ICPR 2004

The characterization of classification problems by classifier disagreements

Robert P.W.Duin, Elżbieta Pękalska, David Tax

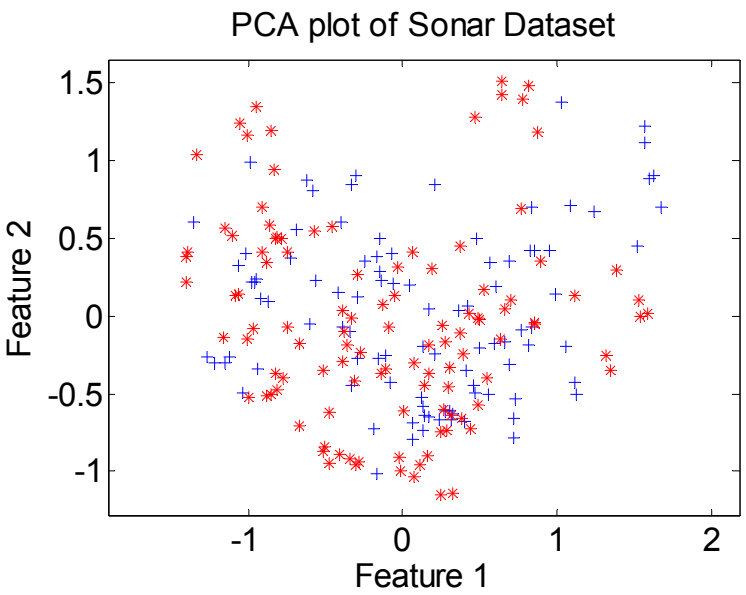
Faculty of Electrical Engineering, Mathematics and Computer Science

Delft University of Technology, The Netherlands

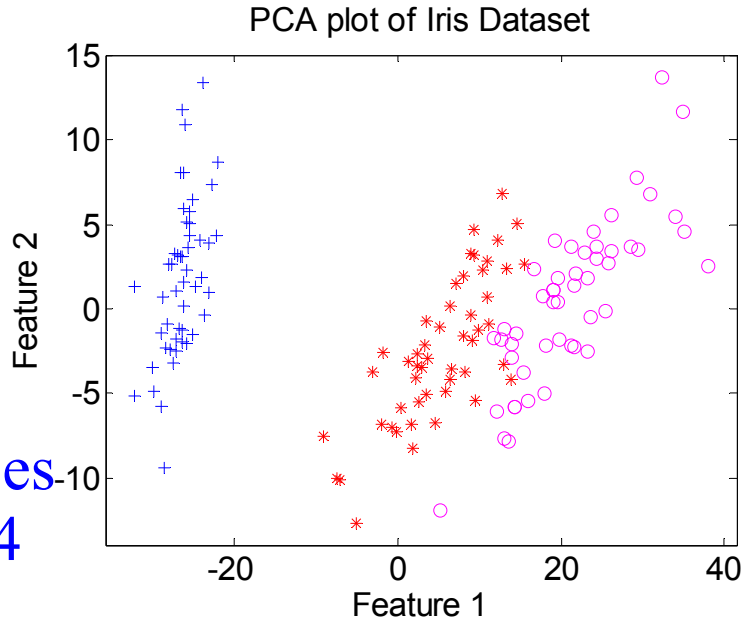
Cambridge, August 2004

P.O. Box 5031, 2600GA Delft, The Netherlands.
Phone: +(31) 15 2786143, FAX: +(31) 15 2781843,
E-mail: r.p.w.duin@ewi.tudelft.nl

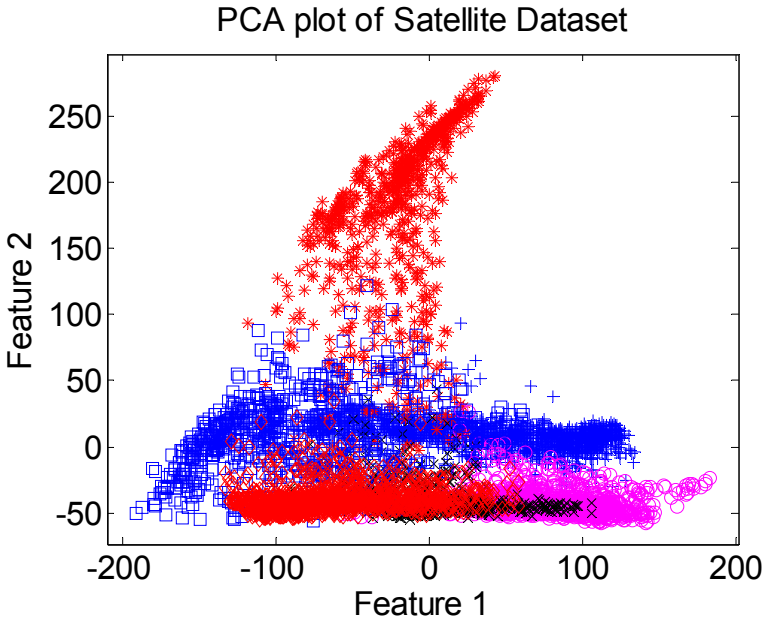
How to characterize a classification problem?



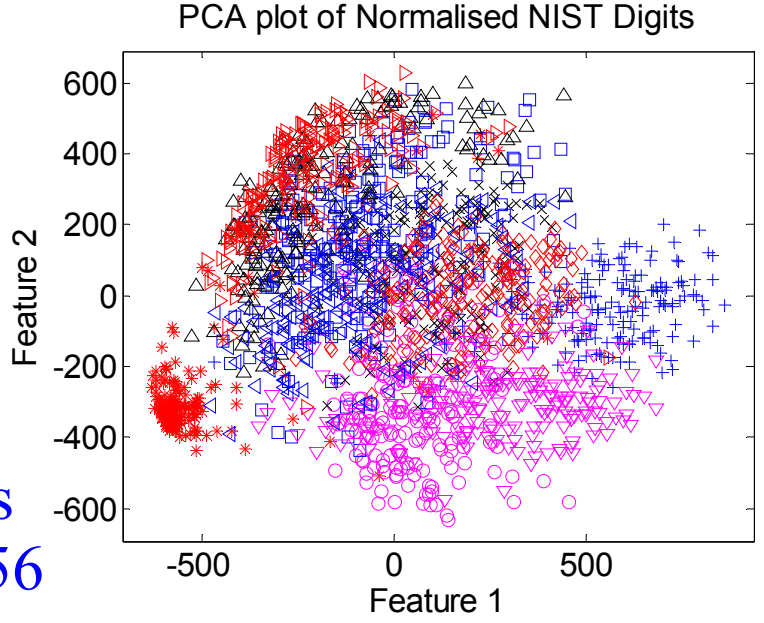
2 classes
208 x 60



3 classes
150 x 4



6 classes
6435 x 36



10 classes
2000 x 256

How to characterize a classification problem?

How difficult is it?

What performance may be expected?

How many samples do we need?

Do we suffer from the dimensionality problem?

Is feature reduction needed?

What classifiers are to be preferred?

The Answer

A problem is characterized by its **Behaviour** w.r.t. the available **Tools** .

Behaviour: similarity to known problems

Tools: classifiers

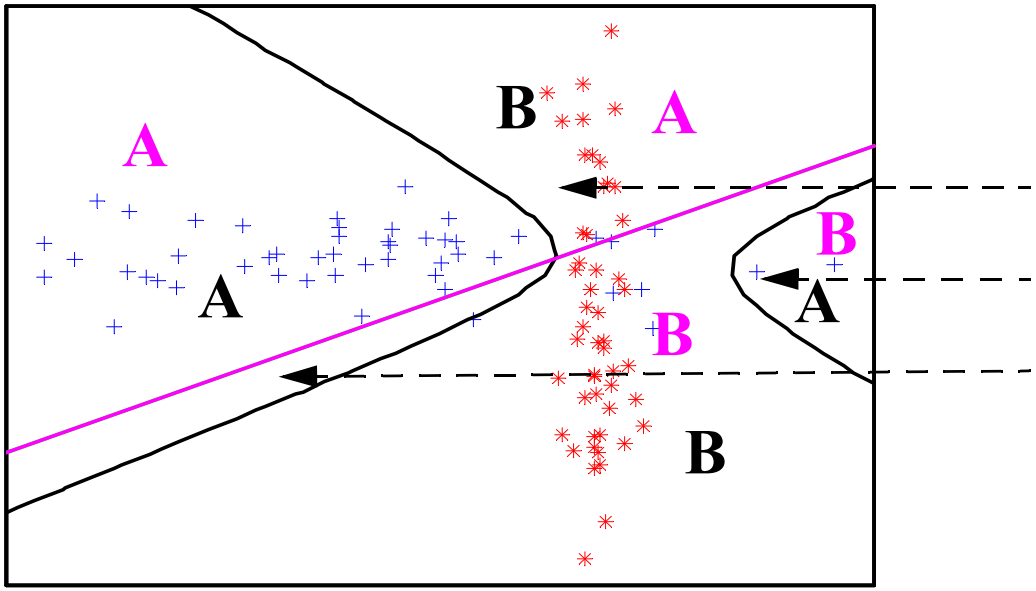
--> Problem similarity based on classification results

on a set of standard classifiers.

