Experiments with Classifier Combining Rules

Robert P.W.Duin and David M.J. Tax

Pattern Recognition Group, Department of Applied Physics

Delft University of Technology, The Netherlands

Cagliari, June 2000

P.O. Box 5046, 2600GA Delft, The Netherlands.
Phone: +(31) 15 2786143, FAX: +(31) 15 2786740,
E-mail: duin@ph.tn.tudelft.nl

		Feature Set							Fixed Combiners						Trained Combiners			
Classifiers		All (649)	Fourier (76)	Profiles (216)	KL-coef (64)	Pixels (240)	Zernike (47)	Morph (6)	Maximum	Median	Mean	Minimum	Product	Majority	Bayes-2	Bayes-1	NMean	1-NN
Basic Classifiers	Bayes-2	52	257	58	128	62	212	310	200	28	63	84	63	68	701	799	67	50
	Bayes-1	183	213	34	57	99	180	291	52	37	34	44	31	51	75	99	39	42
	NMean	87	224	181	99	96	278	540	540	62	45	80	46	75	124	20	103	46
	1-NN	36	192	90	44	37	197	570	569	26	17	35	17	40	46	29	113	30
	k-NN	36	189	92	44	37	193	510	192	54	60	82	42	51	97	27	36	26
	Parzen	36	171	79	37	37	185	521	37	29	32	29	27	51	36	37	31	31
	Fisher	408	248	47	82	153	210	282	39	32	33	65	52	57	48	45	35	36
	Dec. Tree	480	454	403	400	549	598	329	275	134	113	262	110	218	283	104	102	108
	ANN-20	896	900	46	146	852	900	328	900	177	784	900	900	327	45	27	26	21
	ANN-50	900	245	130	823	810	265	717	692	244	290	805	807	163	42	31	55	33
	SVC-1	500	246	66	61	77	294	848	123	108	59	190	101	47	74	69	60	58
	SVC-2	96	212	51	40	60	193	811	40	36	37	42	38	38	40	40	40	40

Combining Classifiers		Feature Set							Fixed Combiners						Trained Combiners			
		All (649)	Fourier (76)	Profiles (216)	KL-coef (64)	Pixels (240)	Zernike (47)	Morph (6)	Maximum	Median	Mean	Minimum	Product	Majority	Bayes-2	Bayes-1	NMean	1-NN
Fixed Combiners	Maximum	105	747	42	39	44	839	436	39	22	21	53	20	76	33	27	19	19
	Median	34	190	43	36	45	174	287	50	23	25	39	25	50	195	58	19	50
	Mean	53	190	35	45	56	176	285	34	20	21	37	20	46	57	34	20	19
	Minimum	315	790	137	109	200	737	652	138	165	135	131	135	160	58	71	72	56
	Product	215	294	131	44	82	401	412	86	234	86	84	86	568	47	860	851	685
	Majority	33	175	35	32	37	169	318	34	23	20	122	20	48	198	27	21	20
Trained Combiners	Bayes-2	104	273	49	44	99	195	289	244	17	115	133	115	28	822	897	129	64
	Bayes-1	60	427	51	40	53	190	294	160	24	41	49	41	26	107	656	56	63
	NMean	32	198	37	46	73	181	266	133	20	19	36	19	51	79	42	15	18
	1-NN	33	186	38	41	72	170	328	212	18	18	38	18	41	49	36	19	18

Points of Interest

Two ways of combining classifiers

Combining arbitrary classifiers

Fixed and trained combining rules

Splitting the feature set

Combining combiners

Performance of classifiers and combiners

The Data

Handwritten digits (binary, 30 x 48 pixels)

10 Classes: 0, 1, ... , 9. 200 Objects / class.

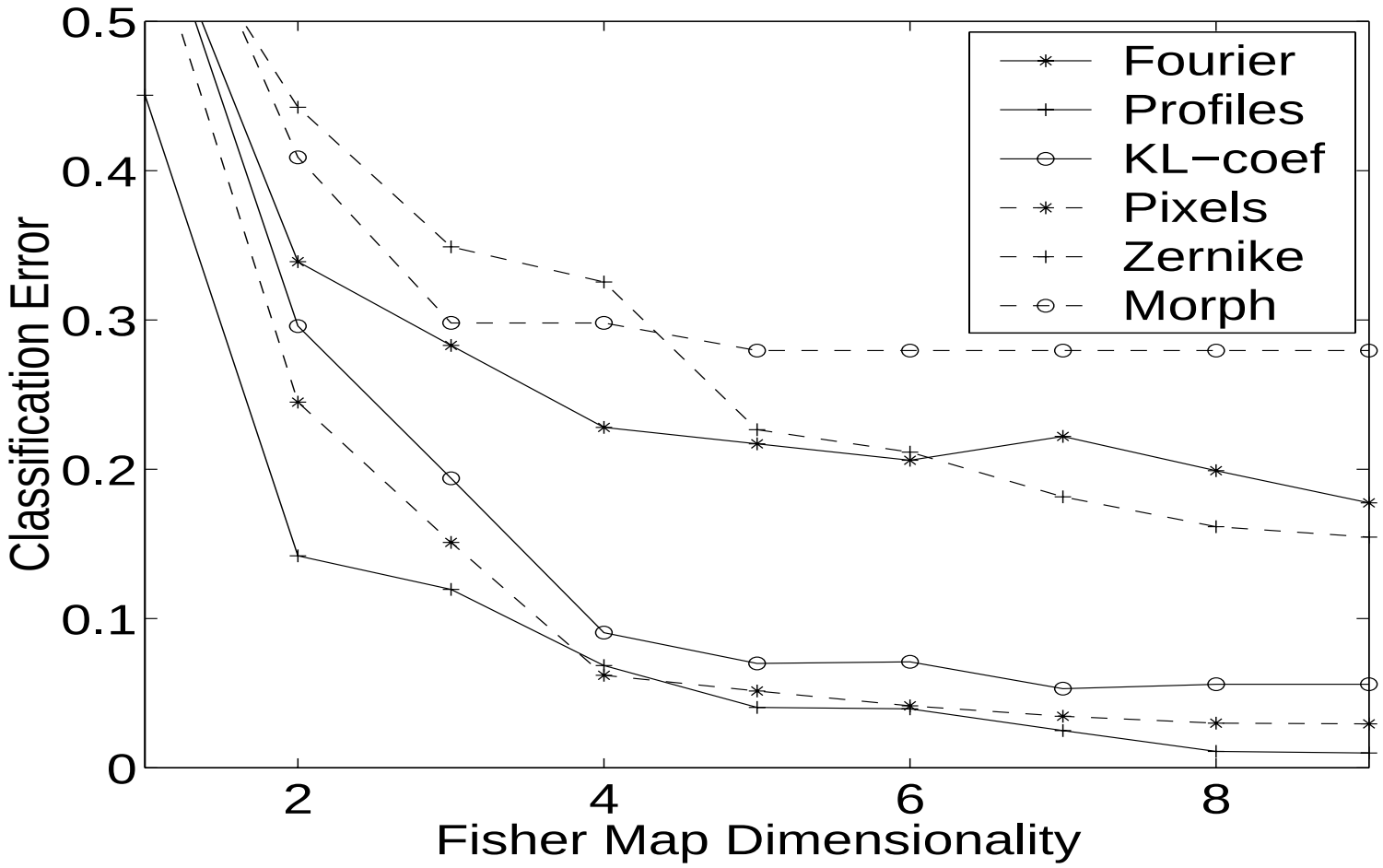
Training: 50 x 10 objects. Testing: 100 x 10 objects.

6 Feature sets: (total of 649 features)

- Fourier: 76 Fourier coefficients of the character shapes.
  6 profile correlations.
- KL-coef: 64 Karhunen-Loève coefficients.
- Pixel: 240 pixel averages in 2 x 3 windows.
- Zernike: 47 Zernike moments.
- Morph: 6 morphological features.

See Machine Learning Repository: mfeat

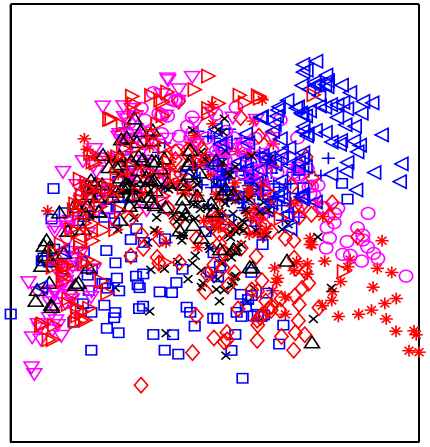
Feature Set Performance



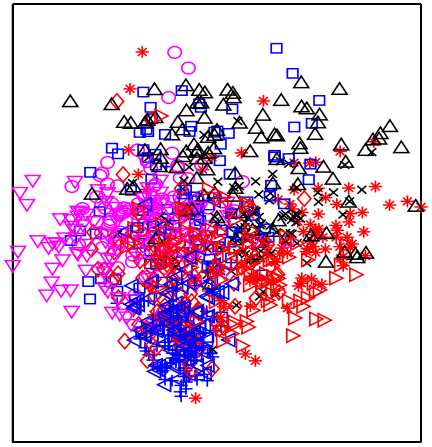
The Fisher classification error for the six feature sets optimally projected on low-dimensional subspaces.

