# Linear and Order Statistics Combiners for Pattern Classification


Kagan Tumer

NASA Ames Research Center

`kagan@ptolemy.arc.nasa.gov`

`http://www.lans.ece.utexas.edu/~kagan/publications.html`



NIPS 1998

# Outline

- Motivation for Combining
- Model Error for Classification
- Linear Combining
  - Combining Classifiers with i.i.d. Errors
  - Correlated Classifier Combining
  - Correlation Reduction
- Order Statistics Combining
  - Linear Combinations of Order Statistics
- Bayes Error Rates and Confidences in Decisions

# Motivation for Combining

- Individual classifiers are sensitive to:
  - Classifier (model; size) selection.
  - Parameter (training method) selection.
  - Feature set selection.
- Different classifiers provide different generalizations.
- Eliminating less successful classifiers discards information.
- Remedy: Pool network outputs before making a decision.

# Historical Perspective

- Similar Concepts
  - Evidence Combining (Expert Reasoning).
  - Forecasting (Economics).
  - Multisensor Fusion (Pattern Recognition).
- Machine Learning
  - Neural Network Ensemble (Hansen and Salamon 90)
  - Boosting (Schapire 90)
  - Stacking (Wolpert 92)
  - Bagging (Breiman 93)

# Issues in Combining

- What to Combine?
  - Classifiers that approximate posteriors.
  - Weak Learners.
  - Classifiers trained on particular subsamples of the data.
- How to Combine?
  - Averaging.
  - Weighted averaging.
  - Entropy based.
  - Confidence based.
  - Order Statistics.