Set of Classifiers

- NMC: Nearest Mean Classifier
- Fisher: Fisher's Linear Discriminant
- UNormalBC: Bayes Classifier assuming uncorrelated normal densities
- NormalBC: Bayes Classifiers assuming arbitrary normal densities
- NaiveBC: Naive Bayes Classifier based on 10-bin histograms per feature
- ParzenC: Parzen classifier, LOO optimization of h
- 1-NN: 1-Nearest Neighbor Rule
- k-NN: k-Nearest Neighbor Rule, LOO optimization of k
- LogC: Logistic Classifier
- SVC-1: Support Vector Classifier, linear kernel, C = 1
- SVC-2: Support Vector Classifier, quadratic kernel, C = 1
- LM-NeurC: Neural Net, 5 neurons, trained by Levenberg-Marquardt
- CART: CART Decision Tree, maximizing purity, using early pruning

Classifier Disagreement



$D(C_1,C_2)$: Fraction of differently labeled test samples (independent of true labels)

Classifier 1

Classifier 2

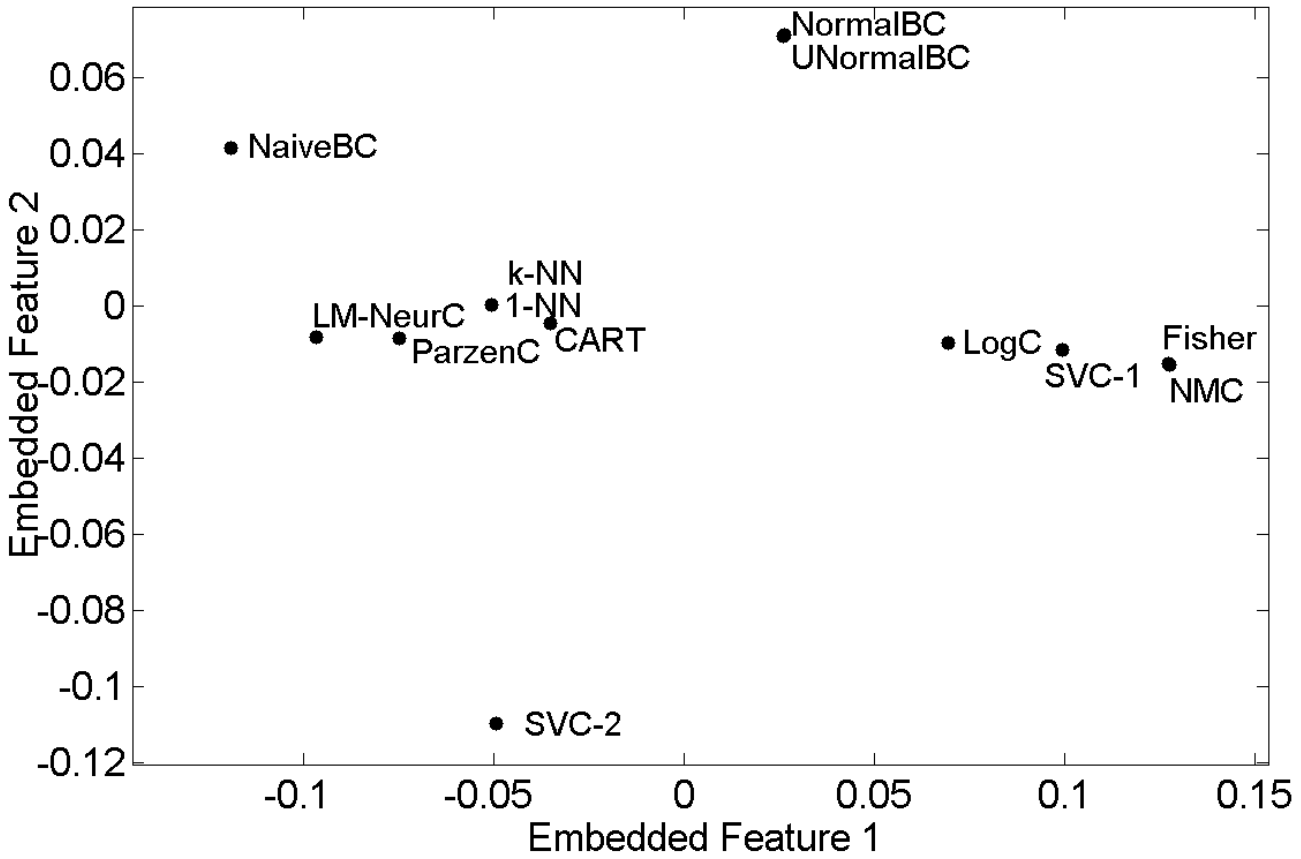
Classifier Disagreement Matrix

$D(C_i,C_j)$: Prob{classifier$_i$(x) ≠ classifier$_j$(x)}

Classifier Disagreement * 100

	NMC	Fisher	UNormalBC	NormalBC	NaiveBC	ParzenC	1-NN	k-NN	LogC	SVC-1	SVC-2	Lm-NeurC	CART
NMC	0	18	18	18	18	17	14	16	17	21	13	22	19
Fisher	18	0	18	18	16	15	18	16	1	3	19	14	15
UNormalBC	18	18	0	0	4	3	6	8	17	21	7	6	5
NormalBC	18	18	0	0	4	3	6	8	17	21	7	6	5
NaiveBC	18	16	4	4	0	3	6	4	15	17	7	10	5
ParzenC	17	15	3	3	3	0	3	5	14	18	8	9	4
1-NN	14	18	6	6	6	3	0	8	17	21	11	12	7
k-NN	16	16	8	8	4	5	8	0	15	17	5	14	7
LogC	17	1	17	17	15	14	17	15	0	4	18	13	14
SVC-1	21	3	21	21	17	18	21	17	4	0	22	17	16
SVC-2	13	19	7	7	7	8	11	5	18	22	0	13	10
Lm-NeurC	22	14	6	6	10	9	12	14	13	17	13	0	9
CART	27	14	14	10	6	8	14	10	10	7	5	13	0

The Classifier Projection Space (CSP)

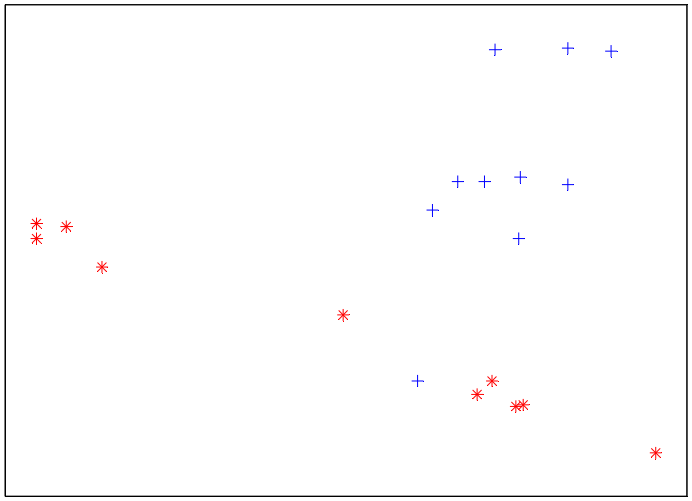


The CSP is a linear approximate embedding of the

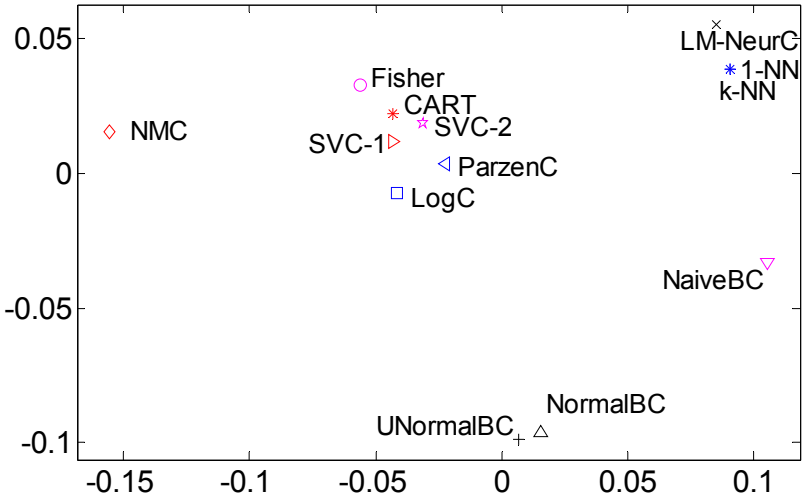
Classifier Disagreement Matrix (MCS 2002)

CSP Examples (1,2)