PCA Plots

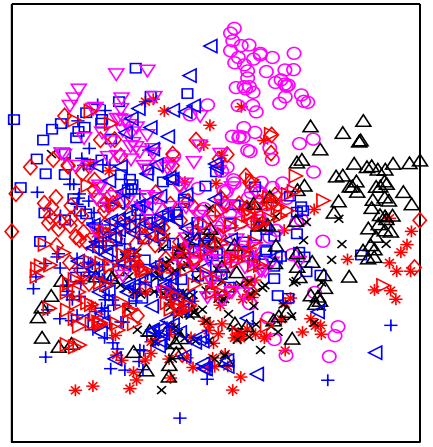
Fourier



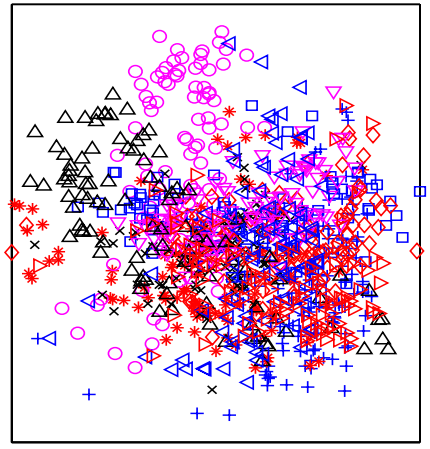
Profiles



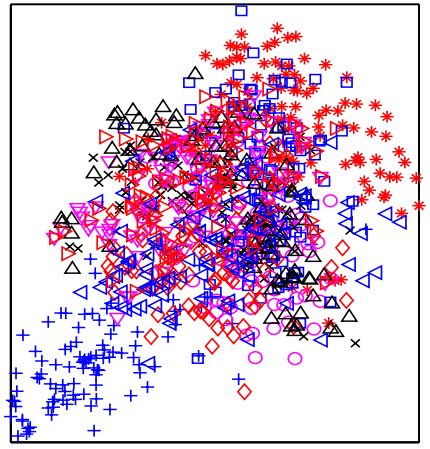
KL



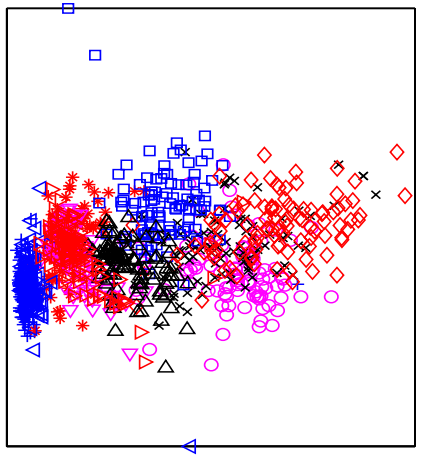
Pixels



Zernike

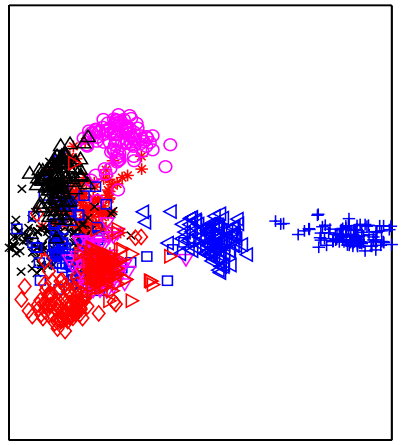


Morph

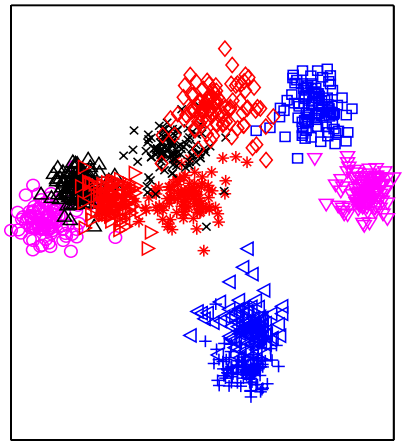


Fisher Map Plots

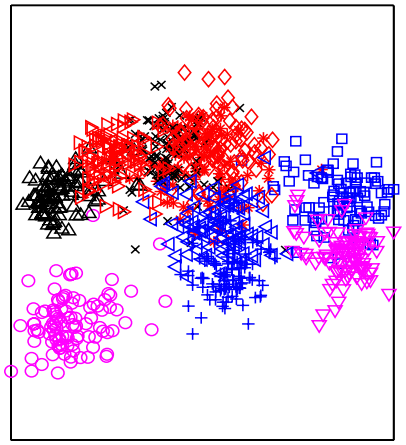
Fourier



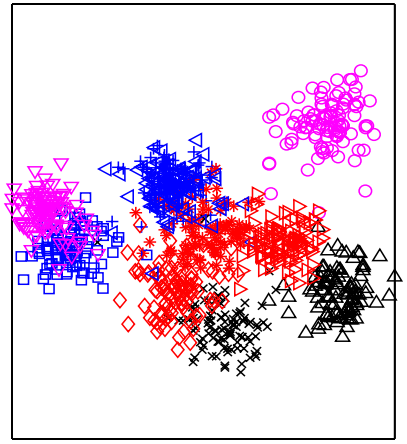
Profiles



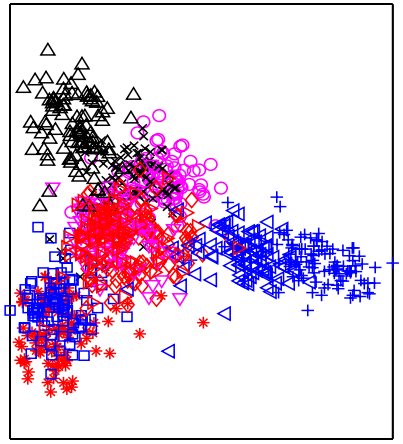
KL



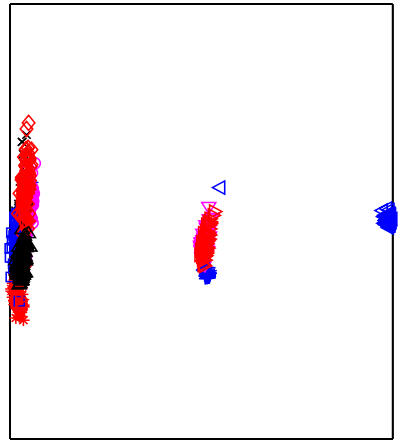
Pixels



Zernike

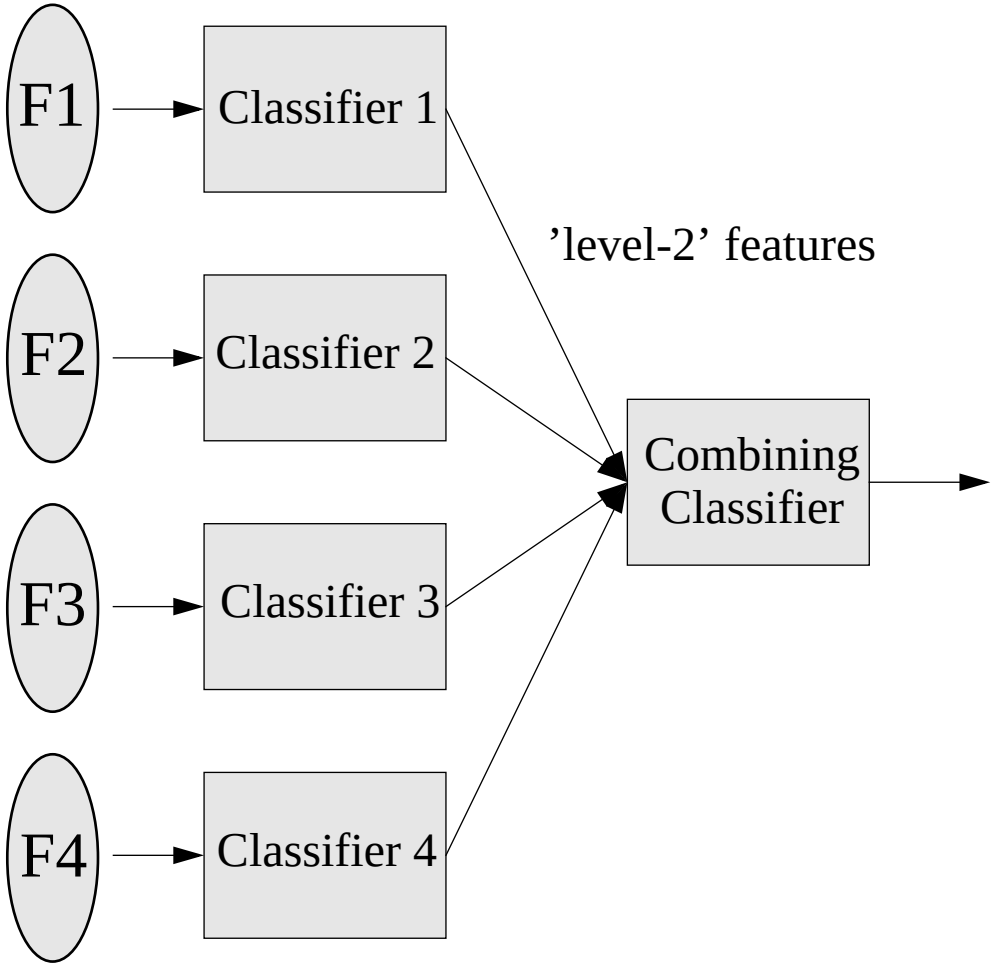


Morph

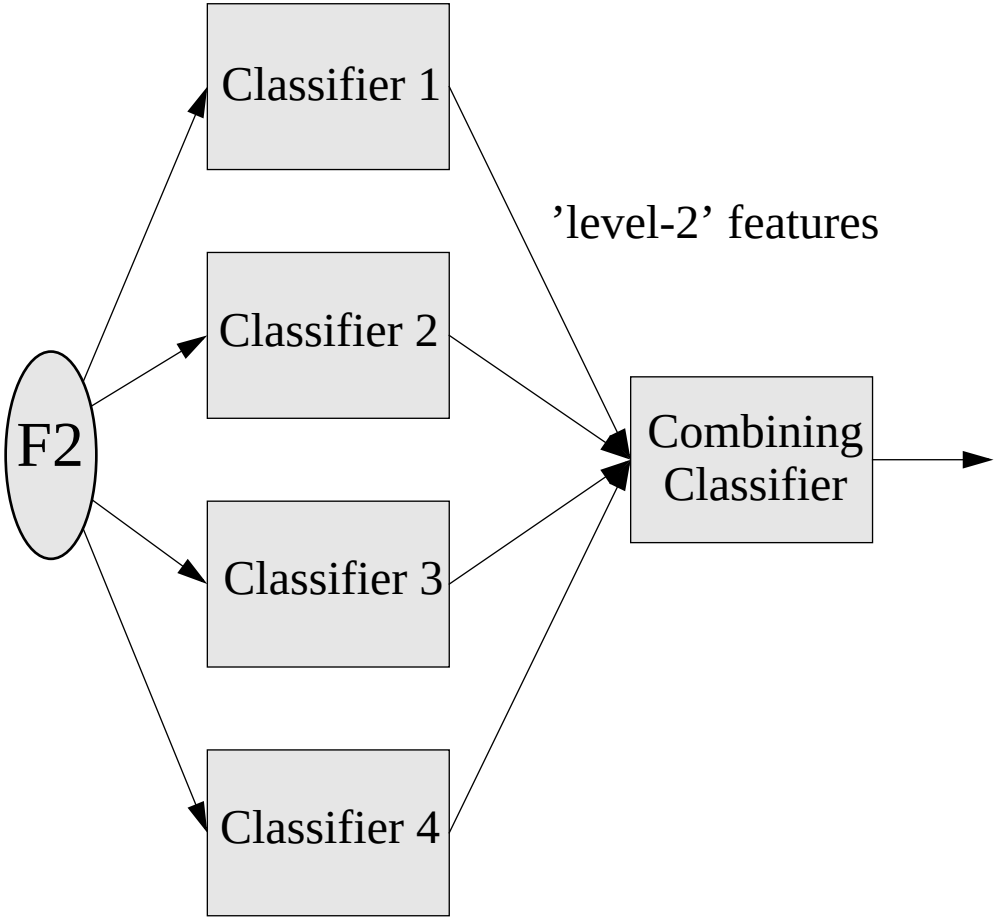


Combining Classifiers over a Different Feature Sets

Different feature sets



Combining Different Classifiers over a Single Feature Set



Single feature set, different classifiers

Combining Different Classifiers

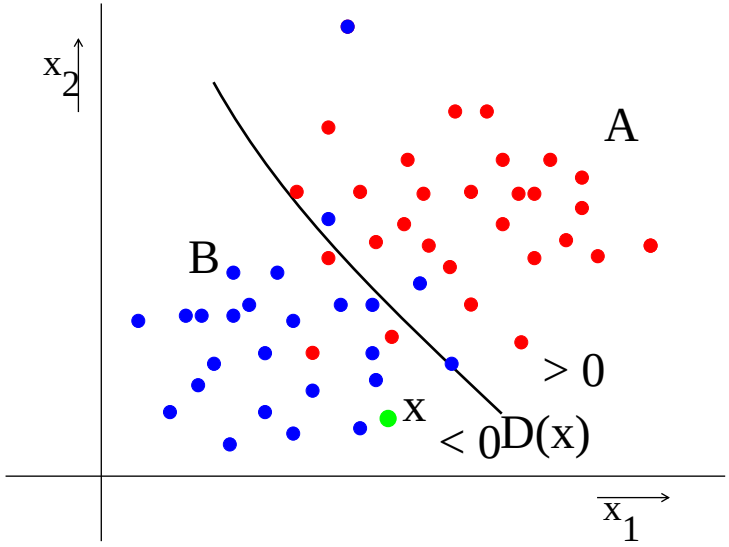
Compute comparable confidence vectors $\mathbf{c(x)} = [c_1(x), c_2(x), ..., c_{10}(x)]$.

1. Density based classifiers: $c_j(\mathbf{x}) = \dfrac{P_j f_j(\mathbf{x})}{\sum_i P_i f_i(\mathbf{x})}$

2. Confidence output classifiers, e.g. neural networks.

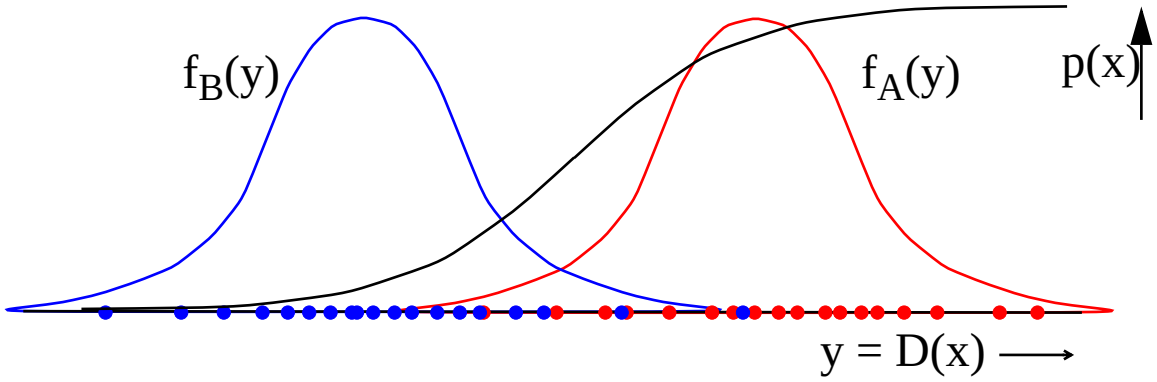
3. Distance based classifiers: discriminants, $c_j(x) = S(D(x))$

Confidences for Discriminants



Classifier Conditional Posterior Probabilities

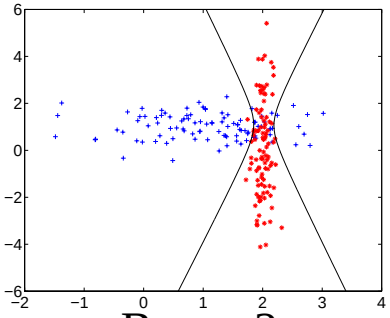
$$c(x) = \text{Prob}\,(\, x \in A \mid y=D(x)\,) = \frac{p_A f_A(y)}{f(y)}$$

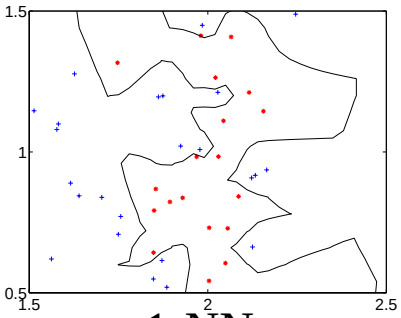


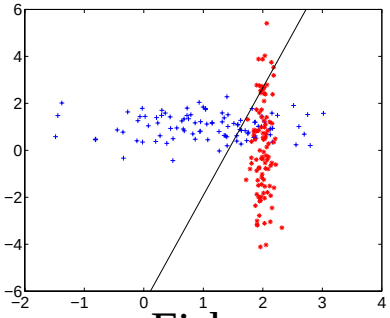
Fit a sigmoid, or a logistic function to the data y = D(x),

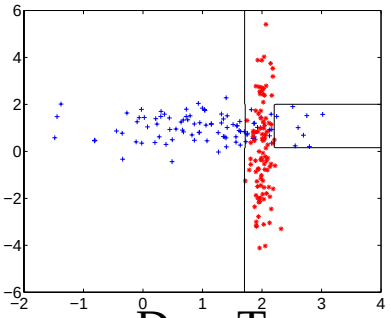
such that $\prod_i c(x_i)$ is maximized restricted to p(x) = 0.5 for D(x) = 0.

