-fold cross validation were performed.
- The average accuracies and corrected standard deviations are shown.
- For reference, we display the accuracy of a single J48 tree as well.
- The results for which a significant difference (5 percent) with Rotational Forest was found are marked with a bullet (better) or an open circle (worse) next to them.

Corrected standar deviation

- Instead of taking $\sigma_{\tilde{\mu}} = \frac{\sigma_{\mu}}{\sqrt{T}}$ where T is the number of experiments, the authors propose:

$$\sigma_{\tilde{\mu}} = \sqrt{\frac{1}{T} + \frac{N_{\text{testing}}}{N_{\text{training}}}}$$

  where N_{training} and N_{testing} are the sizes of the training and the testing sets respectively.
- The new estimate is more conservative.
- Note that the comparison was done using all the $T = 150$ testing accuracies per method and data set (15×10-fold CV).

Outline

With pruning

Classification Accuracy and Standard Deviation of J48 and Ensemble Methods with Pruning

Data Set	Rotations J48	J48	Bagging J48	Boosting J48
anneal	98.93±0.95	98.61±1.06	98.89±0.92	99.58±0.71
audiology	79.80±6.92	77.24±7.04	81.03±7.36	84.90±7.07 ○
autos	82.50±8.66	82.34±9.22	82.69±8.60	85.31±6.99
balance-scale	90.33±2.52	77.82±3.69 •	81.85±3.74 •	78.46±4.07 •
breast-cancer	72.66±6.71	74.19±6.05	72.65±6.12	66.88±7.37 •
cleveland-14-heart	82.85±6.26	76.71±6.84 •	79.21±6.74	79.38±6.99
credit-rating	86.13±3.88	85.63±4.12	85.78±4.02	83.86±4.35
german-credit	74.10±3.93	71.09±3.53 •	73.75±3.62	71.01±3.93 •
glass	74.27±8.11	67.55±9.33 •	73.97±9.41	75.20±8.26
heart-statlog	82.25±6.43	78.22±7.20	80.74±6.66	78.27±7.20
hepatitis	82.80±8.91	79.58±9.28	81.24±8.22	82.46±8.00
horse-colic	84.73±5.44	85.16±5.70	85.41±5.70	81.63±6.11
hungarian-14-heart	80.28±6.33	80.08±7.65	79.62±6.70	78.75±6.65
hypothyroid	99.56±0.35	99.53±0.35	99.58±0.32	99.64±0.30
ionosphere	93.88±3.68	89.91±4.57 •	92.25±3.80	93.18±4.02
iris	95.73±5.20	94.89±5.03	94.67±5.12	94.27±5.18
labor	91.56±11.91	79.56±15.78•	83.13±15.20	87.31±13.36
letter	95.48±0.47	88.04±0.73 •	92.72±0.63 •	95.53±0.47
lymphography	83.99±8.33	76.37±11.09•	77.97±10.22•	81.73±8.61
pendigits	99.20±0.26	96.46±0.56 •	97.93±0.47 •	99.02±0.30
pima-diabetes	76.48±4.44	74.38±4.91	75.65±4.45	71.96±4.53 •
primary-tumor	45.06±6.40	41.71±6.83	43.74±6.76	41.87±6.53
segment	98.05±0.95	96.79±1.28 •	97.49±1.07	98.14±0.89
sonar	83.56±7.84	73.98±8.67 •	78.31±9.11	79.79±8.63
soybean	94.77±2.36	91.90±3.11 •	92.73±2.87 •	92.74±2.82 •
splice	95.47±1.15	94.17±1.22 •	94.43±1.26 •	94.60±1.15 •
vehicle	78.05±3.64	72.33±4.42 •	74.45±4.18 •	75.78±4.19
vote	96.26±2.79	96.49±2.65	96.37±2.54	95.34±3.11
vowel-c	96.89±1.74	79.62±4.17 •	90.20±3.16 •	92.77±2.77 •
vowel-n	95.68±1.95	79.16±4.58 •	89.45±3.22 •	92.13±2.84 •
waveform	83.93±1.69	75.27±2.00 •	81.75±1.70 •	81.34±1.88 •
wisconsin-breast-cancer	97.04±1.94	94.87±2.69 •	95.99±2.44	96.06±2.27
zoo	92.15±8.22	92.56±7.04	93.30±7.07	96.38±5.75
(Win/Tie/Loss)		(0/16/18)	(0/24/10)	(1/24/9)

○ Rotation Forest is significantly worse, • Rotation Forest is significantly better, level of significance 0.05

Without pruning

Classification Accuracy and Standard Deviation of J48 and Ensemble Methods *without* Pruning