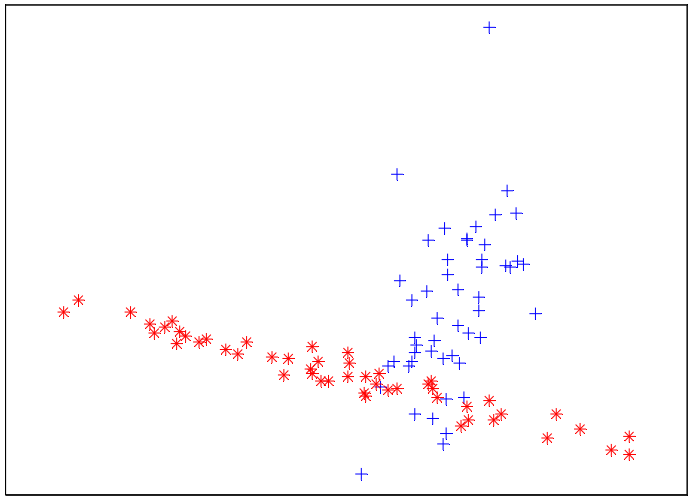
Highleyman-20, 20 2



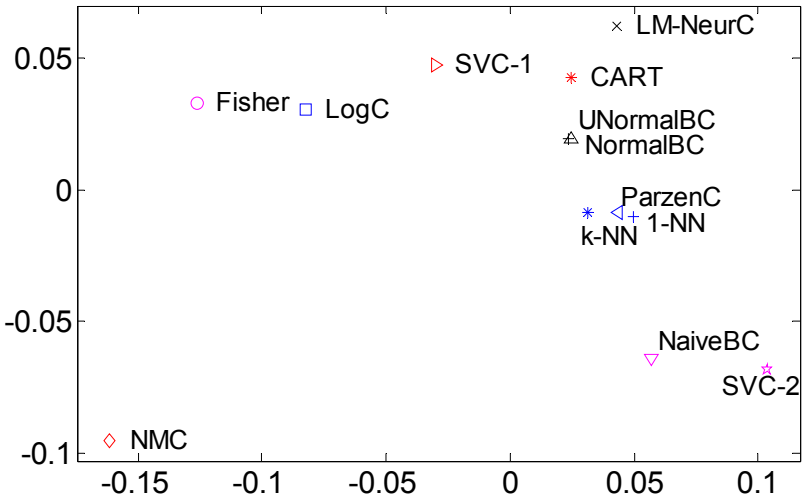
Highleyman-20



Highleyman-100, 100 2

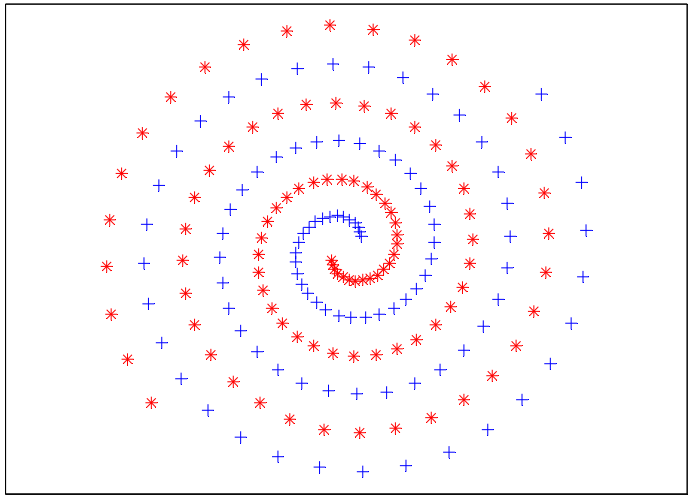


Highleyman-100

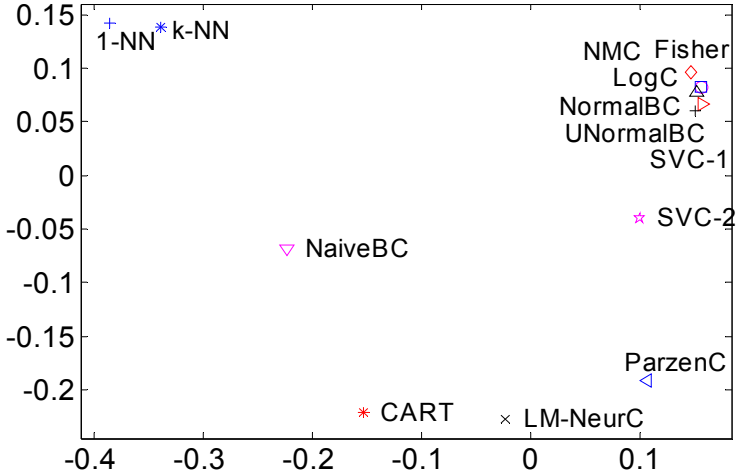


CSP Examples (9,10)