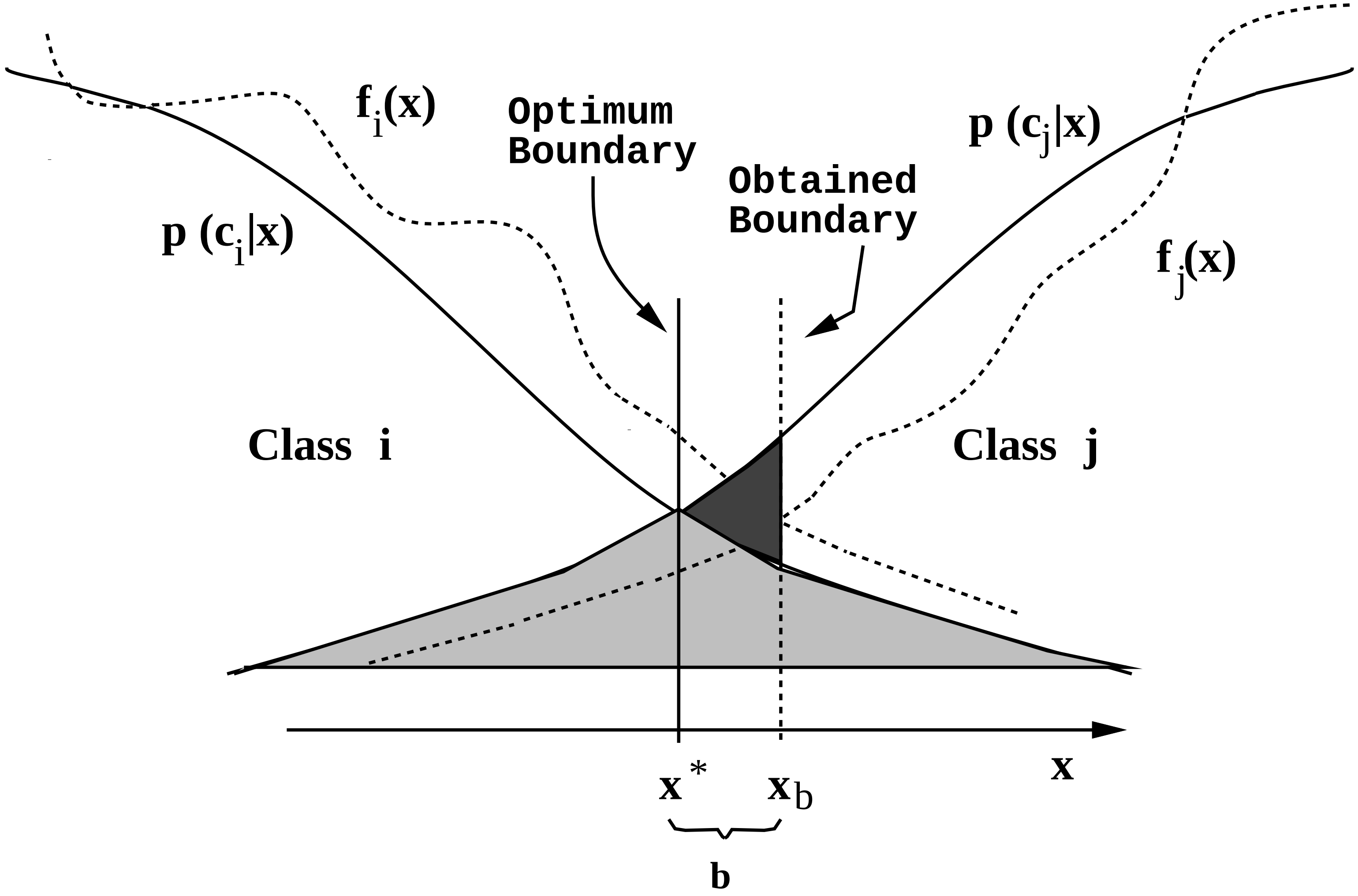
# Classifier Model

- Outputs of certain well-trained networks approximate *a posteriori* probabilities.
- Model classifier outputs as:

$$\boxed{f_i^m(x) = p(c_i|x) + \epsilon_i^m}$$

  - $f_i^m(x)$: response of the $m^{th}$ classifier on the $i^{th}$ class.
  - $p(c_i|x)$: a posteriori probability distribution of the $i^{th}$ class.
  - $\epsilon_i^m$: error of the $m^{th}$ classifier on the $i^{th}$ class.
    * Error can be split into bias and noise: $\epsilon_i^m(x) = \beta_i^m + \eta_i^m(x)$.

# Boundary Analysis (Single Classifier)



## Single Network Analysis

- Bayes boundary, $x^*$, given by: $p(c_i|x^*) = p(c_j|x^*)$ .
- Classifier provides $f_i(\cdot)$ instead of $p(c_i|\cdot)$ .
- Obtained boundary implies:

$$p(c_i|\, x^* + b) + \epsilon_i(x_b) \;=\; p(c_j|\, x^* + b) + \epsilon_j(x_b)$$

- First order approximation yields:

$$b \;=\; \frac{\epsilon_i(x_b) - \epsilon_j(x_b)}{s}$$

  – where: $s \;=\; p'(c_j|x^*) \;-\; p'(c_i|x^*)$.

# Model Error

- The model error:

$$A(b) \;=\; \int_{x^*}^{x^*+b} \left(p(c_j|x) - p(c_i|x)\right) dx \;=\; \frac{1}{2}\, b^2\, s$$

- Expected model error is the expected value of $A(b)$:

$$E_{model} = \frac{s}{2} \int_{-\infty}^{\infty} b^2 f_b(b) db$$

- For classifier without bias: $\boxed{E_{model} = \frac{s\sigma_b^2}{2}}$

- For classifier with bias: $\boxed{E_{model}(\beta) \;=\; \frac{s}{2}\left(\sigma_b^2 + \beta^2\right)}$

# Averaging Combiner ($N$ classifiers)

- Combiner output given by:

$$f_i^{ave}(x) = \frac{1}{N} \sum_{m=1}^{N} f_i^m(x) \; = \; p(c_i|x) \; + \; \bar{\beta}_i \; + \; \bar{\eta}_i(x)$$

- Model error (i.i.d. errors):
  - Classifiers are Unbiased ($\beta_i \; = \; 0$)

$$E_{model}^{ave} = \frac{s}{2} \, \sigma_{b^{ave}}^2 \; = \; \frac{s}{2 \, N} \, \sigma_b^2 \; = \; \frac{E_{model}}{N}$$

  - Classifiers are Biased ($\beta_i \; \neq \; 0$; $z = max(\frac{1}{\sqrt{N}}, \frac{\bar{\beta}}{\beta})$ )

$$E_{model}^{ave}(\bar{\beta}) \; \leq \; \frac{s}{2} \, z^2 \; \left(\sigma_b^2 + \beta^2\right) \; = \; z^2 \, E_{model}(\beta)$$

# Description of Oceanic Sonar Data.

- Two feature sets extracted from *same* raw data.
- Feature Set 1 (FS1) is 25-dimensional.
- Feature Set 2 (FS2) is 24-dimensional.

