Boosting and Bagging with Classification Noise

Thomas G. Dietterich
Department of Computer Science
Oregon State University
Corvallis, Oregon 97331

Currently on sabbatical leave at
Institut D'Investigació en Intel·ligéncia Artificial
Consejo Superior de Investigaciones Científicas
Bellaterra, Spain
tgd@cs.orst.edu

Classification Noise

Goal: Compare Adaboost and Bagging with classification noise

- **Base algorithm: C4.5 release 1**
- **10-fold cross-validation**
- **Randomly choose $x\%$ of the training (and test) examples, and force their class to be incorrect (randomly choosing among the other classes).**
- **9 UCI domains chosen based on previous study that showed interesting performance differences.**
- **100 iterations of Adaboost.M1 (boosting by weighting);**
  200 iterations of bagging
- **Pruning (at 10%) vs. no pruning chosen using** *test set* (contamination)

Domains

audiology
hypo
king-rook-vs-king-pawn
satimage
segment
sick
splice
vehicle
waveform

These were chosen because they were domains where a randomized variant of C4.5 gave different performance than bagging.

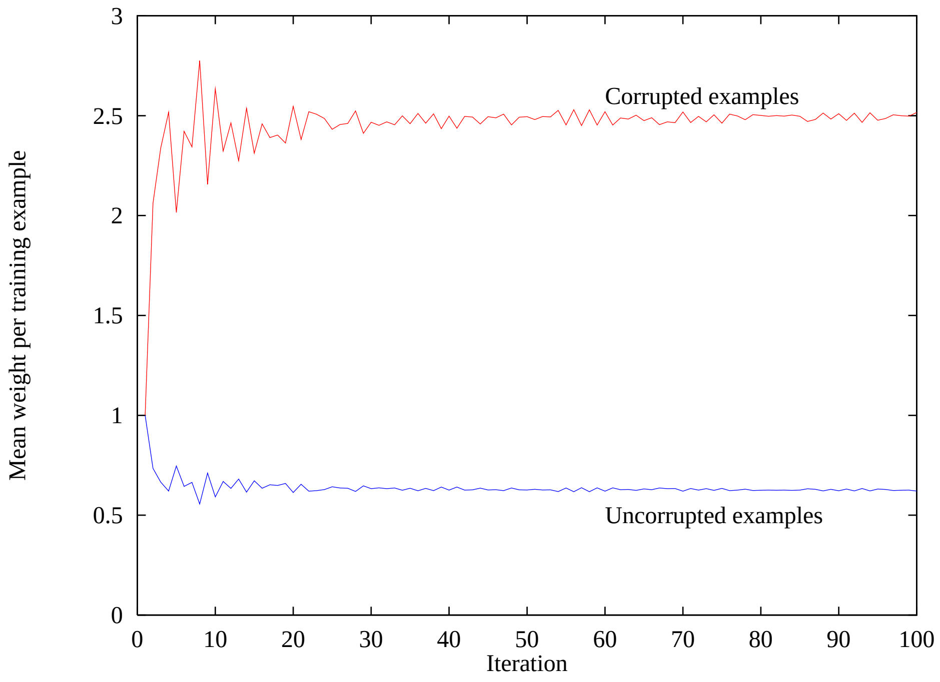
Results

Each cell of the following table gives Win-Lose-Tie (based on a 10-fold cross-validated t test, $p < 0.05$).

noise rate	adaboost vs. C4.5	adaboost vs. bagging	bagging vs. C4.5
0%	$6 - 0 - 3$	$5 - 0 - 4$	$4 - 0 - 5$
5%	$3 - 3 - 3$	$1 - 5 - 3$	$6 - 0 - 3$
10%	$2 - 3 - 4$	$1 - 6 - 2$	$5 - 0 - 4$
20%	$3 - 6 - 0$	$0 - 6 - 3$	$7 - 0 - 2$

- **Adaboost goes from being the best method to being the worst; worse than C4.5 in 6 domains!**
- **Bagging is very robust to classification noise.**

Adaboost is putting large weights on the "noisy examples"



Mean weight per training example for the 560 corrupted training examples and the remaining 2,240 uncorrupted training examples in the sick data set.

Kappa-Error Diagrams

- **Kappa is a measure of agreement among different "judges"**.

  $= 1$ if there is complete agreement.

  $= 0$ if agreement/disagreement is random.