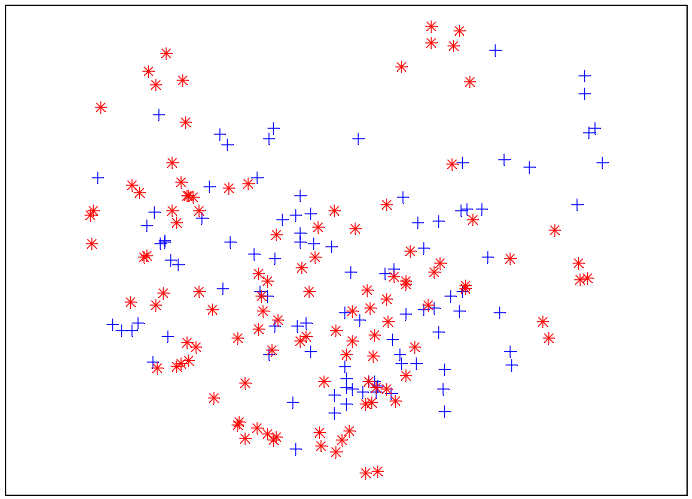
Spirals, 194 2



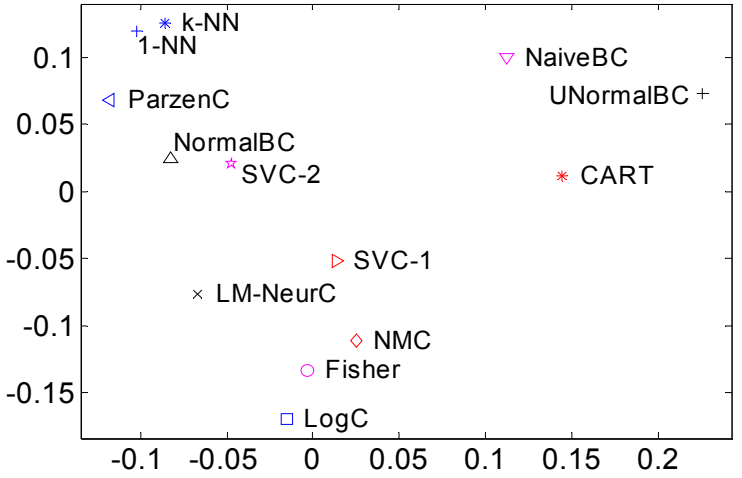
Spirals



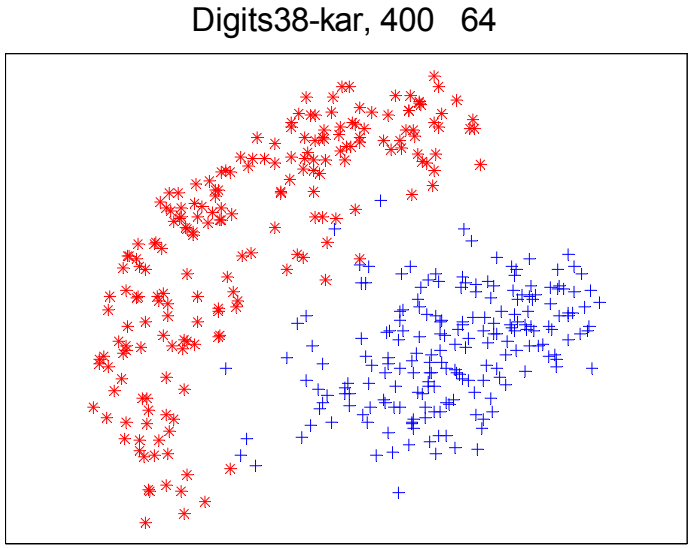
Sonar, 208 60

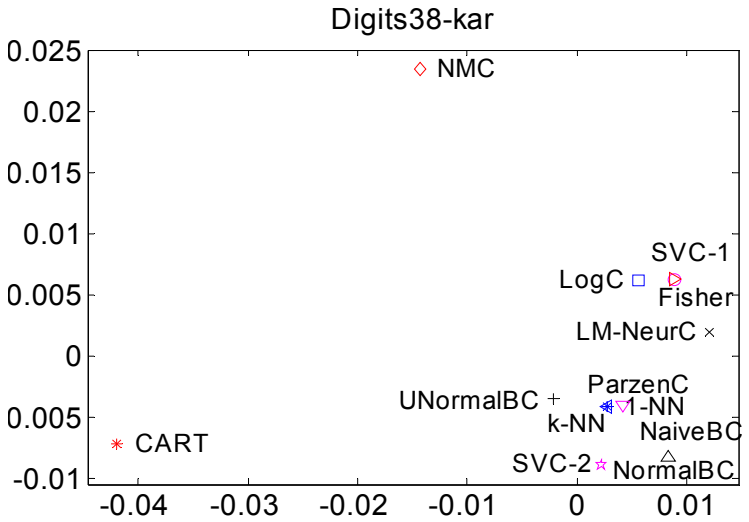


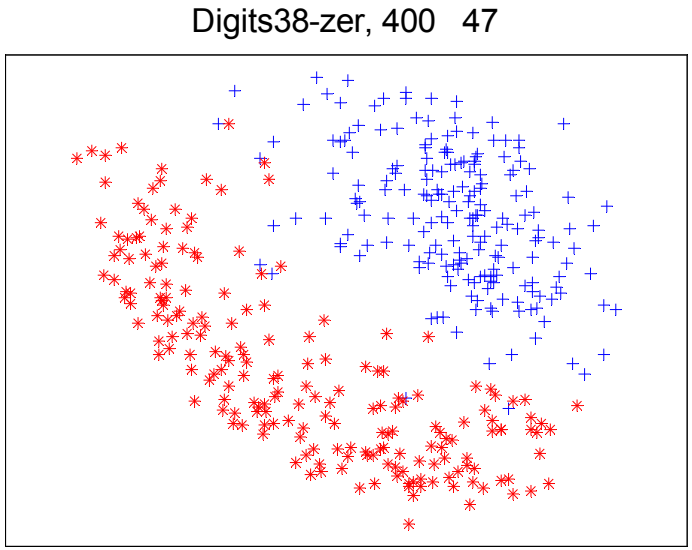
Sonar

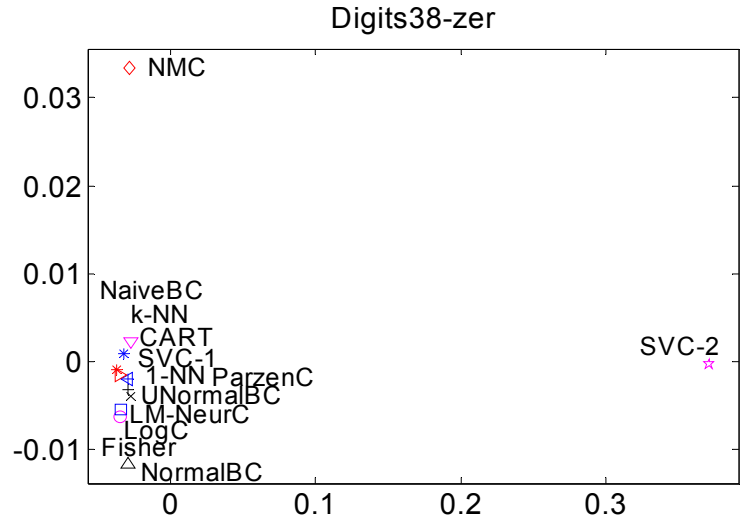


CSP Examples (17,18)









Problem Dissimilarities

$$D(\text{Problem\_p}, \text{Problem\_q}) = \sum_{(i,j) \in \text{Classifiers}} \left| \text{ClassfDisagreement}_p(i,j) - \text{ClassfDisagreement}_q(i,j) \right|$$

Two problems are as different as their classifier disagreement patterns

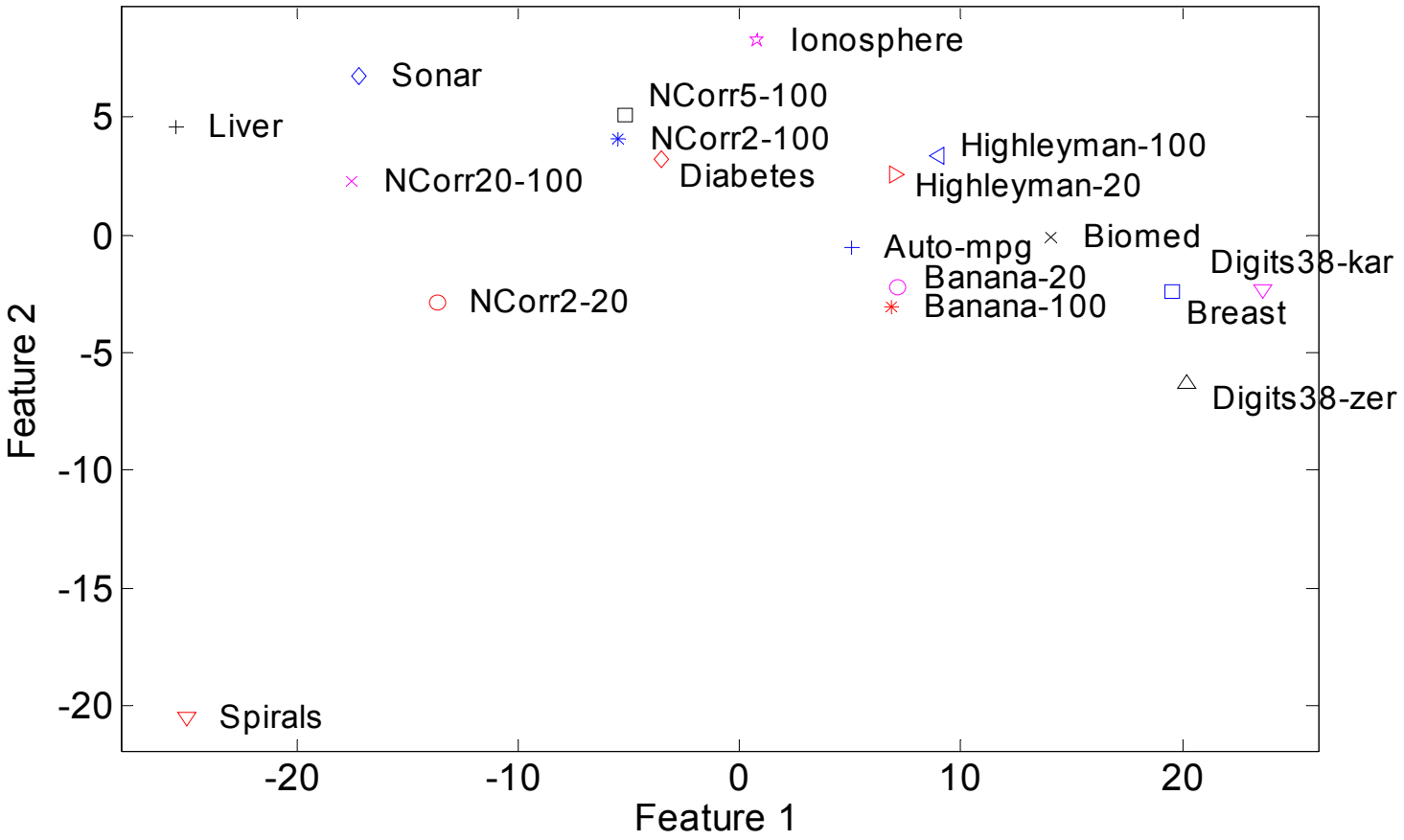
Problems

Dataset name	#features	#objects	R	B
Highleyman-20*	2	10 + 10		
Highleyman-100*	2	50 + 50		
Banana-20*	2	10 + 10		
Banana-100*	2	50 + 50	X	
NCorr2-20*	2	10 + 10		
NCorr2-100*	2	50 + 50		
NCorr5-100*	5	50 + 50		
NCorr20-100*	20	50 + 50		
Spirals*	2	97 + 97	X	X
Sonar	60	97+ 111		
Biomed	5	127 + 67		
Diabetes	8	500 + 268	X	
Auto-mpg	6	229 + 169		
Ionosphere	34	225 + 126		
Liver	6	145 + 200		X
Breast	9	444 + 239		
Digit38-kar	64	200 + 200		X
Digit38-zer	47	200 + 200	X	X