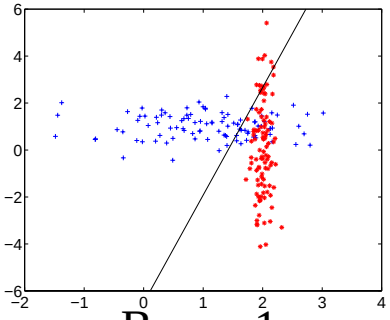
The Classifiers

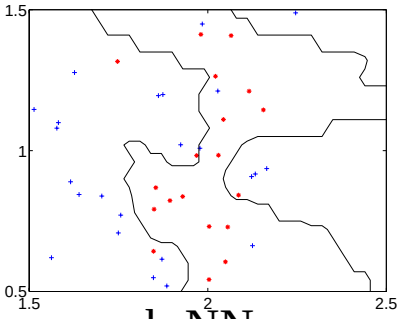

Bayes-2

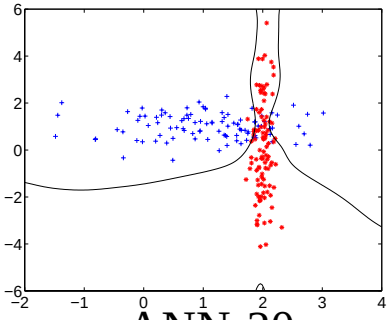

1-NN

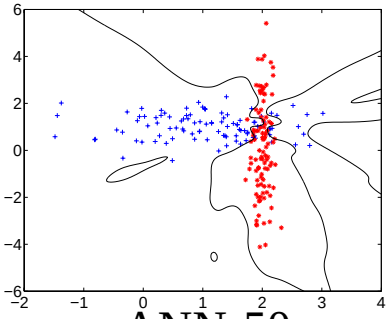

Fisher

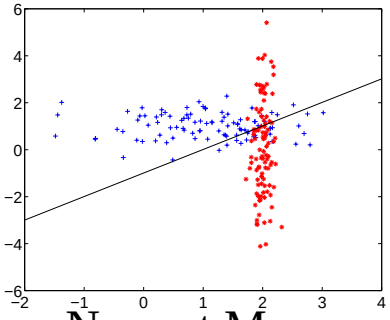

Dec. Tree

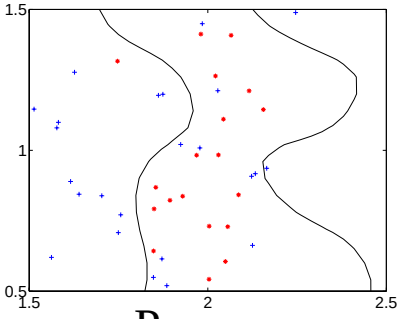

Bayes-1

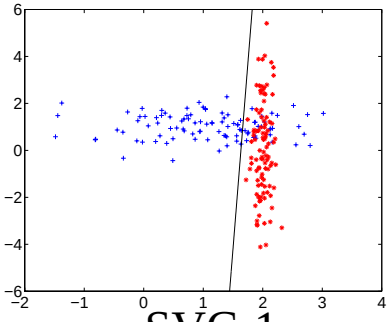

k-NN

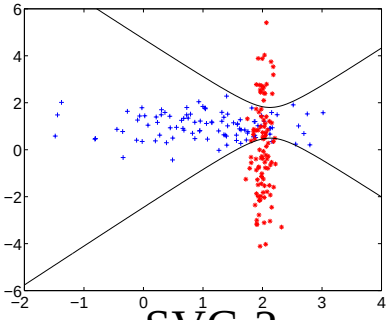

ANN-20


ANN-50

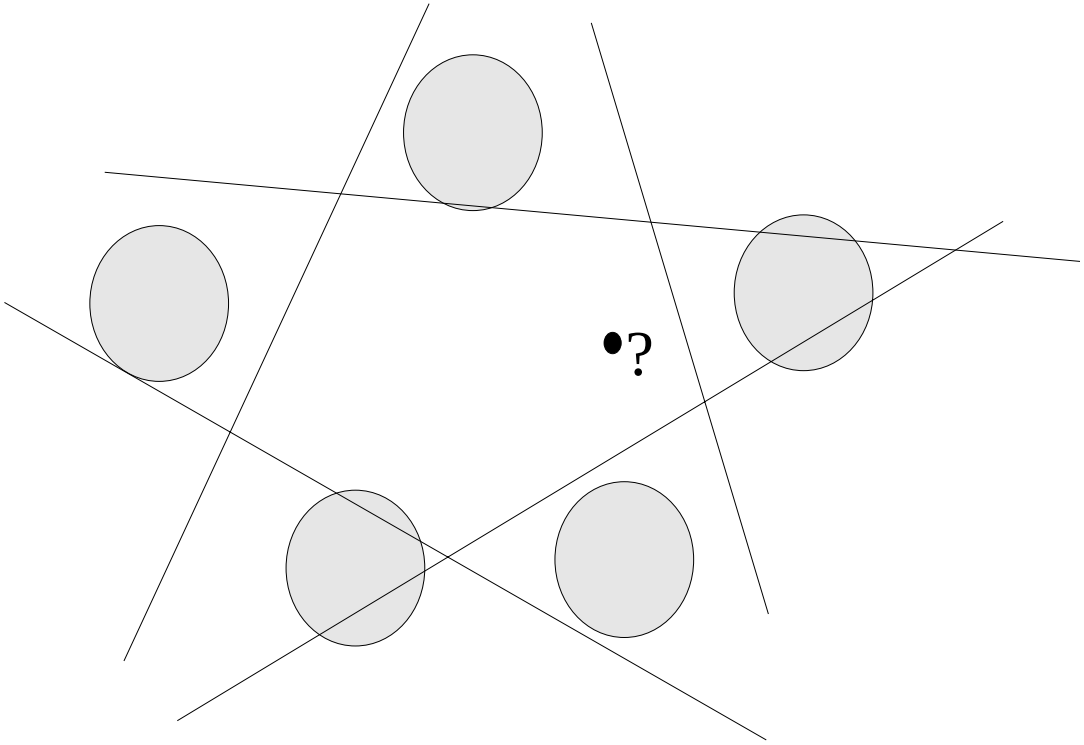

Nearest Mean

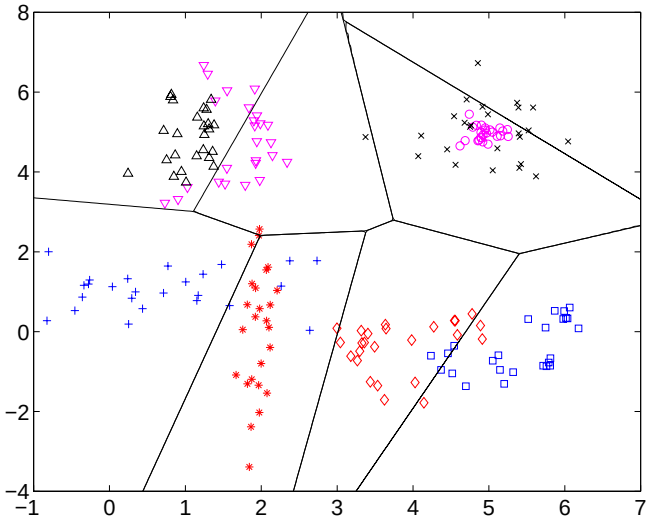

Parzen


SVC-1

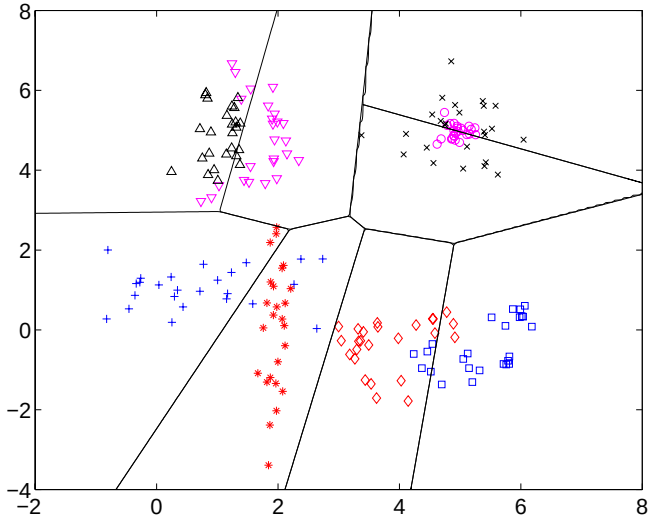

SVC-2

Multi-Class Classification by 2-Class Discriminants





Fisher



Bayes-1

Combine 2-class discriminants and assign to class with maximum confidence

Fixed Combining Rules

Combining n classifiers over 10 classes:

$c_{ij}(\mathbf{x})$ is the confidence of classifier i (i=1,n) for class j (j=1,10):

$$q_j'(\mathbf{x}) = \max_i (c_{ij}), \quad q_j(x) = \frac{q_j'(\mathbf{x})}{\sum_{j=1}^{10} q_j'(\mathbf{x})}$$