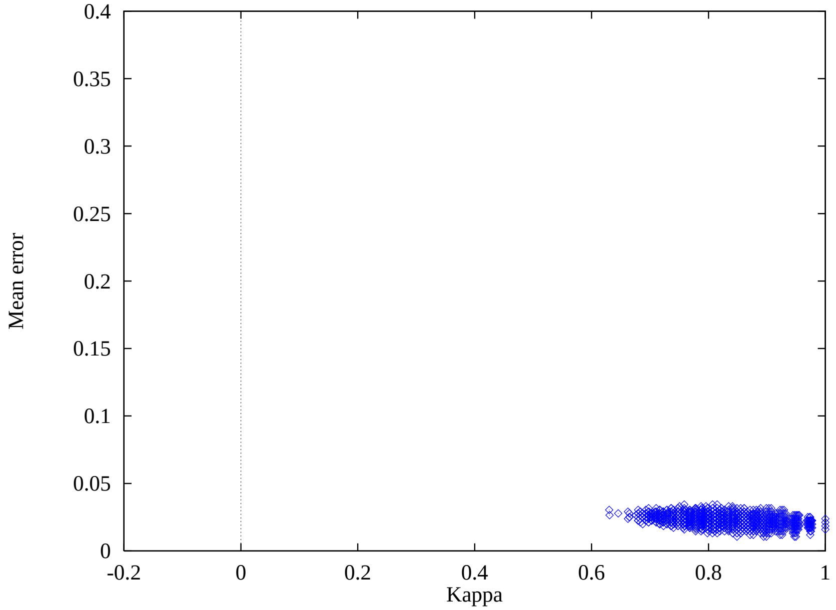
  $= -1$ if there is complete disagreement

- **For each pair of classifiers in an ensemble, we can plot a data point.**

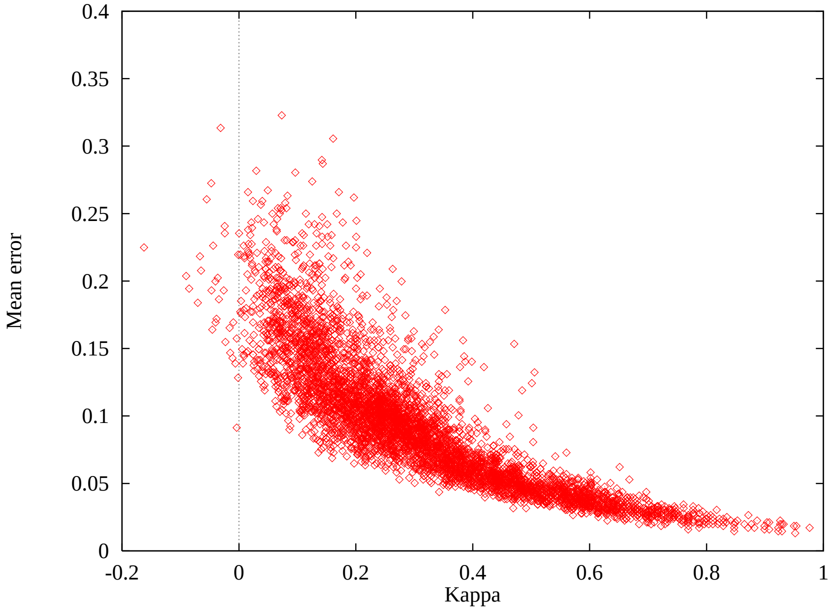
  The error rate is the average of the error rates of the two classifiers.

  The kappa value measures the degree of agreement between the two classifiers.