Problem Dissimilarities Matrix

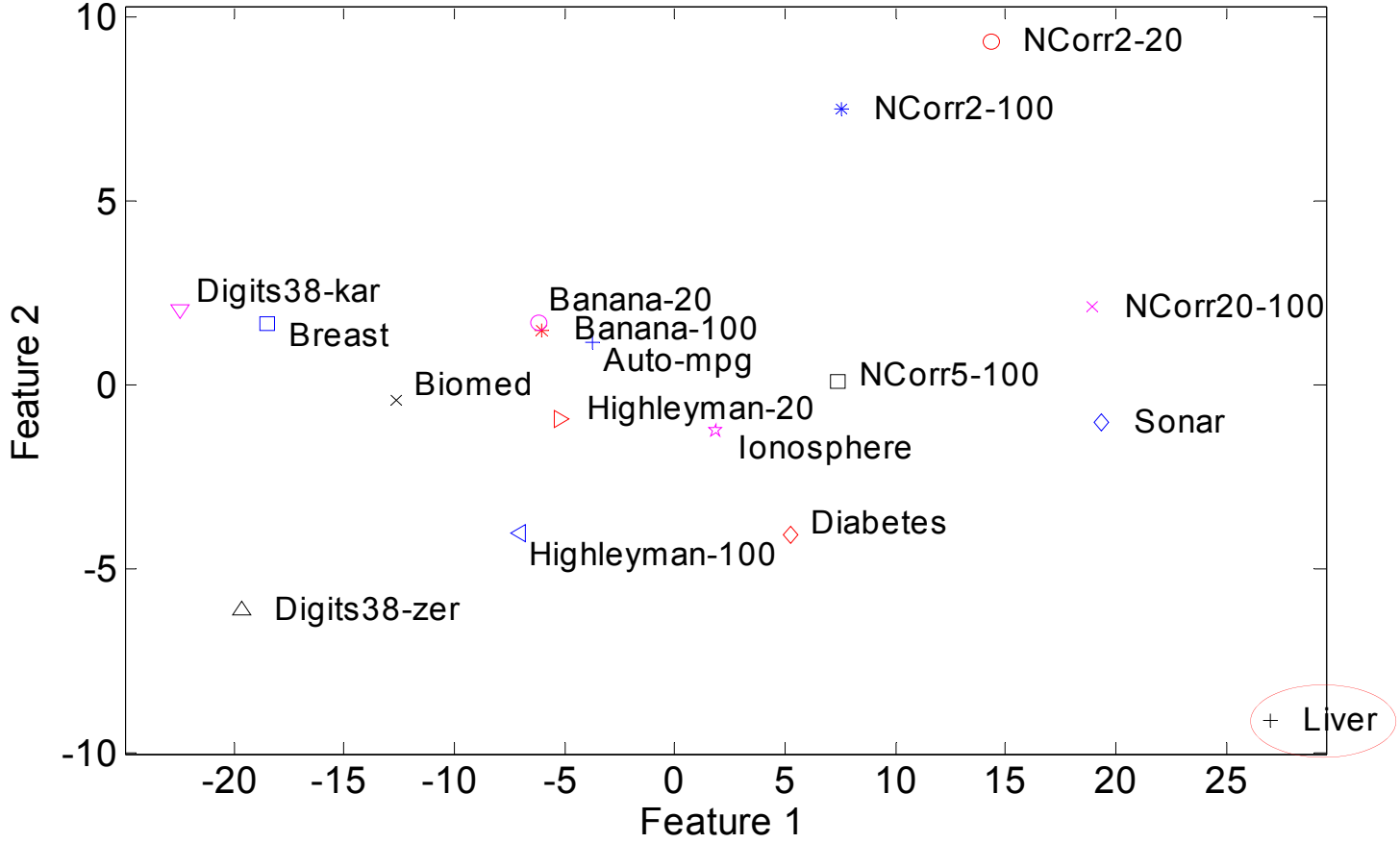
	Highleyman-20*	Highleyman-100*	Banana-20*	Banana-100*	NCorr2-20*	NCorr2-100*	NCorr5-100*	NCorr20-100*	Spirals*	Sonar	Sonar	Biomed	Diabetes	Auto-mpg	Ionosphere	Liver	Digit38-kar	Digit38-zer
Highleyman-20*	0	9	9	9	23	20	17	27	40	26	10	14	12	11	34	15	17	23
Highleyman-100*	9	0	11	11	27	22	18	27	42	27	9	16	13	13	36	13	16	19
Banana-20*	9	11	0	7	23	20	16	27	37	27	10	16	12	13	35	13	15	21
Banana-100*	9	11	7	0	22	20	16	27	37	27	10	15	11	15	35	13	16	21
NCorr2-20*	23	27	23	22	0	14	16	12	23	15	29	20	22	20	22	34	37	39
NCorr2-100*	20	22	20	20	14	0	12	15	34	17	24	15	18	15	26	26	29	33
NCorr5-100*	17	18	16	16	16	12	0	14	33	16	21	13	17	12	23	26	30	30
NCorr20-100*	27	27	27	27	12	15	14	0	25	11	33	18	24	20	18	39	42	39
Spirals*	40	42	37	37	23	34	33	25	0	29	44	33	37	39	26	48	52	49
Sonar	26	27	27	27	15	17	16	11	29	0	33	17	25	19	15	38	42	42
Biomed	10	9	10	10	29	24	21	33	44	33	0	19	10	16	41	7	10	16
Diabetes	14	16	16	15	20	15	13	18	33	17	19	0	12	13	23	24	28	30
Auto-mpg	12	13	12	11	22	18	17	24	37	25	10	12	0	16	33	15	18	23
Ionosphere	11	13	13	15	20	15	12	20	39	19	16	13	16	0	28	21	24	27
Liver	34	36	35	35	22	26	23	18	26	15	41	23	33	28	0	47	51	48
Breast	15	13	13	13	34	26	26	39	48	38	7	24	15	21	47	0	4	12
Digit38-kar	17	16	15	16	37	29	30	42	52	42	10	28	18	24	51	4	0	10
Digit38-zer	23	19	21	21	39	33	30	39	49	42	16	30	23	27	48	12	10	0

Problem Projection Space



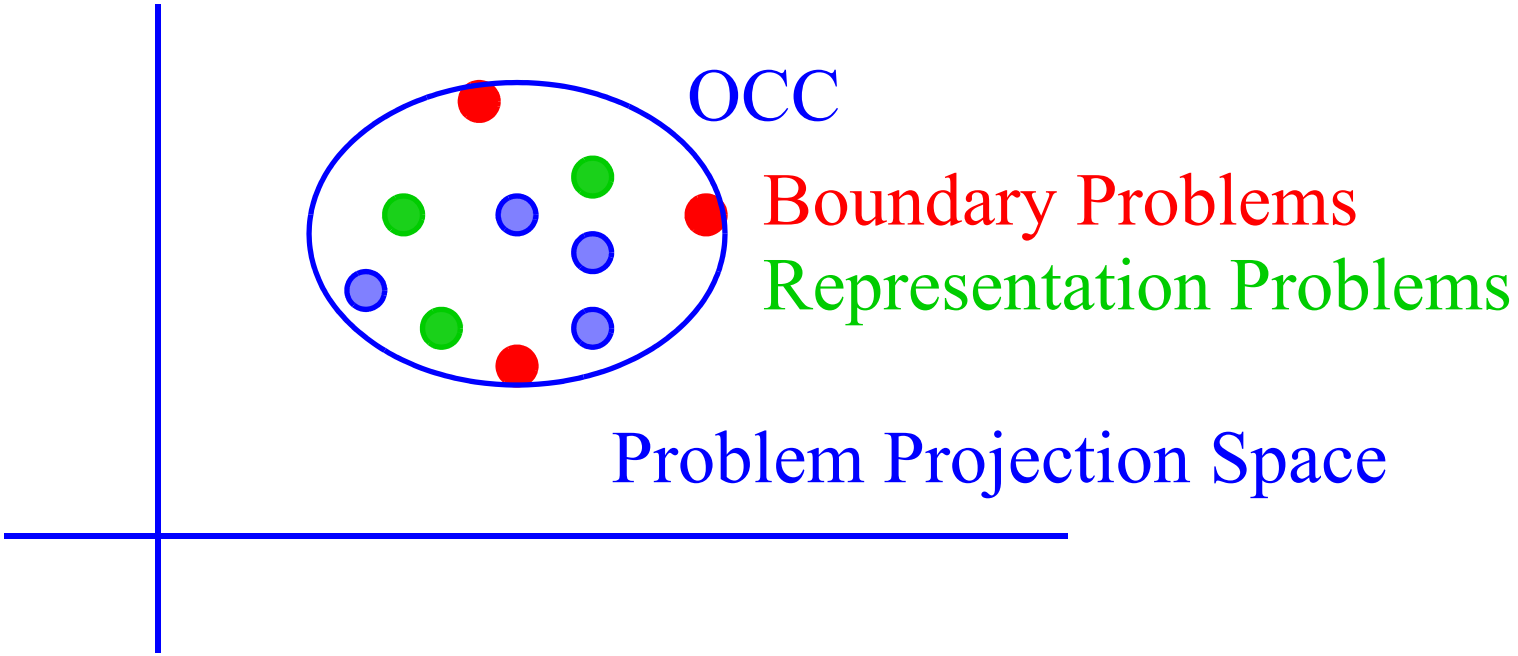
Spirals appears to be an outlier problem

Problem Projection Space (2)