Combining Rules: Max, Median, Mean, Min, Product, Majority

Trained Combining Rules

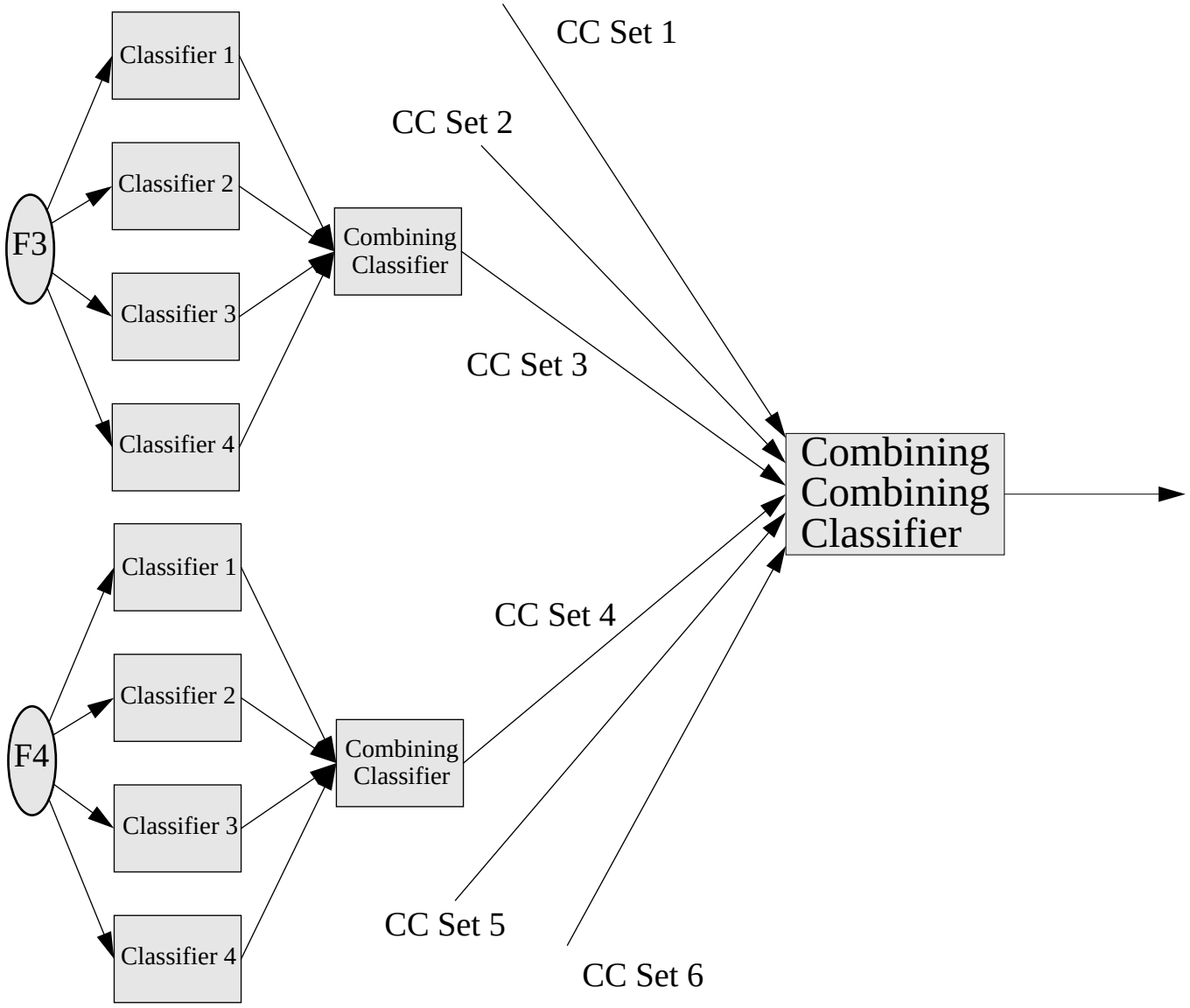
Combining n classifiers over 10 classes:

$c_{ij}(\mathbf{x})$ is the confidence of classifier i (i=1,n) for class j (j=1,10).

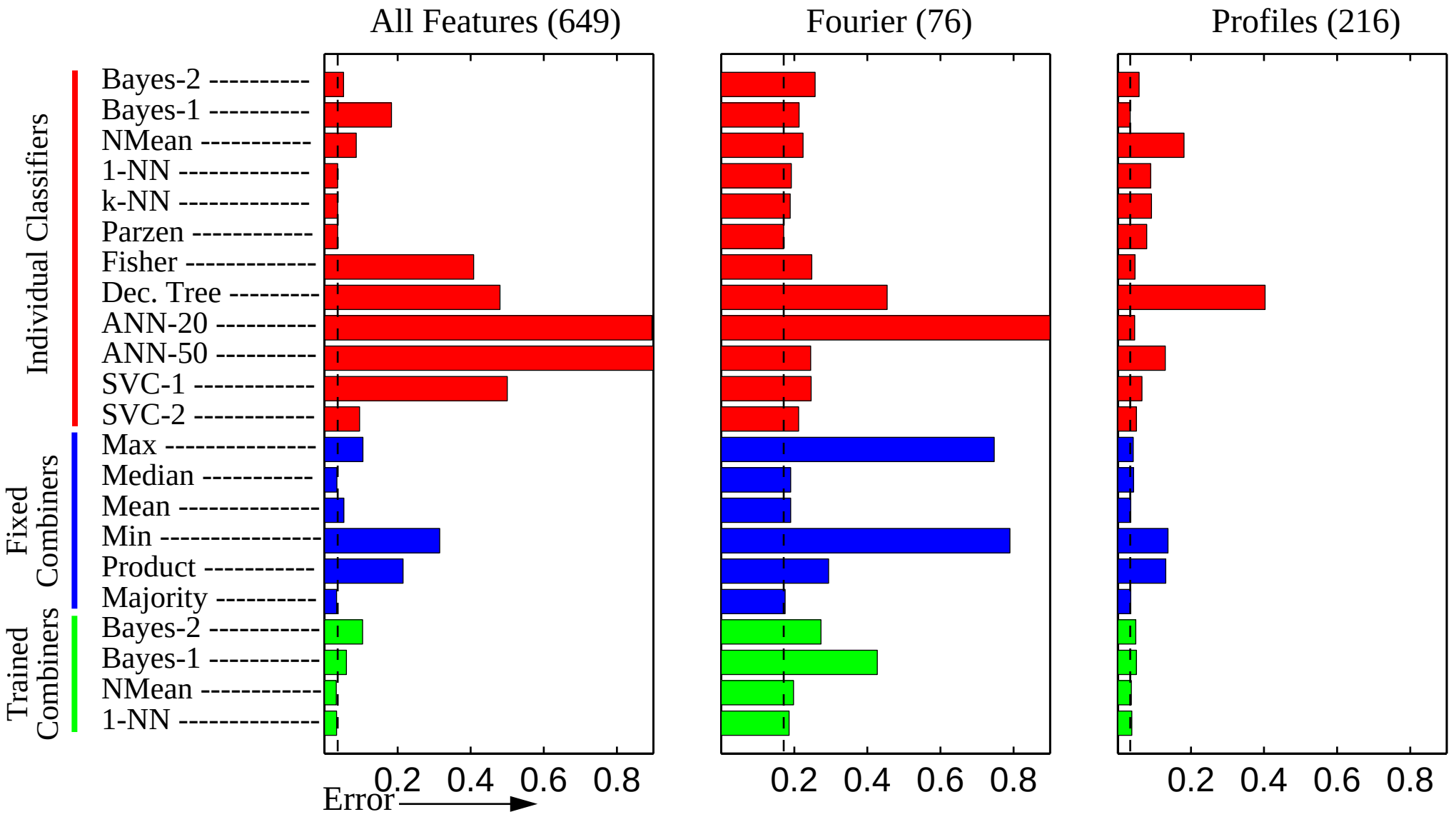
Use $c_{ij}(\mathbf{x})$ as a set of 10 x n features for object **x**.

Train by Bayes-2, Bayes-1, Nearest Mean and 1-NN

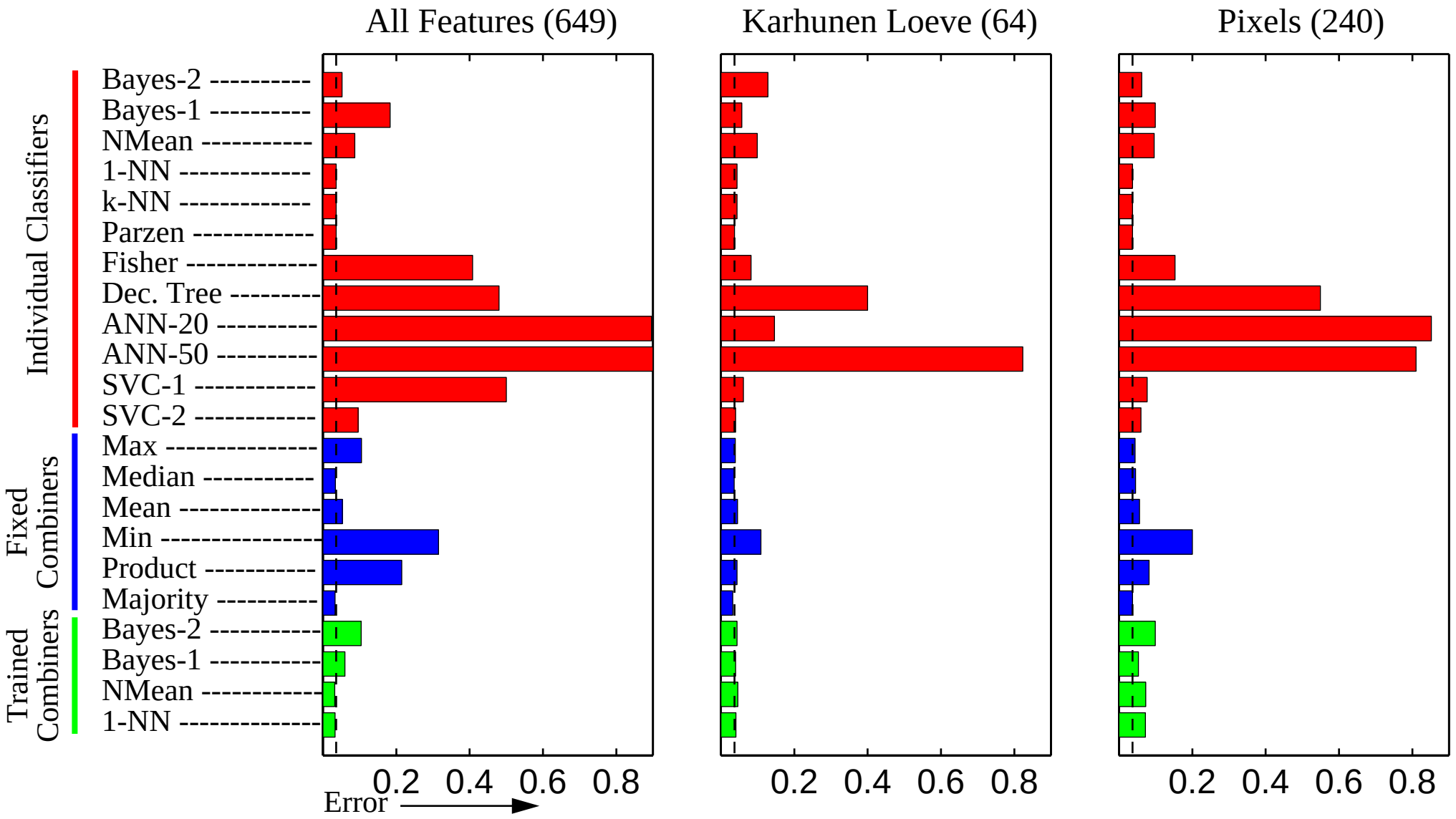
Combining Combiners



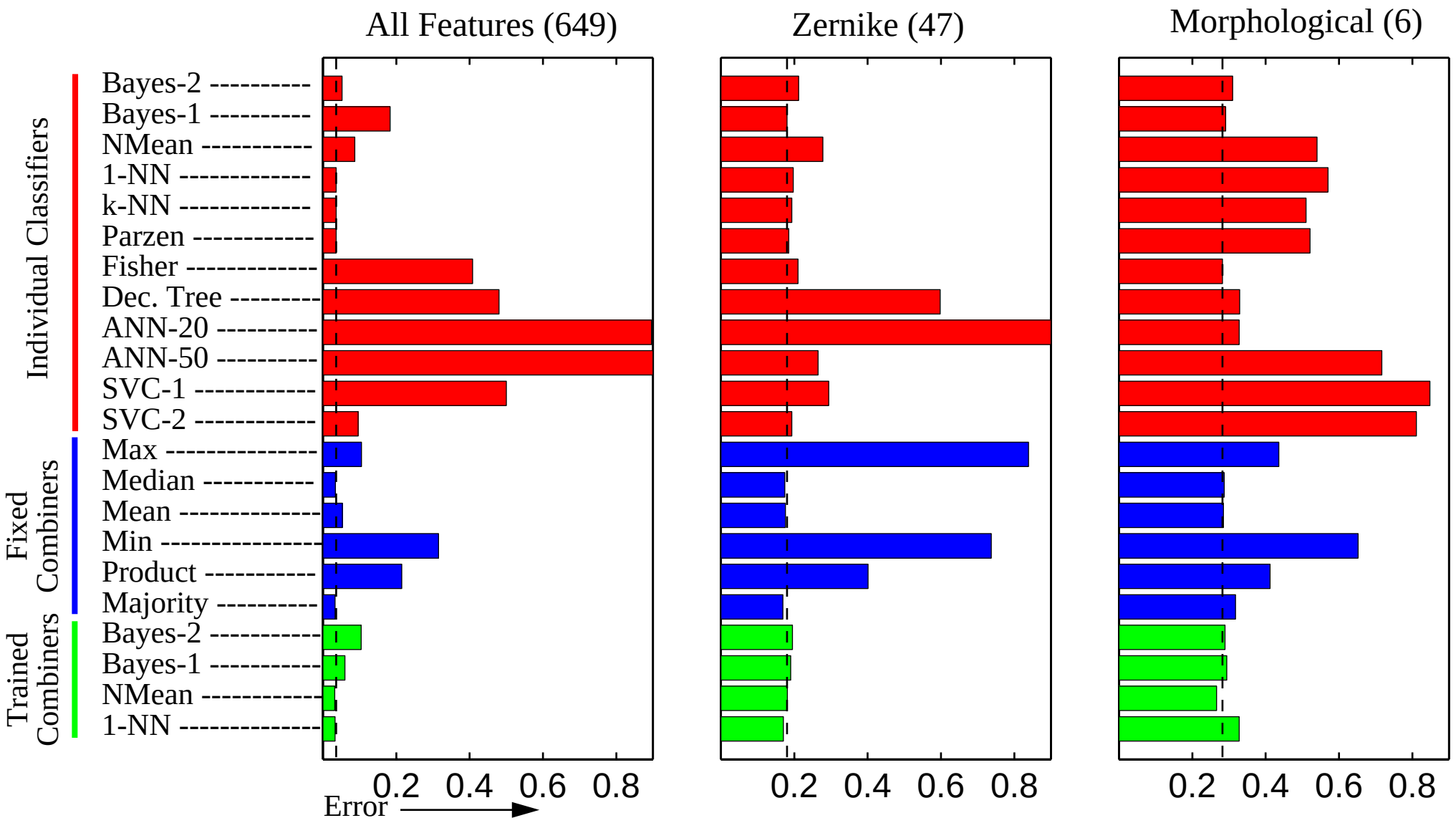
Combining Different Classifiers (1)