Data Set	Rotations J48	J48	Bagging J48	Boosting J48	Random Forest
anneal	99.01±0.93	98.62±1.01	98.98±0.93	99.54±0.68	99.38±0.78
audiology	79.83±6.93	76.33±7.45 •	81.12±7.35	83.30±6.99	76.58±7.94
autos	82.56±8.66	82.86±9.25	84.12±8.42	84.61±7.93	81.95±7.85
balance-scale	90.26±2.62	79.43±4.01 •	81.39±3.70 •	76.82±4.14 •	80.28±3.80 •
breast-cancer	72.07±6.54	68.00±7.43	69.48±7.17	66.12±7.81 •	69.00±7.31
cleveland-14-heart	82.61±6.12	76.49±6.91 •	79.70±6.01	79.20±7.25	80.34±6.47
credit-rating	86.00±3.90	82.50±4.24 •	85.17±4.34	84.02±3.98	85.15±4.23
glass	74.33±8.06	67.77±9.70 •	73.85±9.34	76.23±9.09	75.65±8.42
german-credit	73.87±3.89	67.89±3.95 •	72.08±3.63	71.95±4.32	73.57±3.38
heart-statlog	82.37±6.45	76.69±7.51 •	80.44±6.84	79.38±7.40	80.86±6.53
hepatitis	82.92±8.88	78.95±9.27	80.68±8.89	82.45±8.17	83.04±8.07
horse-colic	84.80±5.35	82.16±5.89	84.80±5.96	81.05±6.20	84.96±5.43
hungarian-14-heart	79.57±6.45	78.85±7.30	78.74±6.65	79.08±7.00	79.28±6.31
hypothyroid	99.57±0.33	99.51±0.37	99.59±0.30	99.65±0.30	99.18±0.46 •
ionosphere	93.88±3.76	89.97±4.55 •	92.29±3.79	93.01±3.97	92.84±3.89
iris	95.73±5.20	94.93±4.99	94.58±5.15	94.36±5.22	94.13±5.18
labor	91.69±11.89	79.84±14.57•	84.31±14.44	87.20±13.81	87.00±13.45
letter	95.54±0.47	88.02±0.75 •	92.85±0.65 •	95.44±0.50	94.52±0.49 •
lymphography	84.27±8.35	75.64±11.12•	78.97±10.32	82.40±9.73	81.28±8.58
pendigits	99.21±0.25	96.46±0.57 •	97.99±0.44 •	99.01±0.28 •	98.81±0.29 •
pima-diabetes	76.39±4.43	73.85±4.94	75.59±4.54	72.49±5.08 •	74.78±4.42
primary-tumor	44.37±6.56	42.42±7.57	42.79±6.92	41.64±6.94	41.56±6.50
segment	98.05±0.95	96.81±1.26 •	97.58±1.05	98.25±0.80	97.71±1.06
sonar	83.49±7.88	73.82±8.71 •	78.34±9.14	79.95±9.51	80.75±7.84
soybean	94.17±2.47	90.67±3.34 •	91.88±3.15 •	92.44±2.76	91.92±2.83 •
splice	95.49±1.13	92.20±1.37 •	94.25±1.20 •	94.11±1.23 •	90.07±1.79 •
vehicle	77.95±3.74	72.38±4.25 •	74.70±4.07 •	76.44±4.01	74.37±4.43 •
vote	96.08±2.88	95.71±2.93	96.43±2.47	95.22±3.19	95.74±2.75
vowel-c	96.87±1.76	81.26±4.18 •	91.72±2.89 •	94.15±2.42 •	95.59±2.23
vowel-n	95.77±1.94	79.22±4.59 •	89.52±3.27 •	91.93±2.72 •	92.37±2.73 •
waveform	83.94±1.72	75.14±1.99 •	81.78±1.74 •	81.45±1.71 •	81.89±1.74 •
wisconsin-breast-cancer	97.02±1.93	94.30±2.74 •	95.82±2.54 •	95.97±2.11	95.75±2.14 •
zoo	92.35±8.04	93.42±6.93	93.50±7.11	97.04±5.21 ○	95.83±6.02
(Win/Tie/Loss)		(0/13/21)	(0/24/10)	(1/25/8)	(0/24/10)

○ Rotation Forest is significantly worse, • Rotation Forest is significantly better, level of significance 0.05