| Class Description | Feature Set 1 | | Feature Set 2 | |
|---|---|---|---|---|
| | Training | Test | Training | Test |
| Porpoise Sound | 116 | 284 | 142 | 284 |
| Cracking Ice | 116 | 175 | 175 | 175 |
| Whale Sound 1 | 116 | 129 | 129 | 129 |
| Whale Sound 2 | 148 | 235 | 118 | 235 |
| Total | 496 | 823 | 564 | 823 |

# Ave Combining Results for Sonar Data

| Classifier | | FS1 | | FS2 | |
|---|---|---|---|---|---|
| | N | Err (%) | $\sigma$ x$10^{-2}$ | Err (%) | $\sigma$ x$10^{-2}$ |
| MLP | 1 | 7.47 | 0.44 | 9.95 | 0.74 |
| | 3 | 7.19 | 0.29 | 9.32 | 0.35 |
| | 5 | 7.13 | 0.27 | 9.20 | 0.30 |
| | 7 | 7.11 | 0.23 | 9.07 | 0.36 |
| RBF | 1 | 6.79 | 0.41 | 10.94 | 0.93 |
| | 3 | 6.15 | 0.30 | 10.55 | 0.45 |
| | 5 | 6.05 | 0.20 | 10.43 | 0.30 |
| | 7 | 5.97 | 0.22 | 10.44 | 0.32 |
| BOTH | 3 | 6.11 | 0.34 | 8.46 | 0.57 |
| | 5 | 6.11 | 0.31 | 8.17 | 0.41 |
| | 7 | 6.08 | 0.32 | 8.14 | 0.28 |

# Average Combiner (Correlated)

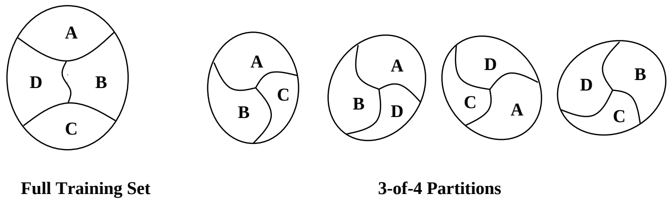
- Classifier errors in estimating posteriors are rarely independent.
- Boundary variance is affected by $\rho$, the correlation between classifiers.
- Model error is given by:

$$E^{ave}_{model} = \frac{1 + \rho\,(N-1)}{N}\, E_{model}.$$

## Decreasing Classifier Correlation

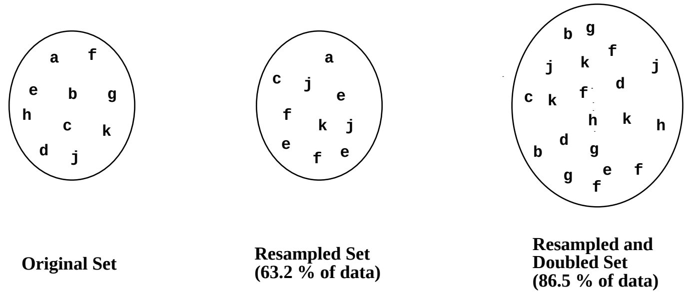
- $k$-fold cross validation.
- Bootstrapping: Randomly construct training set by resampling the original training set *with substitution.*
- Weighted Experts: Use spatial similarity to partition data.
- Input decimation: Present only some features to the classifier.
- Conclusion: Reduced correlation often compromised by poorer individual performance.

# $k-1$-of-$k$ Training (Cross-Validation)




- train $k$ classifiers on $\frac{k-1}{k}$ of the data.
- Selection of $k$ is critical for reducing correlations, and keeping individual classifier performances stable.

# Resampled Set Combining (Bootstrapping)




- Resample with substitution.
- Not limited to a specific number of partitions.
- Control correlations through resampled set size.