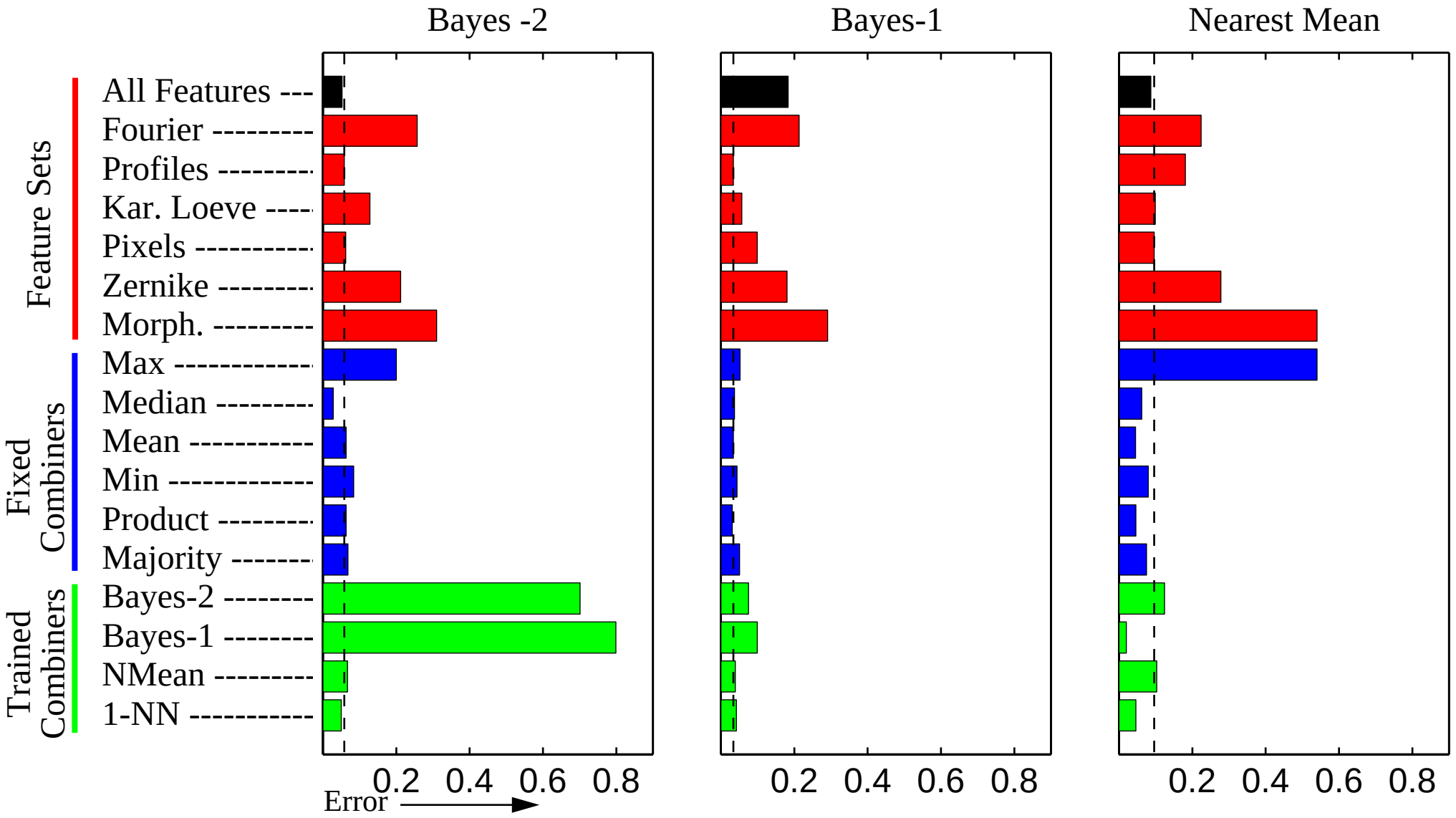
Combining Different Classifiers (2)



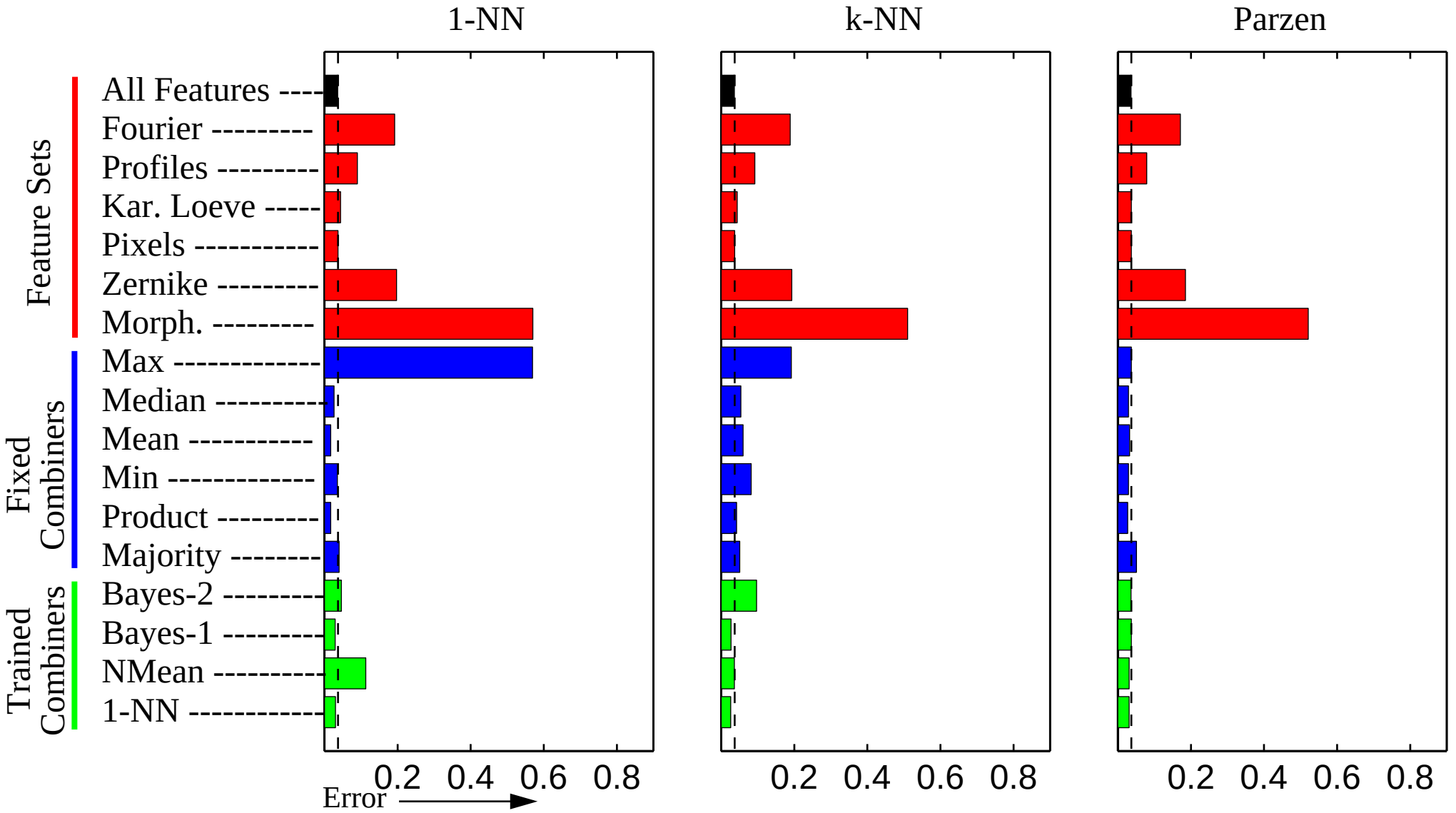
Combining Different Classifiers (3)



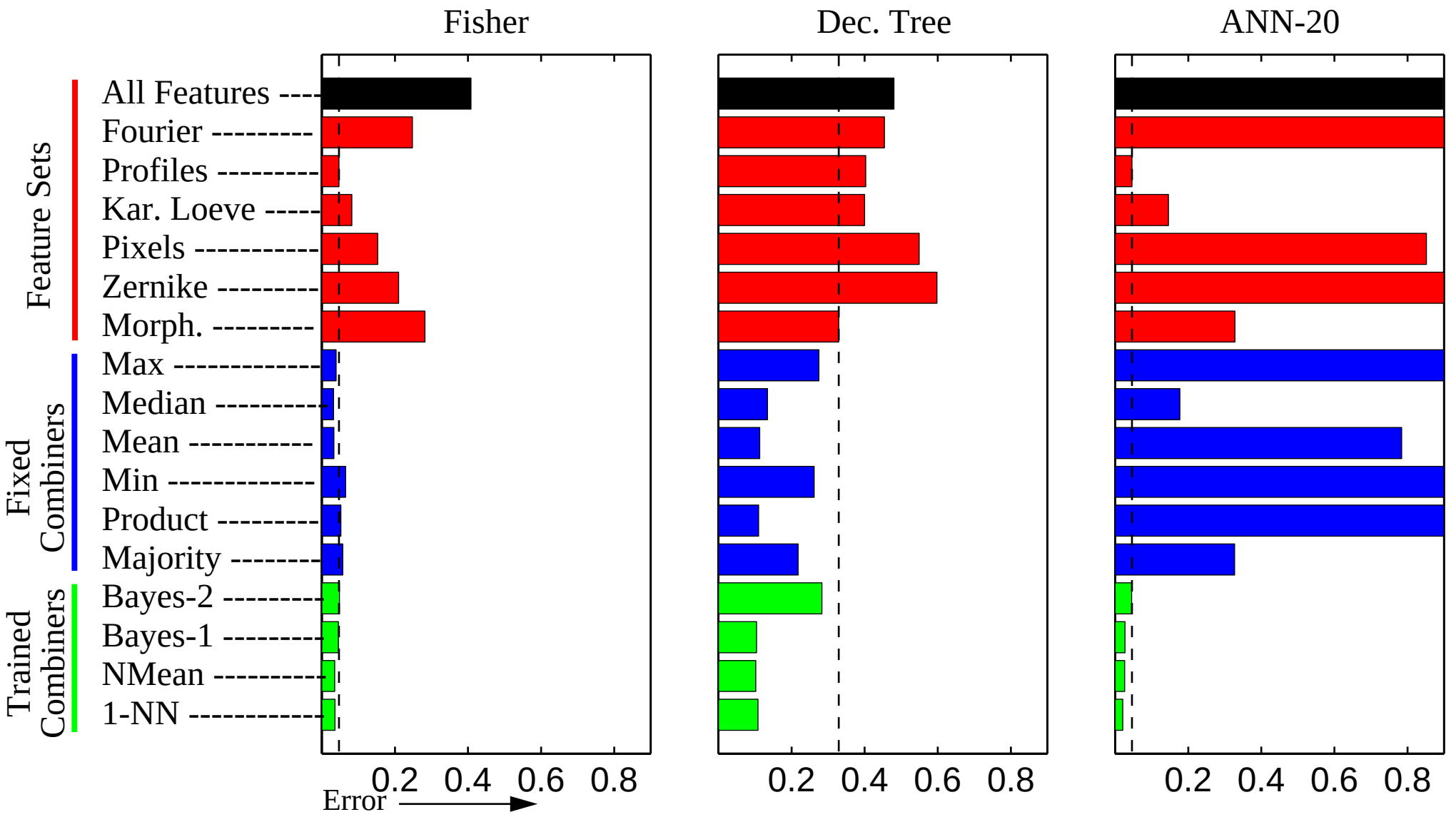
Combining the Same Classifier over Different Feature Sets (1)