Liver appears to be an outlier problem after removing Spirals

A Standard Set of Problems Described by a One-Class Classifier



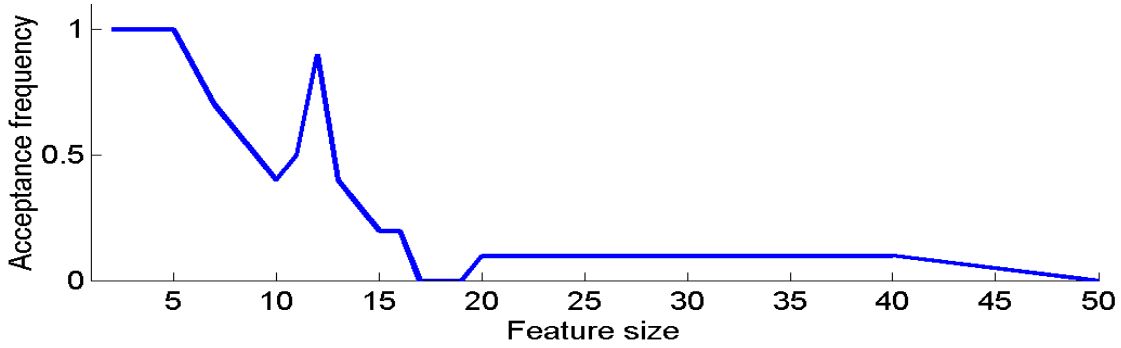
For OCC on non-Euclidean dissimilarity data, see Pekalska, Tax and Duin, NIPS 2002

Representation Problems and Boundary Problems

Dataset name	#features	#objects	Representation	Boundary
Highleyman-20*	2	10 + 10		
Highleyman-100*	2	50 + 50		
Banana-20*	2	10 + 10		
Banana-100*	2	50 + 50	X	
NCorr2-20*	2	10 + 10		
NCorr2-100*	2	50 + 50		
NCorr5-100*	5	50 + 50		
NCorr20-100*	20	50 + 50		
Spirals*	2	97 + 97	X	X
Sonar	60	97+ 111		
Biomed	5	127 + 67		
Diabetes	8	500 + 268	X	
Auto-mpg	6	229 + 169		
Ionosphere	34	225 + 126		
Liver	6	145 + 200		X
Breast	9	444 + 239		
Digit38-kar	64	200 + 200		X
Digit38-zer	47	200 + 200	X	X

Problem Modification Experiment (1)

The NCorrX-20 has a decreasing probability of being rejected as a 'standard' problem for increasing dimensionalities X



Classification of sampled datasets

problem	1	0.9	0.8	0.65	0.5	0.35
Sonar	accept	accept	accept	accept	accept	reject
Biomed	accept	accept	accept	accept	accept	accept
Liver	accept	reject	reject	reject	reject	reject

Decreasing probability of acceptance for decreasing training sample sizes

Problem Modification Experiment (2)

Classification of sub-sampled datasets

problem	1	0.9	0.8	0.65	0.5	0.35
Sonar	accept	accept	accept	accept	accept	reject
Biomed	accept	accept	accept	accept	accept	accept
Liver	accept	reject	reject	reject	reject	reject

Decreasing probability of acceptance for decreasing training sample sizes

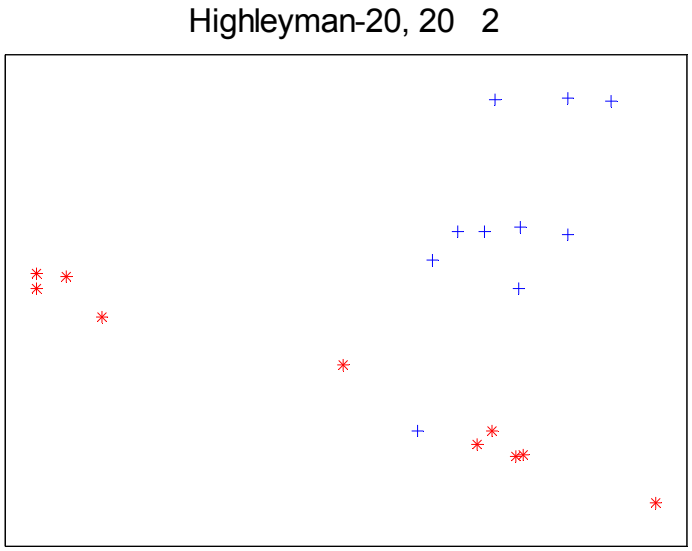
Conclusions

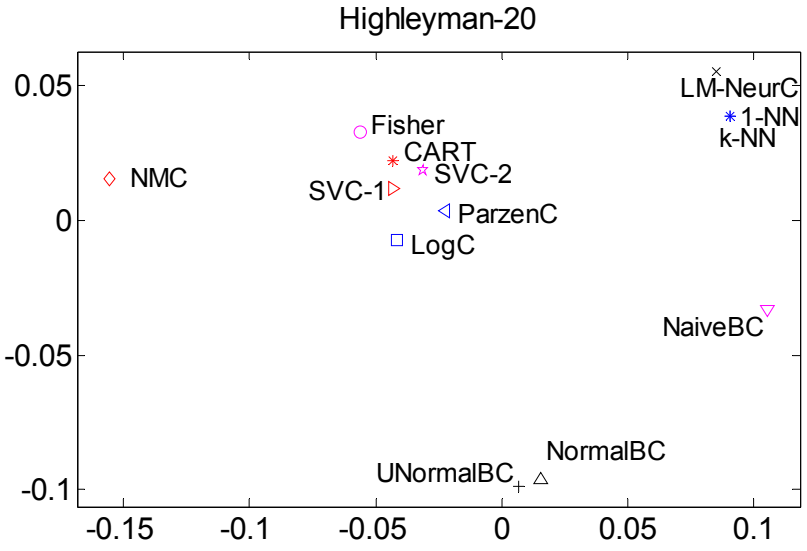
Standard classifiers are used to characterize classification problems.

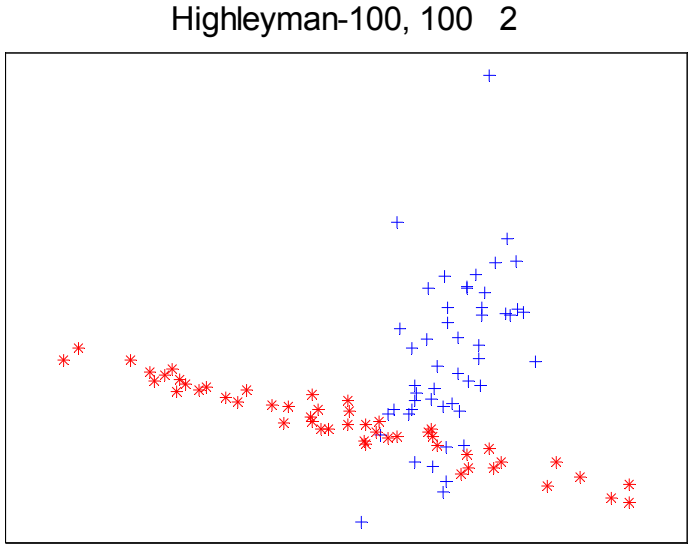
Such characterizations may be used for defining sets of 'standard' problems.

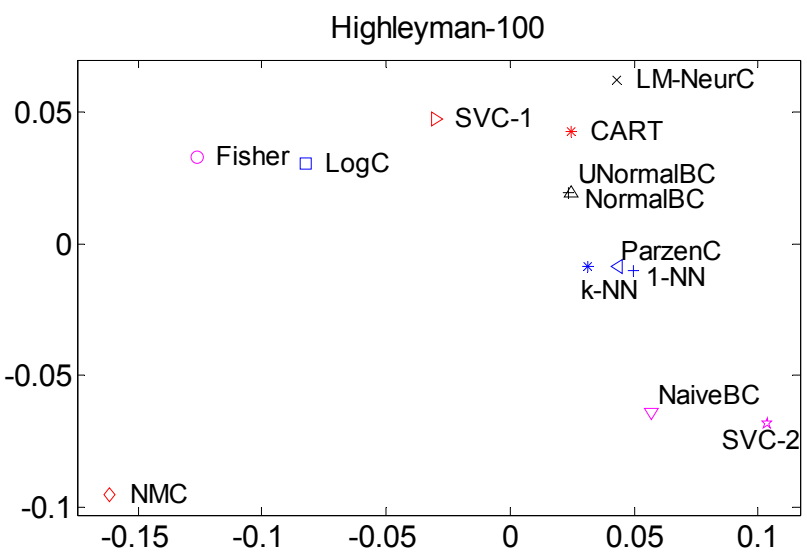
New or modified problems may be tested on their similarity to such a set.