Accuracy comparison

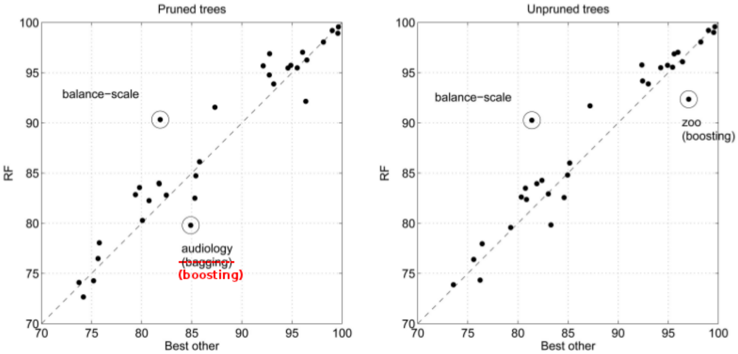


Fig. 4. Comparison of accuracy of Rotation Forest ensemble (RF) and the best accuracy from any of a single tree, Bagging, Boosting, and Random Forest ensembles.

Summary

	Pruned trees				Unpruned trees				
	J48	Bagging	AdaBoost	Rotation Forest	J48	Bagging	AdaBoost	Random Forest	Rotation Forest
Pruned trees									
J48	-	29 (9)	25 (12)	29 (18)	14 (2)	26 (9)	23 (12)	22 (8)	28 (18)
Bagging	4 (0)	-	21 (7)	27 (10)	3 (0)	18 (2)	17 (6)	16 (4)	25 (9)
AdaBoost	8 (1)	12 (3)	-	25 (9)	8 (0)	12 (2)	15 (0)	16 (1)	26 (7)
Rotation Forest	4 (0)	6 (0)	8 (1)	-	2 (0)	7 (0)	7 (1)	5 (0)	17 (0)
Unpruned trees									
J48	19 (5)	30 (14)	25 (14)	31 (19)	-	31 (12)	26 (13)	28 (9)	31 (21)
Bagging	7 (1)	15 (0)	21 (4)	26 (10)	2 (0)	-	20 (5)	20 (4)	28 (10)
AdaBoost	10 (1)	16 (3)	18 (0)	26 (8)	7 (1)	13 (1)	-	15 (1)	26 (8)
Random Forest	11 (2)	17 (2)	17 (5)	28 (10)	5 (2)	13 (2)	18 (4)	-	28 (10)
Rotation Forest	5 (0)	8 (0)	7 (1)	15 (0)	2 (0)	5 (0)	7 (1)	5 (0)	-

The entry $a_{i,j}$ shows the number of times method of the column (j) has a better result than the method of the row (i). The number in the parentheses shows in how many of these differences have been statistically significant.

Outline

Outline

Overview

- Visualization means for classifier ensembles.
- Based on pairwise diversity measures.
- Diversity is intuitively clear for two variables (two classifier outputs).
  - Measured as "deviation from independence" using a correlation coefficient or an appropiate statistic for nominal variables (class labels).
- Difficult to define for more than two variables.

Kappa

- The pairwise diversity measure used is the interrater agreement, kappa (κ).
- Kappa evaluates the level of agreement between two classifier outputs while correcting for chance.
- For c class labels, kappa is defined on the $c \times c$ coincidence matrix $\mathscr{M}$ of the two classifiers.
- The entry $m_{k,s}$ of $\mathscr{M}$ is the proportion of the dataset used for testing, which D_i labels as ω_k and D_j labels as ω_s.

Kappa (II)

- The agreement between D_i and D_j is given by:

$$\kappa_{i,j} = \frac{\sum_k m_{k,k} - ABC}{1 - ABC}$$

where $\sum_k m_{kk}$ is the observed agreement between the classifiers and ABC is “agreement by chance”:

$$ABC = \sum_k \left(\sum_s m_{k,s} \right) \left(\sum_s m_{s,k} \right)$$

Kappa (III)

- Low values of κ signify high disagreement and, hence, high diversity.
- If the classifiers produce identical class labels, $\kappa = 1$.
- If the classifiers are independent, $\kappa = 0$.
  - Independence is not necessarily the best scenario in multiple classifier systems.
- More desirable is "negative dependence", $\kappa < 0$.
  - Classifiers commit related erros.
  - When one classifier is wrong, the other has more than random chance of being correct.

Kappa-Error diagrams

- An ensemble of L classifiers generates $L(L-1)/2$ pairs of classifiers D_i, D_j.
  - Points in the diagram.
- Kappa-Error diagram:
  - x-axis: κ for the pair of classifiers.
  - y-axis: averaged individual error of D_i and D_j, $E_{i,j} = \frac{E_i + E_j}{2}$.
- The most desirable point will lie in the bottom left corner: low kappa and low error.