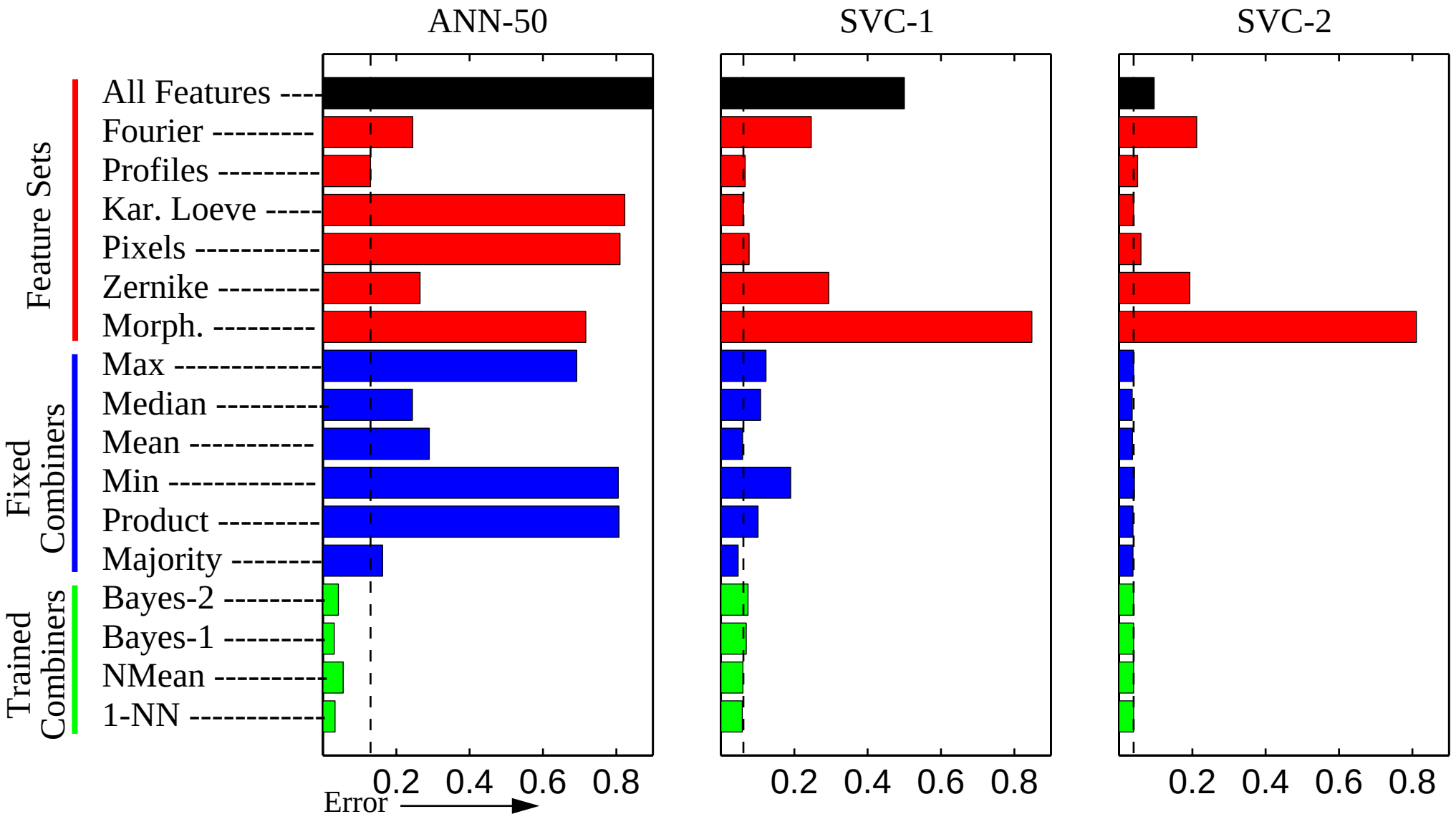
Combining the Same Classifier over Different Feature Sets (2)



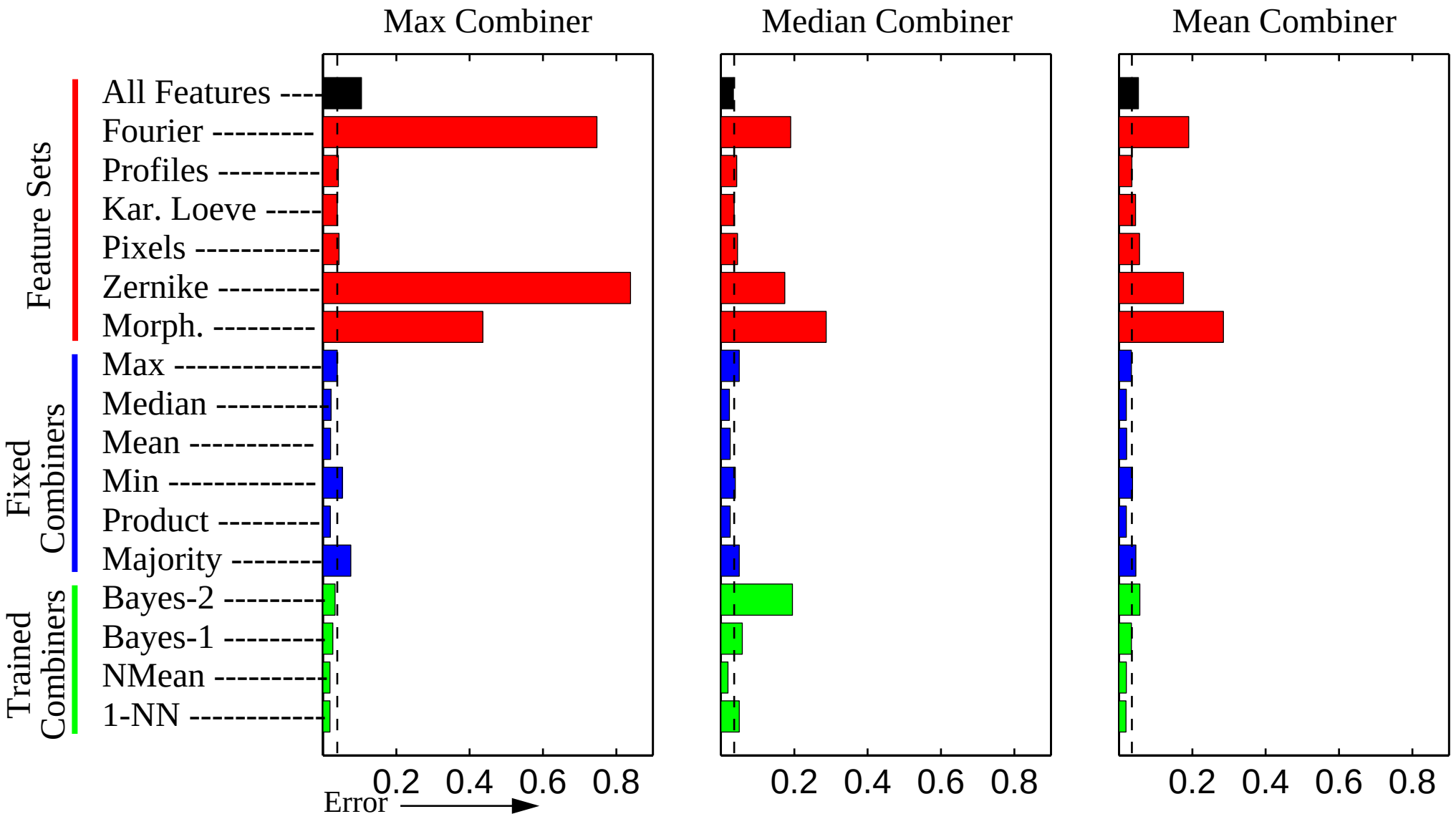
Combining the Same Classifier over Different Feature Sets (3)



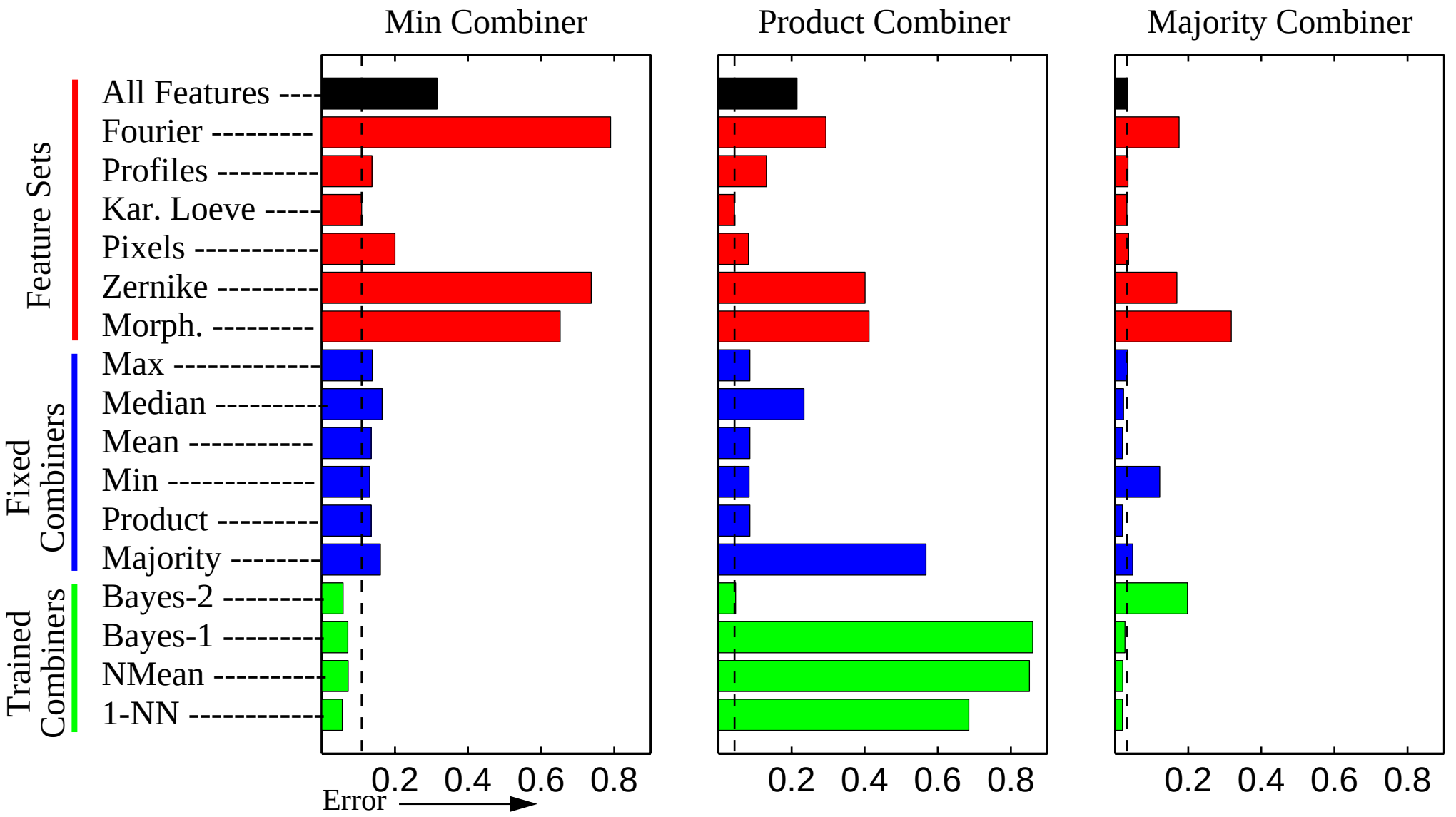
Combining the Same Classifier over Different Feature Sets (4)