# Cross-Validation Combining Results for FS1

| Classifier(s) | | 2-of-3 Ave | | 6-of-7 Ave | | Original Sample | |
|---|---|---|---|---|---|---|---|
| | N | Error | $\sigma$ | Error | $\sigma$ | Error | $\sigma$ |
| MLP | 1 | 9.37 | 0.98 | 8.24 | 0.56 | 7.47 | 0.44 |
| | 3 | 8.21 | 0.35 | 7.59 | 0.38 | 7.19 | 0.29 |
| | 7 | 7.90 | 0.05 | 7.39 | 0.24 | 7.11 | 0.23 |
| RBF | 1 | 7.98 | 0.98 | 6.51 | 0.61 | 6.79 | 0.41 |
| | 3 | 6.16 | 0.34 | 5.95 | 0.27 | 6.15 | 0.30 |
| | 7 | 5.47 | 0.00 | 5.48 | 0.31 | 5.97 | 0.22 |

# Resampling results FS1

| Classifier(s) | | 63.2 Resample | | 85.4 Resample | | 100 (Re)sample | |
|---|---|---|---|---|---|---|---|
| | N | Error | $\sigma$ | Error | $\sigma$ | Error | $\sigma$ |
| MLP | 1 | 8.60 | 0.65 | 8.02 | 0.57 | 7.47 | 0.44 |
| | 3 | 7.47 | 0.41 | 7.36 | 0.34 | 7.19 | 0.29 |
| | 5 | 7.49 | 0.37 | 7.31 | 0.13 | 7.13 | 0.27 |
| | 7 | 7.47 | 0.27 | 7.27 | 0.17 | 7.11 | 0.23 |
| RBF | 1 | 8.15 | 0.71 | 6.70 | 0.76 | 6.79 | 0.41 |
| | 3 | 6.72 | 0.38 | 5.56 | 0.36 | 6.15 | 0.30 |
| | 5 | 6.22 | 0.33 | 5.29 | 0.28 | 6.05 | 0.20 |
| | 7 | 6.01 | 0.35 | 5.07 | 0.25 | 5.97 | 0.22 |

# Outline

- Motivation for Combining
- Model Error for Classification
- Linear Combining
  - Combining Classifiers with i.i.d. Errors
  - Correlated Classifier Combining
  - Correlation Reduction
- Order Statistics Combining
  - Linear Combinations of Order Statistics
- Bayes Error Rates and Confidences in Decisions

# Motivation for Order Statistics Ensemble

- Many ensemble methods, covering the spectrum from:
  - Averaging: Simple, but limited applicability.
  - Meta-learners: Powerful, but costly.
- Both methods are ill-suited for problems where most but not all classifiers perform well.
- Solution: Use Family of Order statistics.
  - Single Order Statistic: robust combining.
  - Trim Combiner: a better way to average.
  - Spread Combiner: a better way to *pick* min or max.
- Order Statistics: Balance between averaging and meta-learners.

## Practicality of Order Statistics

- Suppose a process (i.i.d.) generates $\{X_1, X_2, X_3, X_4, X_5 , X_6\}$
- Repeat process and obtain estimate for the mean and variance of each random variable:
  - Each $X_i$ has same variance ($\sigma_{X_i} = \sigma_{X_j} \; \forall i, j$).
  - Covariance between them is zero ($cov(X_i, X_j) = 0 \; \forall i, j$).
- Now, order each sequence $X^{1:6}$, $X^{2:6}$, $X^{3:6}$, $X^{4:6}$, $X^{5:6}$, $X^{6:6}$ and obtain mean and variance for each "new" variable.
- The variance depends on place in sequence and is less than or equal to that of the i.i.d samples.
- The 6 new variables are no longer independent.

# Order Statistics Combiners

- Order classifier output:

$$f_i^{1:N}(x) \leq f_i^{2:N}(x) \leq \cdots \leq f_i^{N:N}(x).$$

- $f^{j:N}$: $j$th value in progression, or $j$th order statistic.
- Define the *max*, *med*, and *min* combiners:

$$\begin{aligned}
f_i^{max}(x) &= f_i^{N:N}(x), \\
f_i^{med}(x) &= \begin{cases} \dfrac{f_i^{\frac{N}{2}:N}(x) + f_i^{\frac{N}{2}+1:N}(x)}{2} & \text{if } N \text{ is even} \\ f_i^{\frac{N+1}{2}:N}(x) & \text{if } N \text{ is odd,} \end{cases} \\
f_i^{min}(x) &= f_i^{1:N}(x).
\end{aligned}$$

# OS Combiners (Unbiased Classifiers)

- For unbiased classifiers, boundary offset has variance:

$$\sigma^2_{b^{os}} = \frac{2\,\sigma^2_{\eta_k^{os}}}{s^2} = \frac{2\,\alpha\sigma^2_{\eta_k}}{s^2} = \alpha\sigma^2_b$$

  – where $\alpha$ is a reduction factor that depends on the order statistic and on the distribution of $b$

- Model Error :

$$E^{os}_{model} = \alpha\; E_{model}$$

- $\alpha$ tabulated for various distributions.