Kappa-Error Diagrams for SICK with no noise

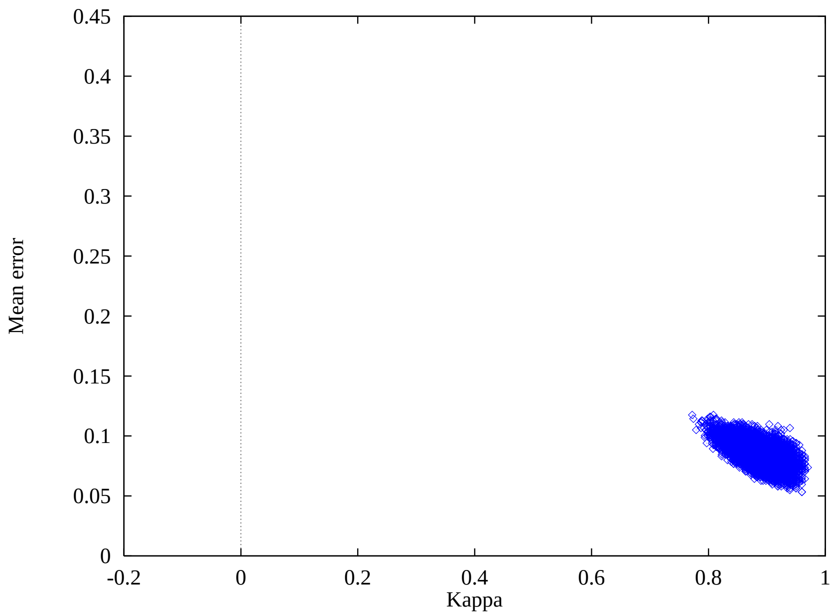


Bagging

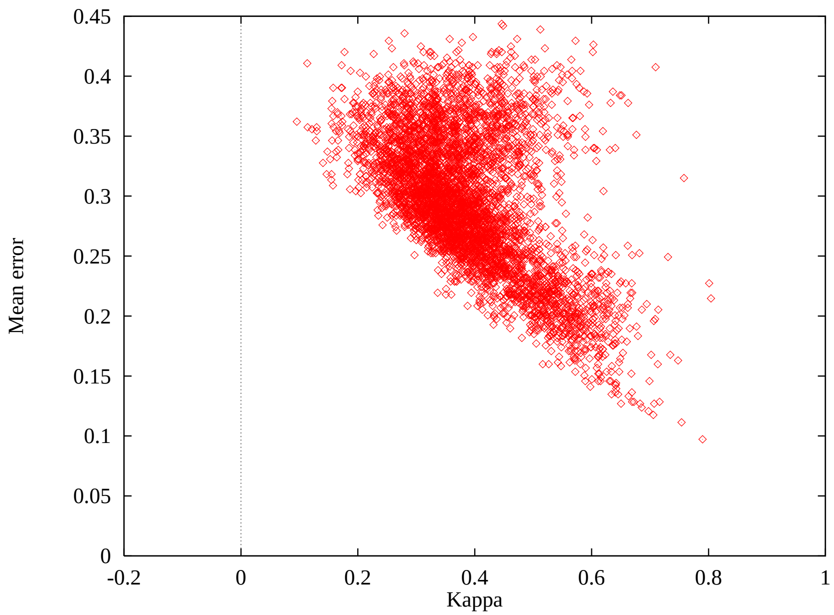


Adaboost

Kappa-Error Diagrams for SPLICE with no noise

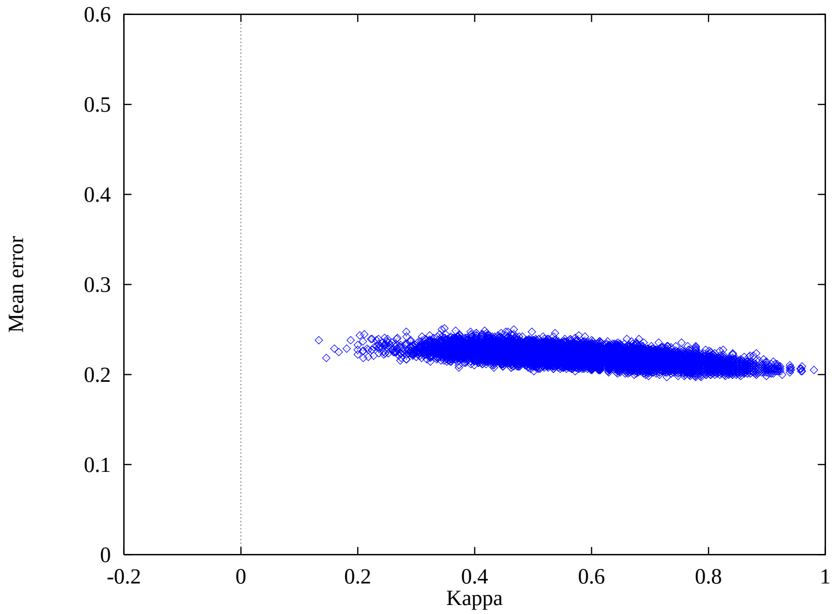


Bagging

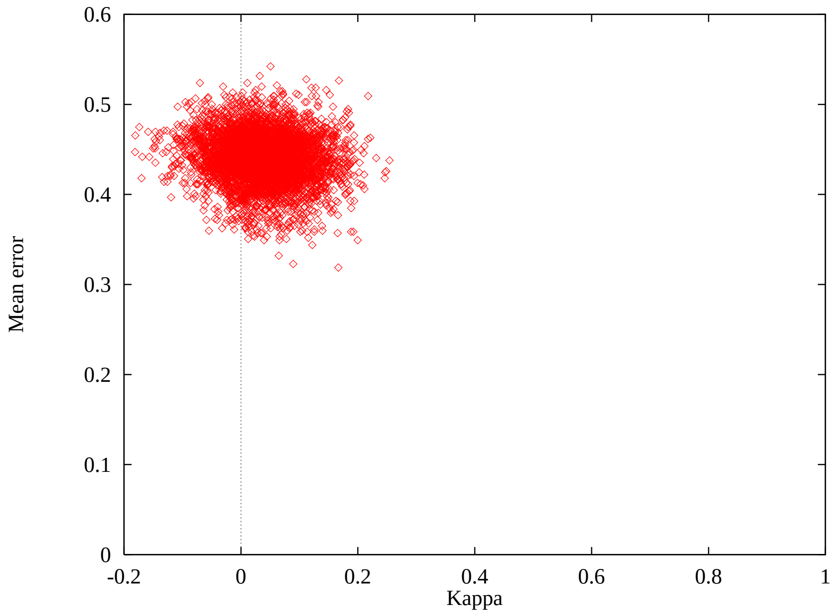


Adaboost

Kappa-Error Diagrams for SICK with 20% Class Noise

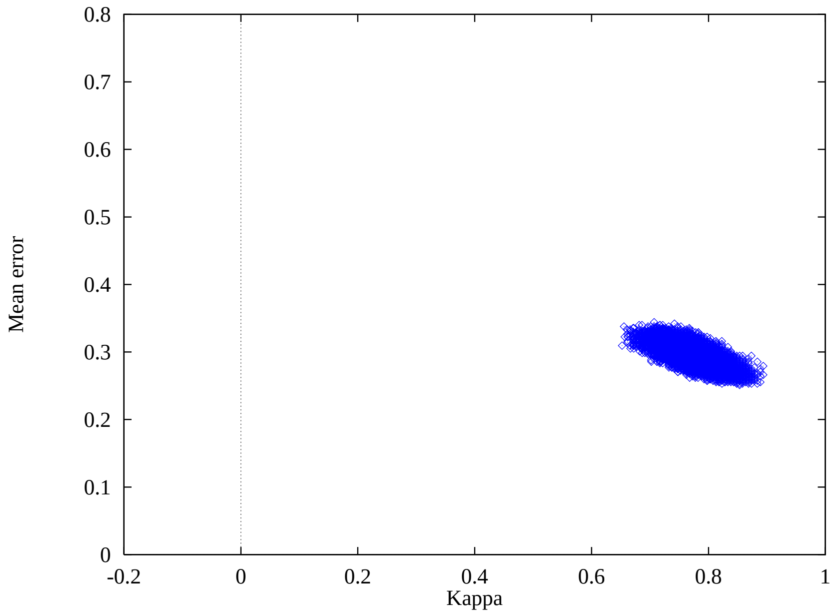


Bagging

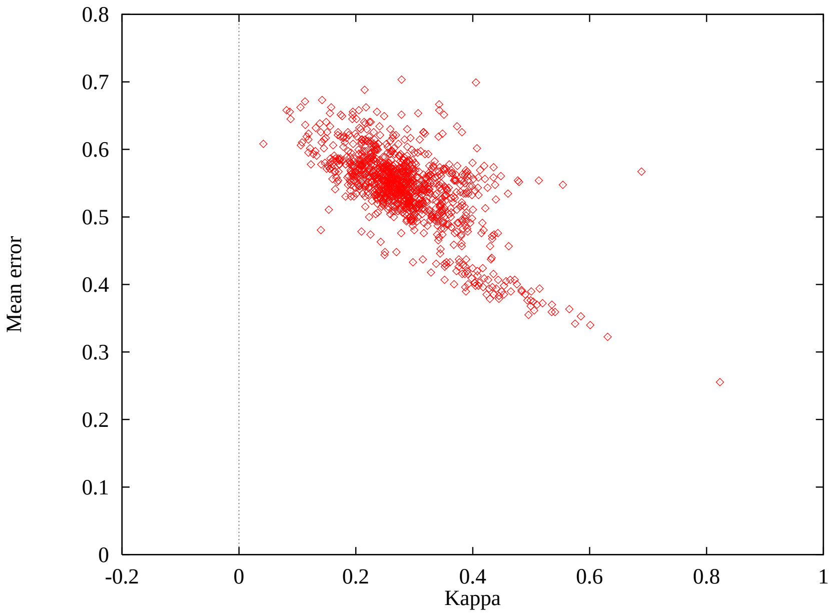


Adaboost

Kappa-Error Diagrams for SEGMENT with 20% Class Noise



Bagging



Adaboost

Comparison with Error-Correcting Coding

We designed good codes and applied the ECOC method in four domains. Here are the results at 20% class noise:

Domain	Bagging	ECOC
audiology	0.3856	0.4434
hypo	0.2062	0.2091
segment	0.2284	0.2297
satimage	0.2850	0.2909

In all cases, Bagging is superior, although the differences are only statistically significant in audiology. So there is no particular reason to favor ECOC in place of Bagging.

Conclusions

- Adaboost is best is low-noise domains
- Adaboost is devastated by high noise
- Bagging is very robust to noise
- Bagging beats or ties ECOC in four domains