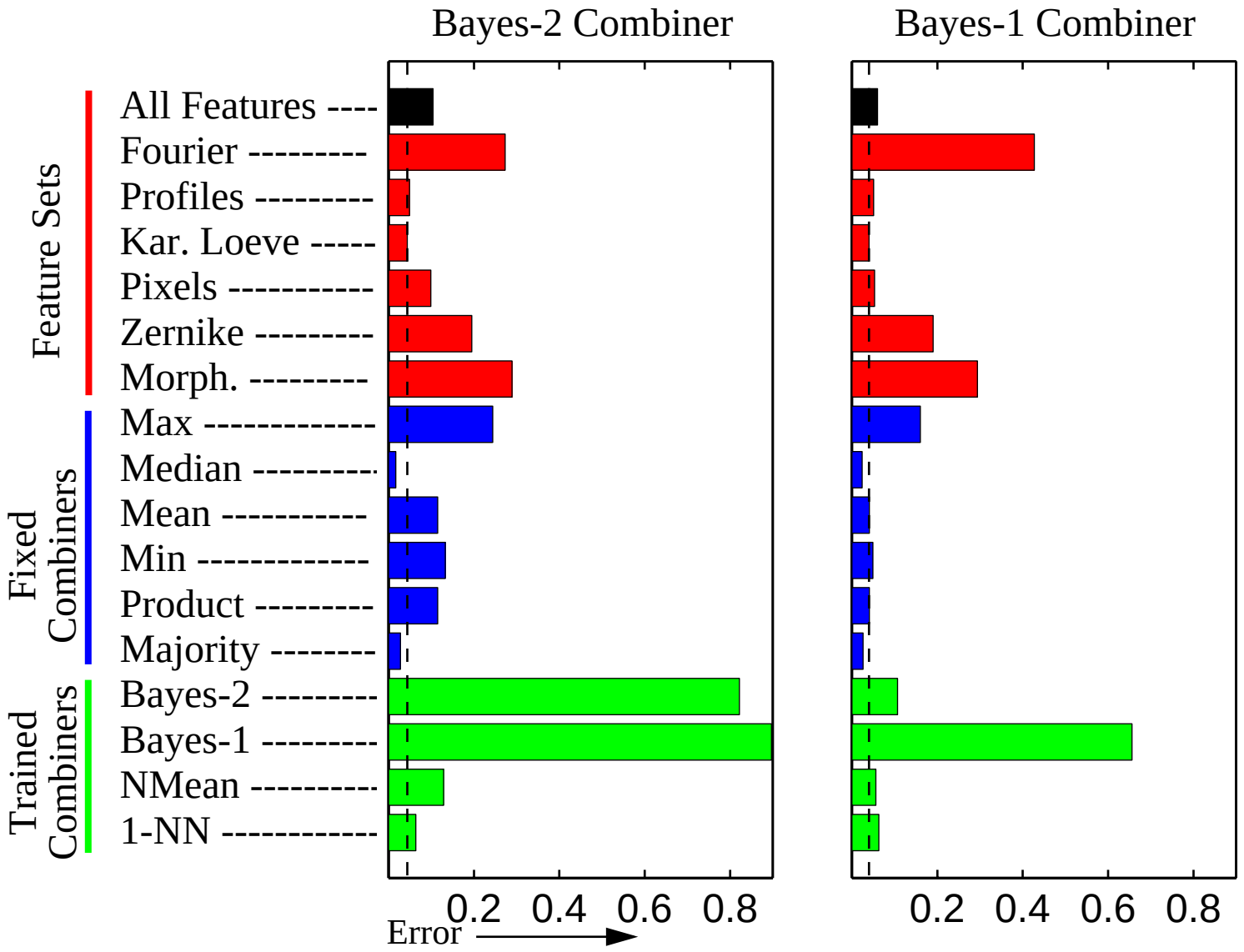
Combining Combined Classifiers over Different Feature Sets (1)



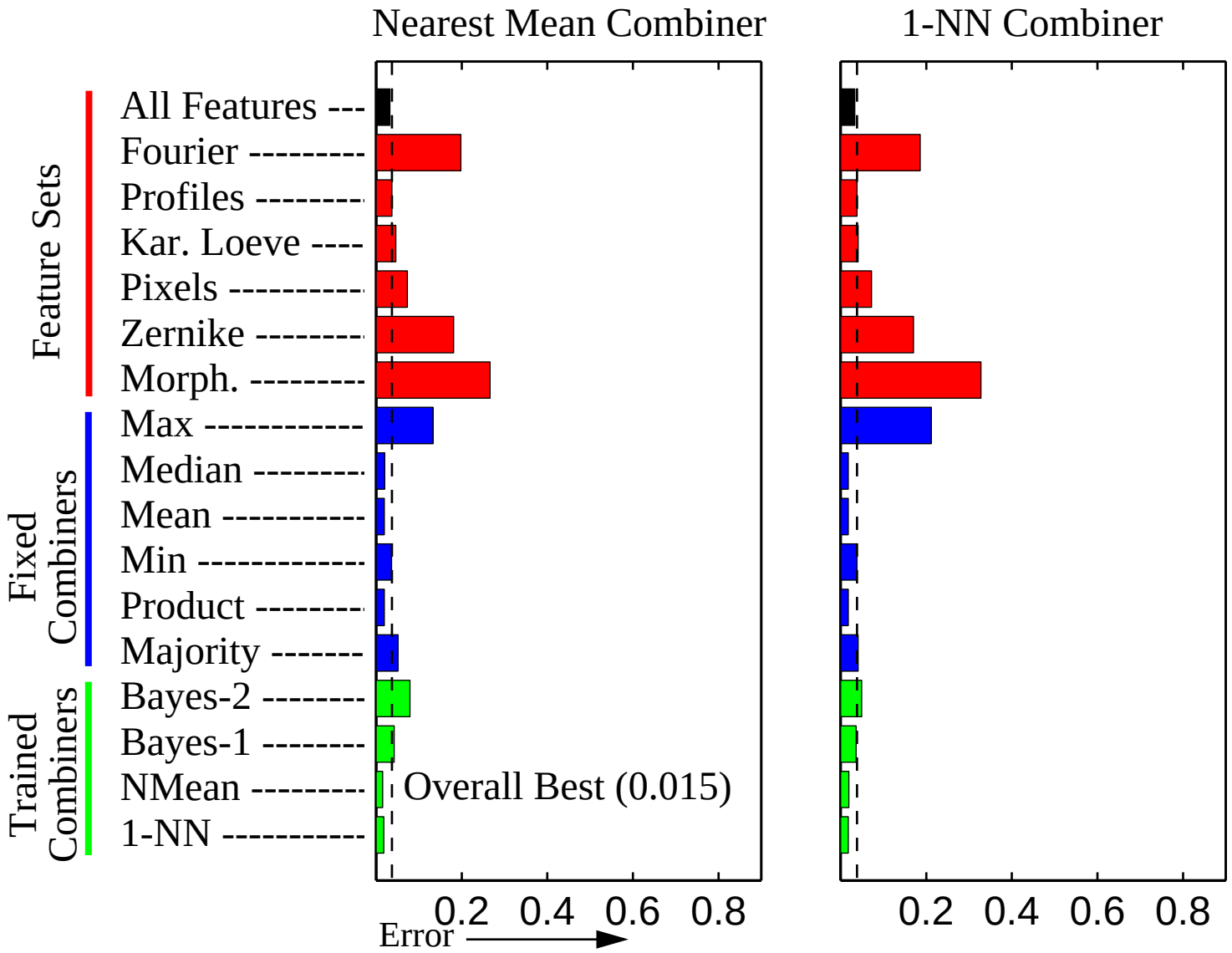
Combining Combined Classifiers over Different Feature Sets (2)



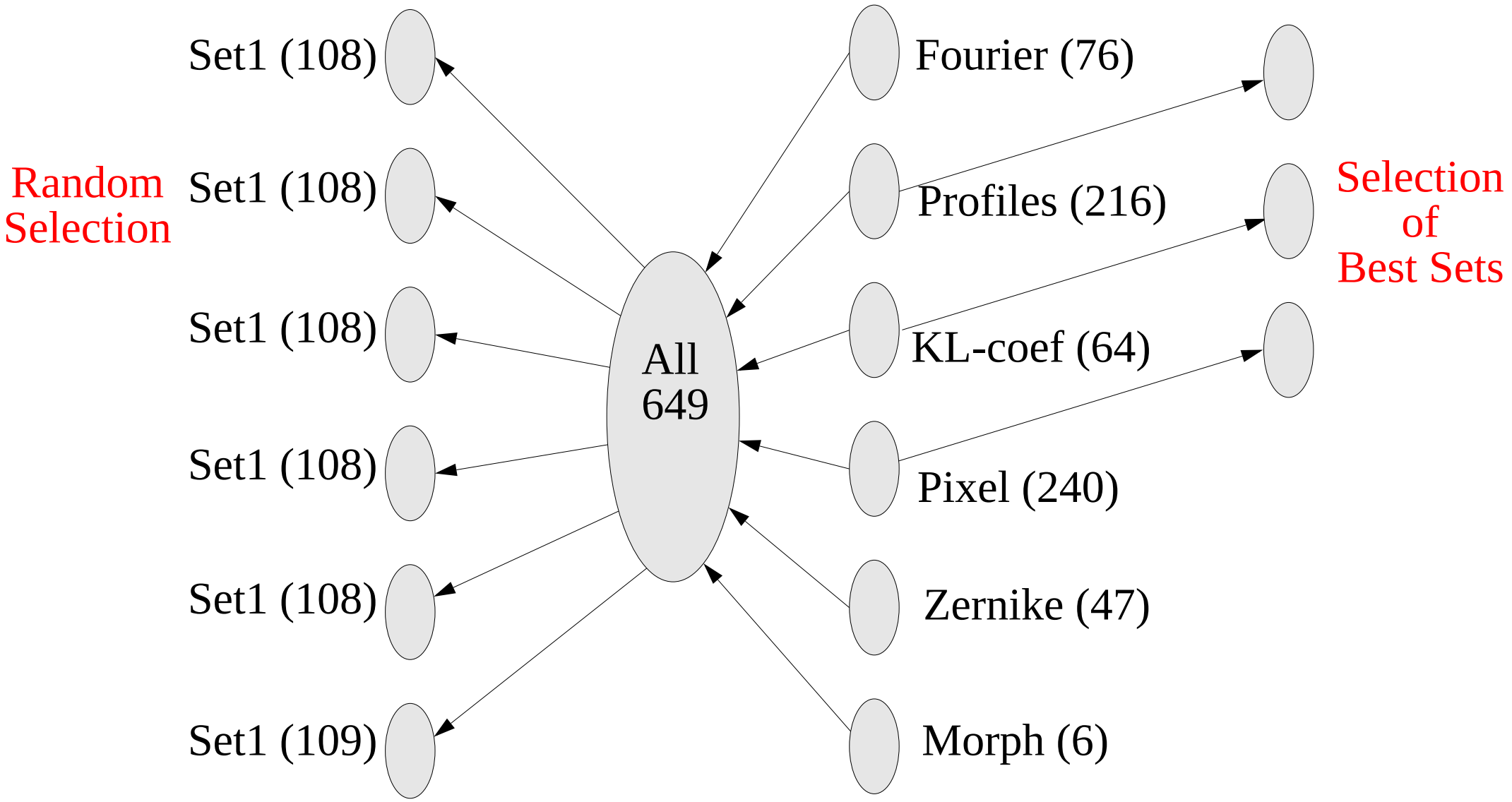
Combining Combined Classifiers over Different Feature Sets (3)