CSP Examples (1,2)

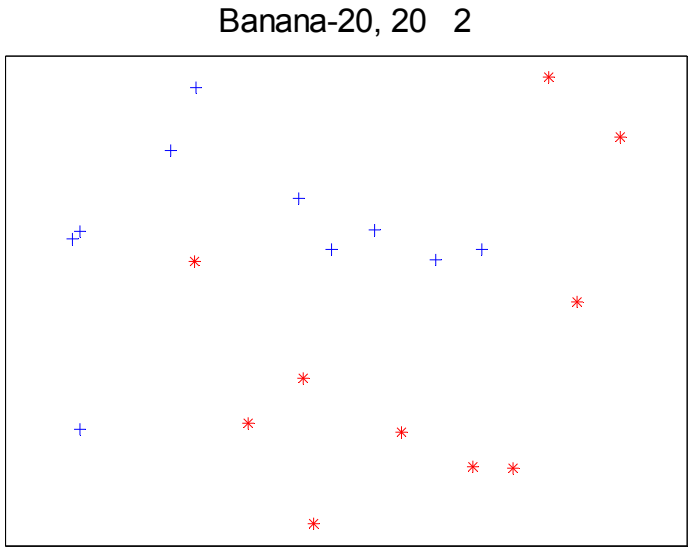


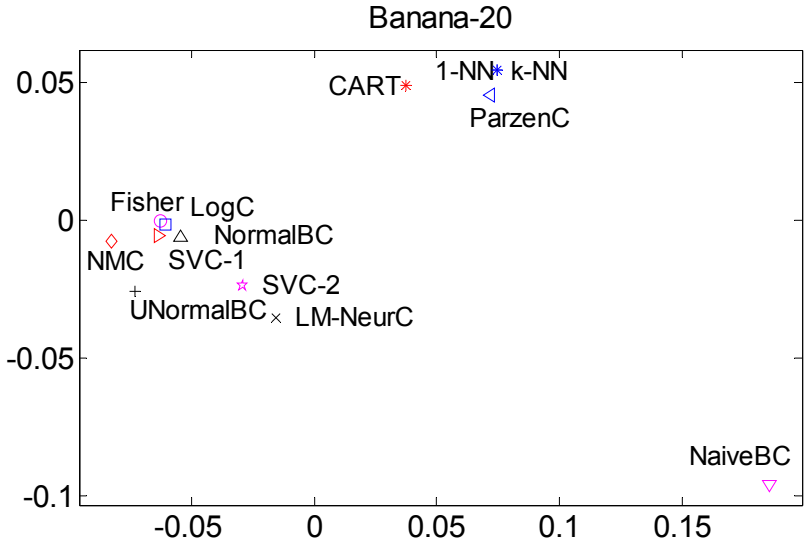


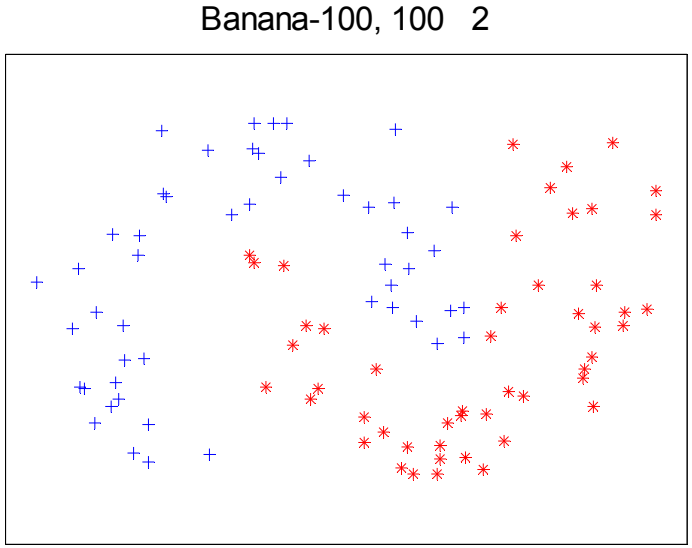


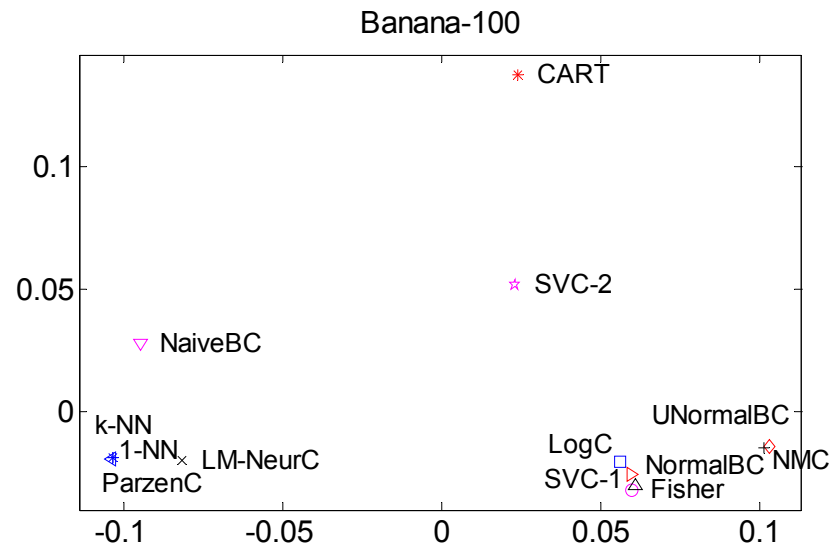


CSP Examples (3,4)

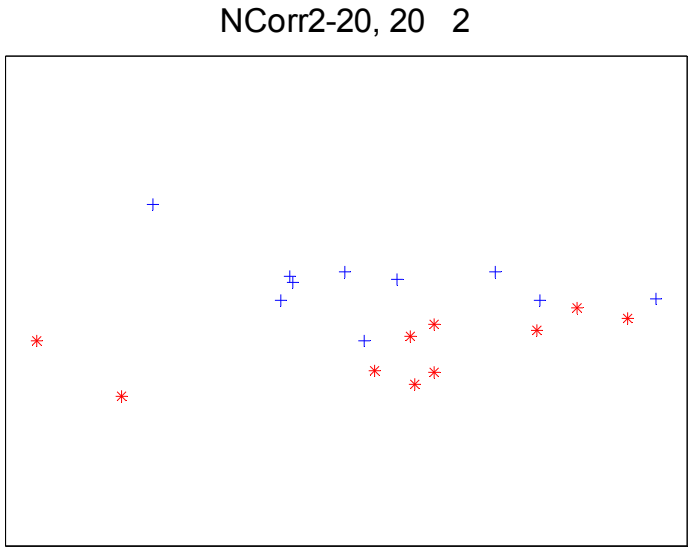


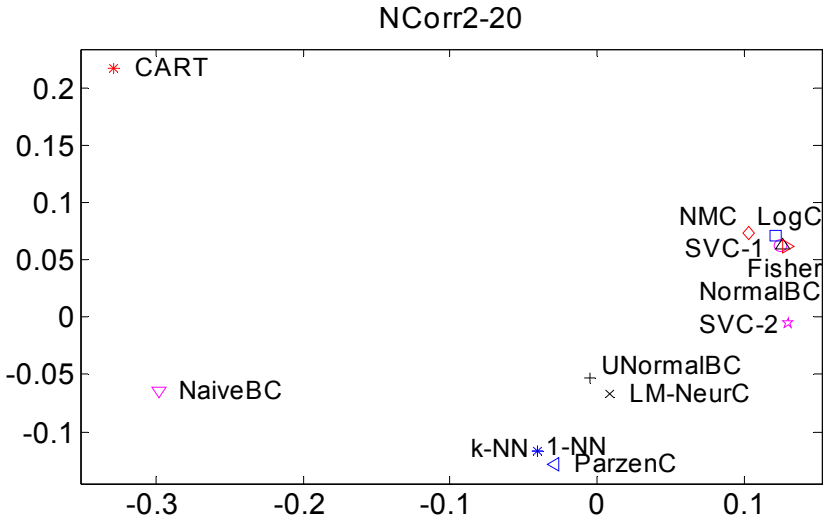


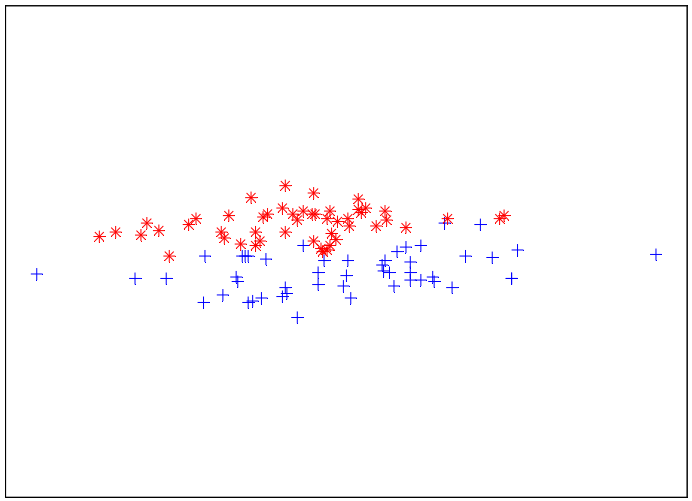


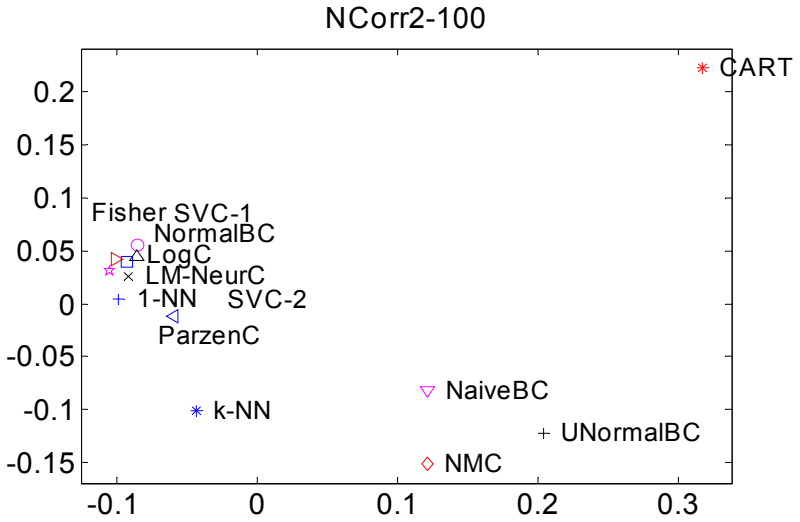


CSP Examples (5,6)