# Reduction Factors ($\alpha$) for Gaussian Distribution

| N | $k$ | $\alpha$ | N | $k$ | $\alpha$ | N | $k$ | $\alpha$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1.00 | 6 | 2 (5) | .280 | 9 | 1 (9) | .357 |
| 2 | 1 (2) | .682 | | 3 (4) | .246 | | 2 (8) | .226 |
| 3 | 1 (3) | .560 | 7 | 1 (7) | .392 | | 3 (7) | .186 |
| | 2 | .449 | | 2 (6) | .257 | | 4 (6) | .171 |
| 4 | 1 (4) | .492 | | 3 (5) | .220 | | 5 | .166 |
| | 2 (3) | .360 | | 4 | .210 | 10 | 1 (10) | .344 |
| 5 | 1 (5) | .448 | 8 | 1 (8) | .373 | | 2 (9) | .215 |
| | 2 (4) | .312 | | 2 (7) | .239 | | 3 (8) | .175 |
| | 3 | .287 | | 3 (6) | .201 | | 4 (7) | .158 |
| 6 | 1 (6) | .416 | | 4 (5) | .187 | | 5 (6) | .151 |

# Med Ensembles for FS1

| Classifier | N | Ave Err (%) | Ave $\sigma$ x$10^{-2}$ | Med Err (%) | Med $\sigma$ x$10^{-2}$ |
|---|---|---|---|---|---|
| MLP | 1 | 7.47 | 0.44 | | |
| | 3 | 7.19 | 0.29 | 7.25 | 0.21 |
| | 5 | 7.13 | 0.27 | 7.30 | 0.29 |
| | 7 | 7.11 | 0.23 | 7.27 | 0.29 |
| RBF | 1 | 6.79 | 0.41 | | |
| | 3 | 6.15 | 0.30 | 6.42 | 0.29 |
| | 5 | 6.05 | 0.20 | 6.23 | 0.18 |
| | 7 | 5.97 | 0.22 | 6.25 | 0.20 |
| BOTH | 3 | 6.11 | 0.34 | 6.02 | 0.33 |
| | 5 | 6.11 | 0.31 | 5.76 | 0.29 |
| | 7 | 6.08 | 0.32 | 5.67 | 0.27 |

# Trimmed Mean Combiner

- Main Idea: Remove extremes and then average.

$$f_i^{trim}(x) = \frac{1}{M_2 - M_1 + 1} \sum_{m=M_1}^{M_2} f_i^{m:N}(x)$$

- Advantages:
  - Use reduction power of linear combination
  - Use reduction power of order statistics

## Error Reduction (Trimmed Mean Combiner)

$$\frac{E_{model}^{trim}}{E_{model}} = \frac{1}{(M_2 - M_1 + 1)^2} \left( \sum_{m=M_1}^{M_2} \alpha_{m:N} + 2 \sum_{m=M_1}^{M_2} \sum_{l>m} \beta_{m,l:N} \right)$$

- In averaging: $\alpha$'s equal 1, $\beta$'s equal 0:
  - Example: $N = 5$, $M_1 = 2$, $M_2 = 4$ (get rid of extremes). Error reduction for varios combiners:
    - Average ($N$): .2 (1/5)
    - Average ($M_2 - M_1 + 1$): .333 (1/3)
    - Trimmed Mean: .214
    - Median: .287

# Spread Combiner

- Main Idea: Use power of max and min

$$f_i^{spr}(x) = \frac{1}{2}\left(f_i^{1:N}(x) + f_i^{N:N}(x)\right)$$

- Advantages:
  - Use best of max min
  - No additional training
  - No additional parameters

# Error Reduction (Spread Combiner)

$$\frac{E^{spr}_{model}}{E_{model}} = \frac{\alpha_{1:N} + \beta_{1,N:N}}{2}$$

- Compare to min or max:
  - Example: For $N = 5$:
    - Min or Max provide error reduction of .448
    - Spread provides error reduction of .261
  - Example: For $N = 9$:
    - Min or Max provide error reduction of .357
    - Spread provides error reduction of .194