Outline

Kappa-Error diagrams

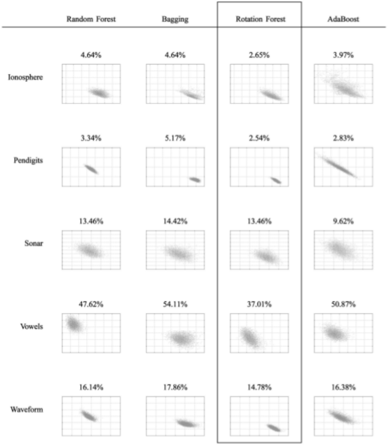


Fig. 5. κ-Error Diagrams. x-axis = κ, y-axis = $E_{i,j}$ (average error of the pair of classifiers). Axes scales are constant for each row. The *ensemble error* on the testing set is displayed above the plot.

Kappa-Error centroids

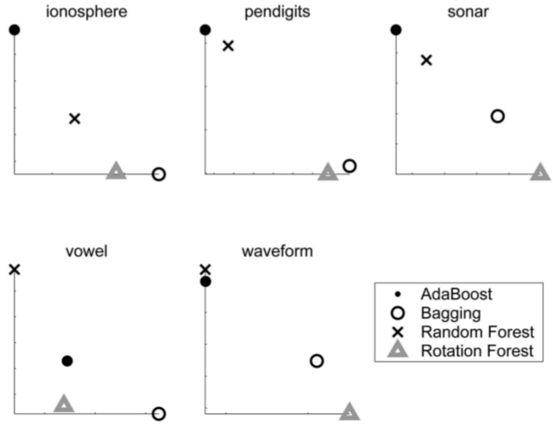


Fig. 6. Centroids of the kappa-error clouds for the five data.

Kappa-Error diagram

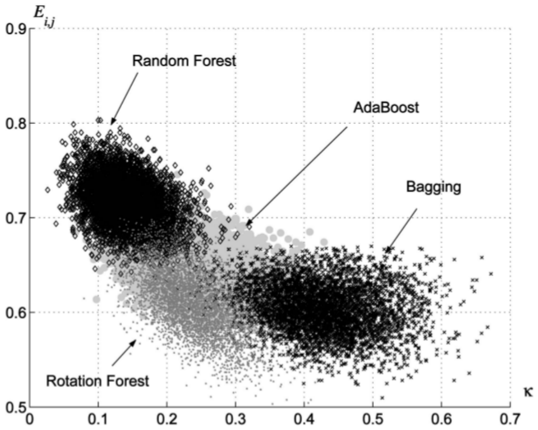


Fig. 7. Kappa-error diagrams for the vowel-n data set.

Kappa-Error diagram

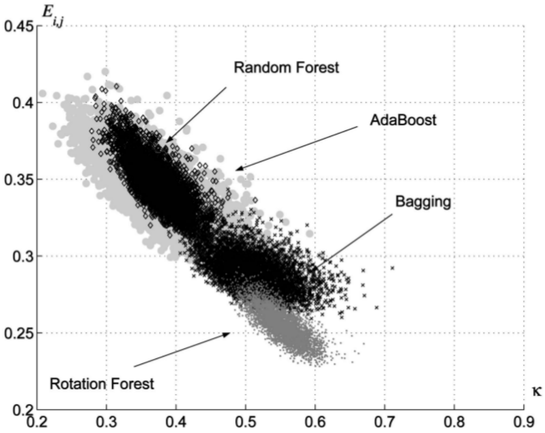


Fig. 8. Kappa-error diagrams for the waveform data set.

Outline

Conclusions

- In general, Rotation Forest is similar to Bagging.
    - Like Bagging, Rotation Forest is more accurate and less diverse than both AdaBost and Random Forest.
- Results show that the minimal improvement on the diversity-accuracy pattern materializes in significant better ensembles.

Caveats

- Rotation Forest has an extra parameter which controls the sizes of the feature subsets or eqivalently the number of feature subsets.
  - We did not tune the hyperparameters of any of the ensemble methods.
- All datasets are from UCI repository.
  - Do not include very large-scale datasets.
- Random Forest offers a way to order the features by their importance.
- We used the same ensemble size L for all methods.

Outlook

- Evaluation of the sensitivity of the algorithm to the choice of M and L.
- Application of Rotation Forest together with other ensemble approaches.
- Trying a different base classifier model.
- Examining the effect of randomly pruning classes and taking a bootstrap sample for each feature subset, prior to applying PCA.
  - Find out whether or not this will have an adverse effect on the performance of Rotation Forest.
- Use a different feature extraction algorithm.