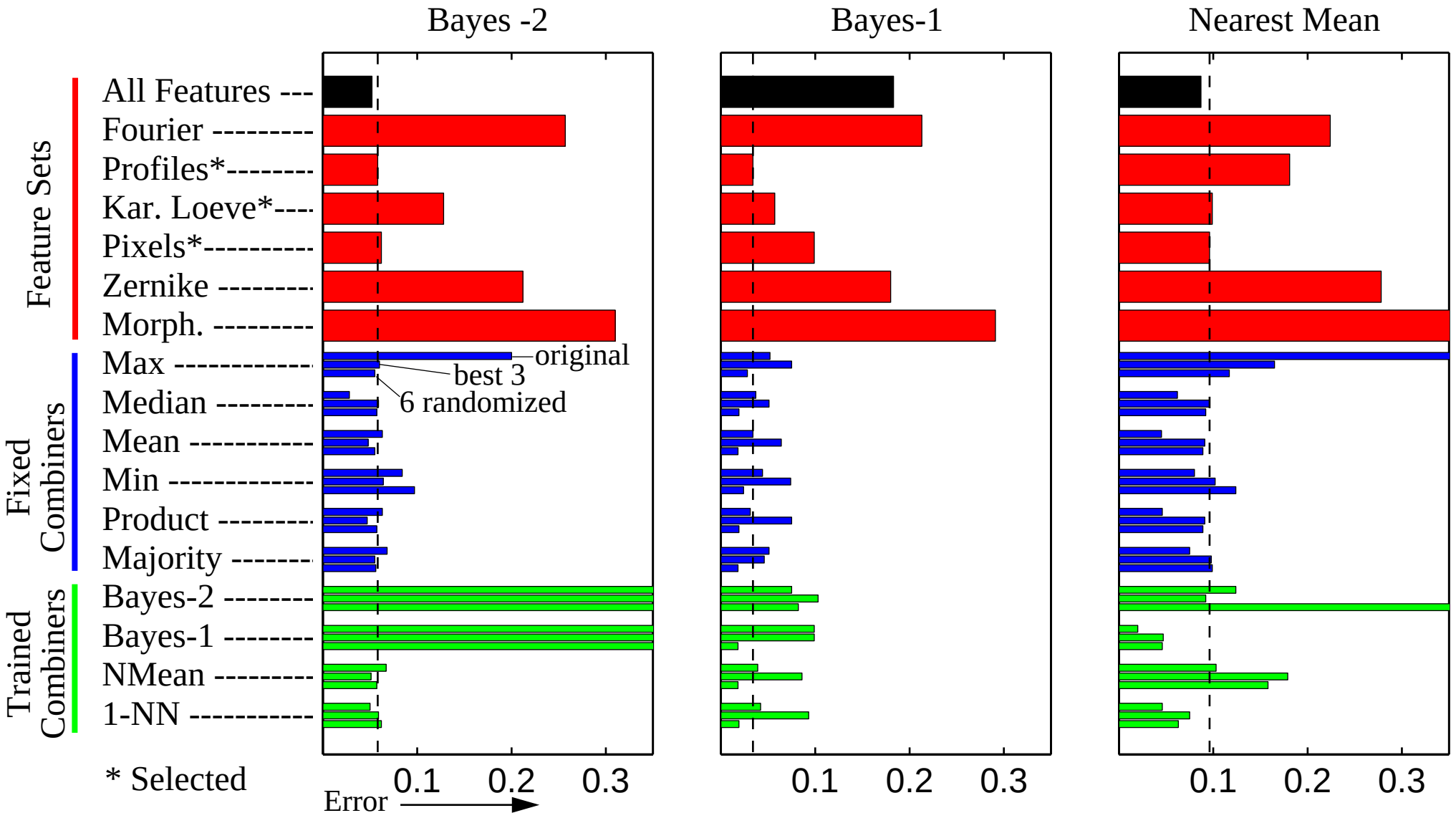
Combining Combined Classifiers over Different Feature Sets (4)



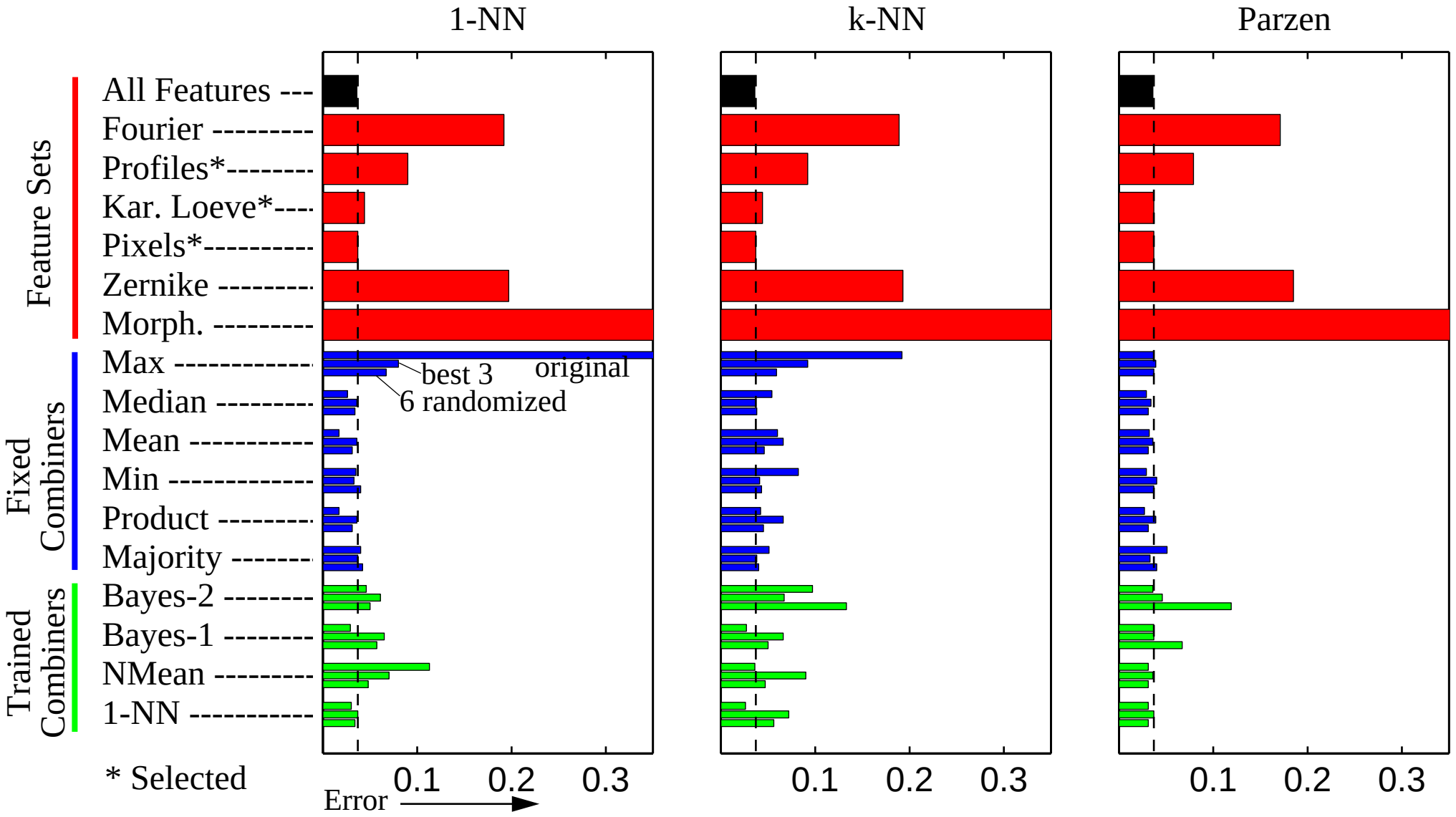
Reorganizing the Feature Sets



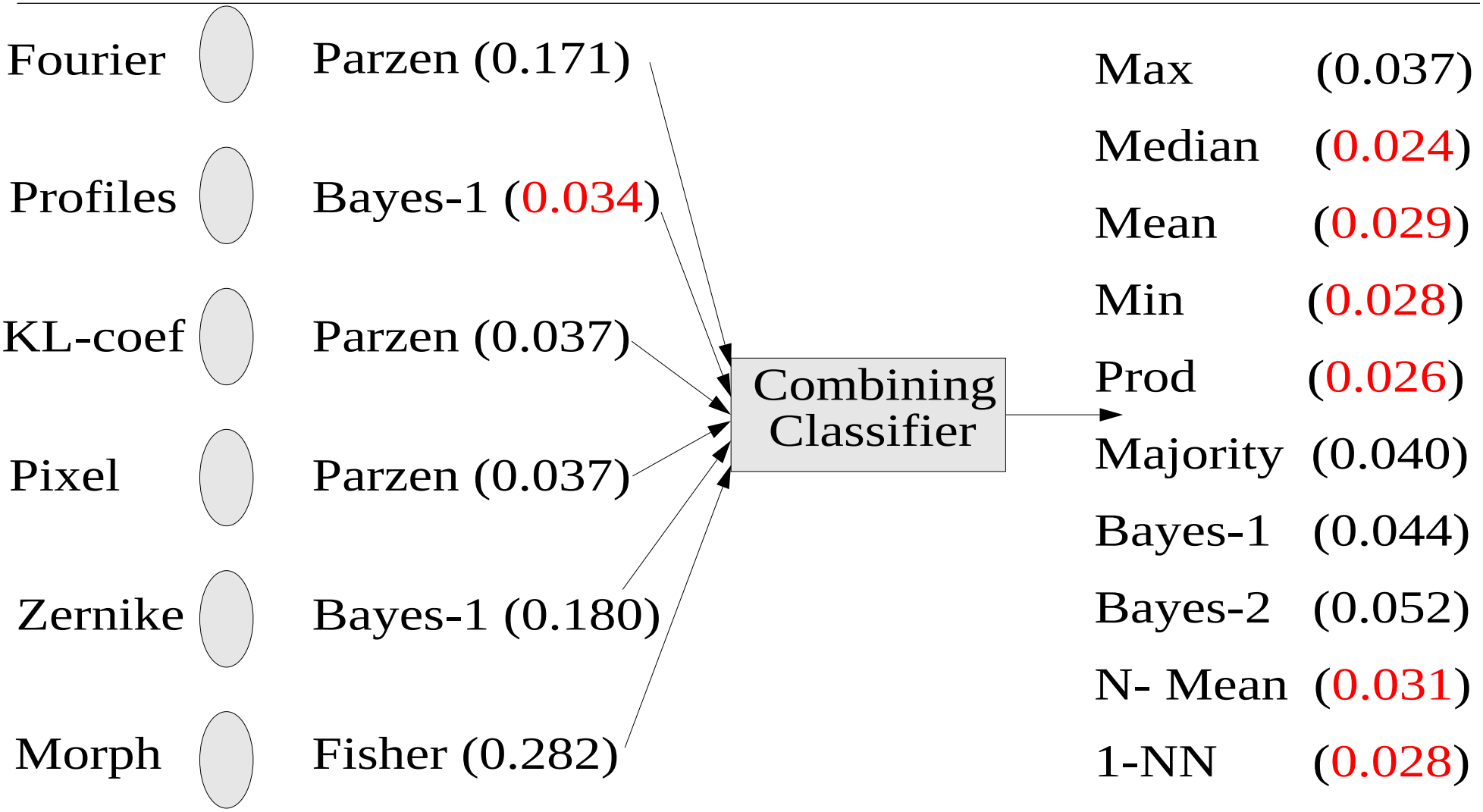
Combining the Same Classifier over Different Feature Sets: Original, Best 3 Selected and 6 Randomized



Combining the Same Classifier over Different Feature Sets: Original, Best 3 Selected and 6 Randomized



Combining Best Classifiers for Each Feature Set



(Compare Overall Best: 0.015)

The Best Results per Feature Set

Features	Best Classifier	Error	Best Combiner	Error
All	1-NN k-NN Parzen	0.036	NMean	0.032
Fourier	Parzen	0.171	Majority	0.175
Profiles	Bayes-1	0.034	Majority	0.035
KL-coef	Parzen	0.037	Majority	0.032
Pixels	1-NN k-NN Parzen	0.037	Majority	0.037
Zernike	Bayes-1	0.180	Majority	0.169
Morph	Fisher	0.282	NMean	0.266
Random	Bayes-1	0.025	Mean, Majority, Bayes-1, NMean	0.018

The Best Results per Classifier

Classifier	Best Feature Set	Error	Best Combiner	Error
Bayes-2	Profiles	0.058	Median	0.028
Bayes-1	Profiles	0.034	Product	0.031
NMean	Pixels	0.096	Bayes-1	0.020
1-NN	Pixels	0.037	NMean, Product	0.017
k-NN	Pixels	0.037	1-NN	0.026
Parzen	Pixels, KL-coef	0.037	Product	0.027
Fisher	Profiles	0.047	Median	0.032
Dec. Tree	Morph	0.329	NMean	0.102
ANN-20	Profiles	0.046	1-NN	0.021
ANN-50	Profiles	0.130	Bayes-1	0.031
SVC-1	KL-coef	0.061	Majority	0.047
SVC-2	KL-coef	0.040	Median	0.036

The Best Combiner Combiners

Combiner	Best Feature Set	Error	Best Combiner Combiner	Error
Maximum	KL-coef	0.039	NMean, 1-NN	0.019
Median	KL-coef	0.036	NMean	0.019
Mean	Profiles	0.035	1-NN	0.019
Minimum	KL-coef	0.109	1-NN	0.056
Product	KL-coef	0.044	Bayes-1	0.047
Majority	KL-coef	0.032	1-NN, Product, Mean	0.020
Bayes-2	KL-coef	0.044	Median	0.017
Bayes-1	KL-coef	0.040	Median	0.024
NMean	Profiles	0.037	NMean	0.015
1-NN	Profiles	0.038	Median, Mean, Product, 1-NN	0.018

Conclusions for this Experiment

Random feature subsets are better than a systematic subset for a single classifier.

This holds for the individual classifiers as well as for combining over the feature subsets.

Combining the best individual classifier per feature set is not better than using the same classifier.

Combining different classifiers over the same feature set is sometimes useful.

Combining combining classifiers are useful.

Weak classifiers combine well.

Simple good solution: Use Nearest Mean over Bayes-1, or Bayes-1 over Nearest Mean