# Spread Combiner Reduction Factors

| N | $E^{spr}_{model}/E_{model}$ | $E^{minormax}_{model}/E_{model}$ |
|---|---|---|
| 2 | .500 | .682 |
| 3 | .362 | .560 |
| 4 | .299 | .492 |
| 5 | .261 | .448 |
| 6 | .236 | .416 |
| 7 | .219 | .392 |
| 8 | .205 | .373 |
| 9 | .194 | .357 |
| 10 | .186 | .344 |

# Results with Variability in Classifiers (N=8)

| Data | Ave | Max | Min | Spread | Trim ($M_1$-$M_2$) |
|---|---|---|---|---|---|
| Cancer | $1.32 \pm .26$ | $1.44 \pm .29$ | $1.44 \pm .29$ | $1.44 \pm .29$ | $1.32 \pm .26$ (2-6) |
| Card | $13.66 \pm .39$ | $13.08 \pm .29$ | $13.02 \pm .29$ | $12.97 \pm .26$ | $13.20 \pm .37$ (7-8) |
| Diabetes | $24.84 \pm .74$ | $25.05 \pm .68$ | $25.05 \pm .68$ | $25.05 \pm .68$ | $24.84 \pm .62$ (6-8) |
| Gene | $12.89 \pm .45$ | $12.76 \pm .51$ | $12.41 \pm .21$ | $12.43 \pm .46$ | $12.56 \pm .42$ (7-8) |
| Glass | $33.96 \pm .00$ | $39.43 \pm .57$ | $33.77 \pm .57$ | $33.40 \pm .36$ | $33.77 \pm .57$ (1-6) |
| Soybean | $7.65 \pm .00$ | $7.82 \pm .27$ | $13.41 \pm 1.1$ | $7.71 \pm .32$ | $7.65 \pm .00$ (4-8) |
| Sonar 1 | $11.64 \pm .38$ | $11.47 \pm .45$ | $11.29 \pm .57$ | $11.51 \pm .37$ | $12.30 \pm .36$ (4-5) |
| Sonar 2 | $8.82 \pm .35$ | $7.68 \pm .47$ | $8.91 \pm .28$ | $8.24 \pm .45$ | $7.81 \pm .33$ (7-8) |

# Results with Homogeneous Classifiers (N=8)

| Data | Ave | Max | Min | Spread | Trim ($M_1$-$M_2$) |
|---|---|---|---|---|---|
| Cancer | $0.69 \pm .23$ | $0.57 \pm .02$ | $0.57 \pm .02$ | $0.57 \pm .02$ | $0.57 \pm .23$ (7-8) |
| Card | $13.14 \pm .47$ | $12.79 \pm .02$ | $12.79 \pm .02$ | $12.79 \pm .02$ | $12.80 \pm .02$ (7-8) |
| Diabetes | $22.92 \pm .47$ | $23.23 \pm .71$ | $23.12 \pm .71$ | $23.23 \pm .71$ | $22.92 \pm .47$ (4-8) |
| Gene | $12.26 \pm .30$ | $12.46 \pm .38$ | $12.16 \pm .17$ | $12.11 \pm .40$ | $12.16 \pm .19$ (1-6) |
| Glass | $32.08 \pm .01$ | $32.08 \pm .01$ | $32.08 \pm .01$ | $32.08 \pm .01$ | $32.08 \pm .01$ (3-6) |
| Soybean | $7.06 \pm .00$ | $7.18 \pm .23$ | $9.06 \pm 1.7$ | $7.06 \pm .00$ | $7.06 \pm .00$ (3-6) |

| | | | | | |
|---|---|---|---|---|---|
| Sonar 1 | $8.94 \pm .13$ | $9.62 \pm .34$ | $9.36 \pm .32$ | $9.48 \pm .37$ | $8.92 \pm .21$ (1-6) |
| Sonar 2 | $7.17 \pm .17$ | $7.19 \pm .25$ | $7.41 \pm .34$ | $7.22 \pm .15$ | $7.07 \pm .21$ (2-6) |

## Summary of Order Statistics Ensembles

- Error reduction is quantified (in tabular form).
- When classifiers are not homogeneous *med* outperforms *ave*.
- Spread and Trim blend simplicity of averaging and the generality of meta-learners.
- When performance if sample dependent, order statistics combiners select the *correct* samples.

# Conclusions

This work:

- Analyzes the effect of combining on classification problems (linear/nonlinear combiners).
- Ties the classifier error the decision boundary variance.
- Introduces order statistics combiners that are
  - powerful;
  - computationally easy to use.
- Explores correlation reduction methods.
- Presents a new way of estimating the Bayes error.