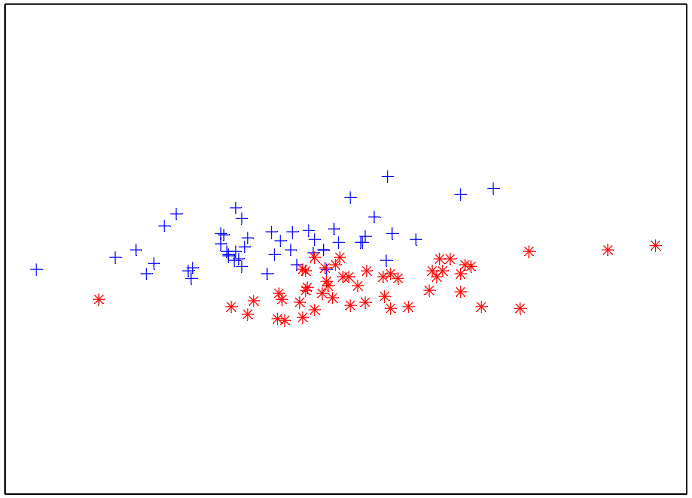


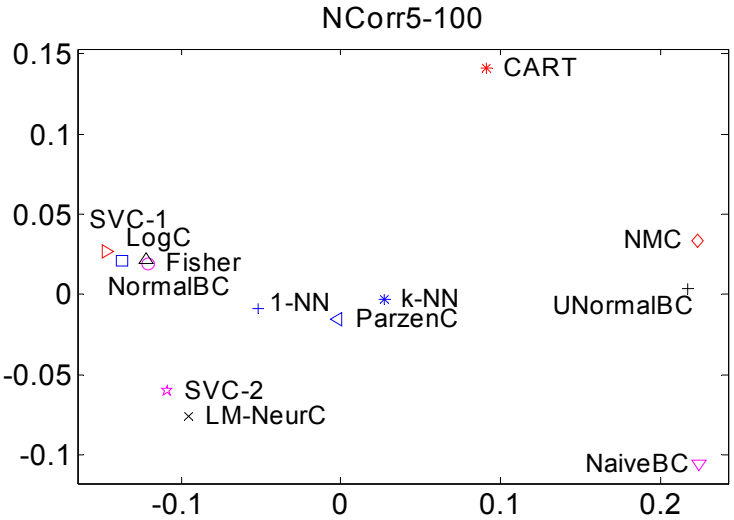


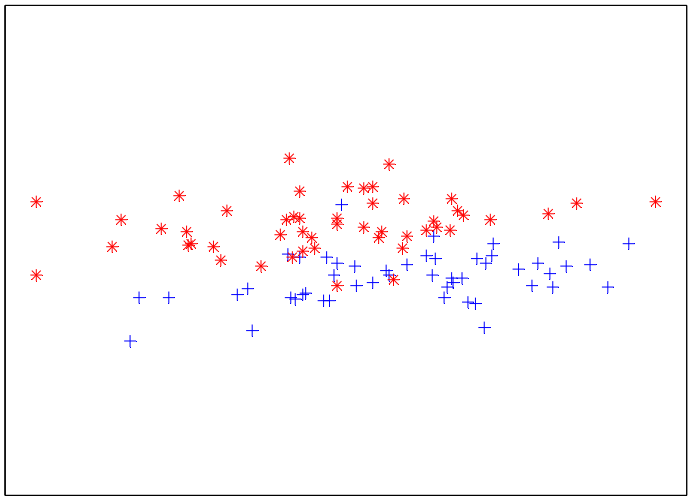


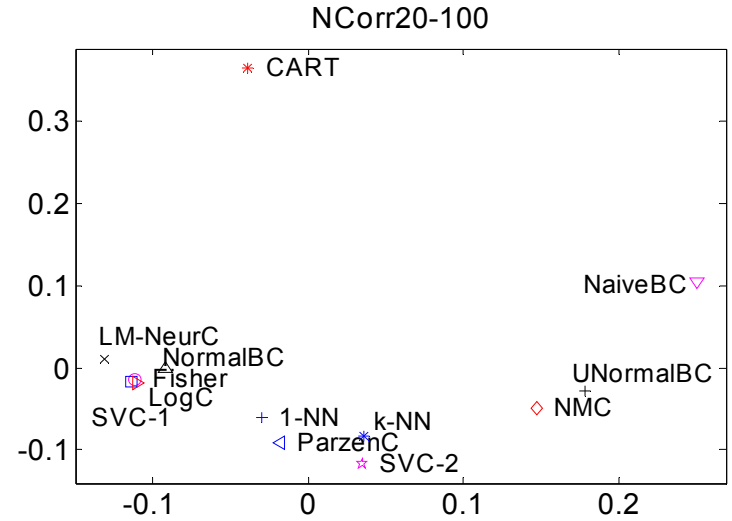


CSP Examples (7,8)



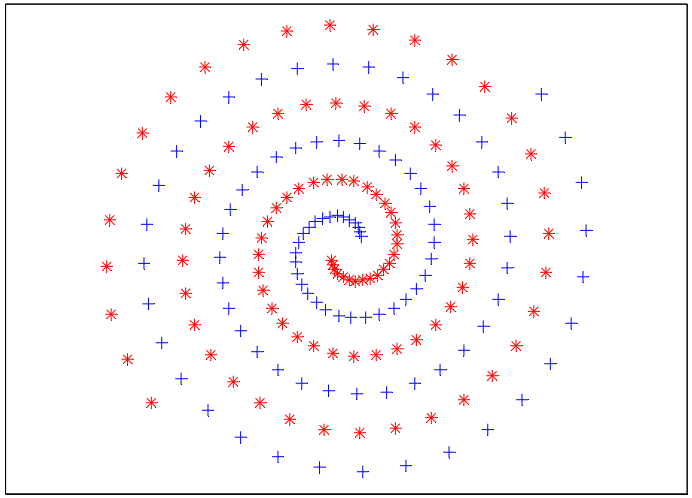




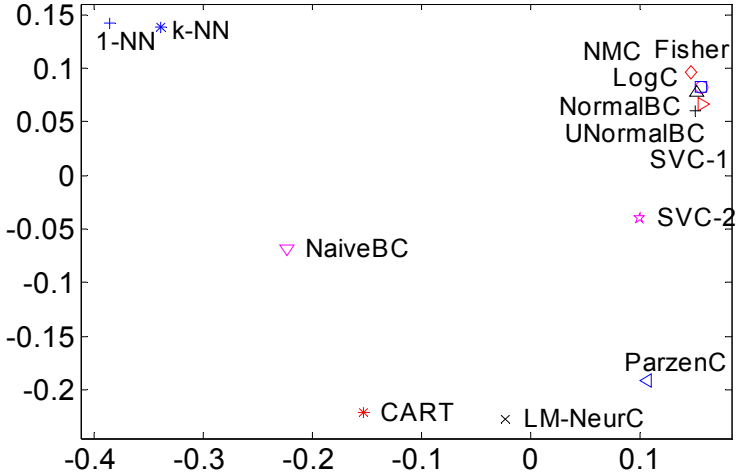


CSP Examples (9,10)

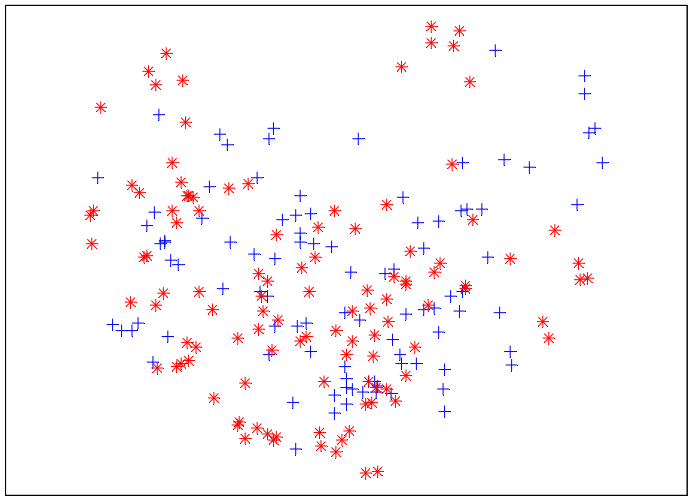
Spirals, 194 2



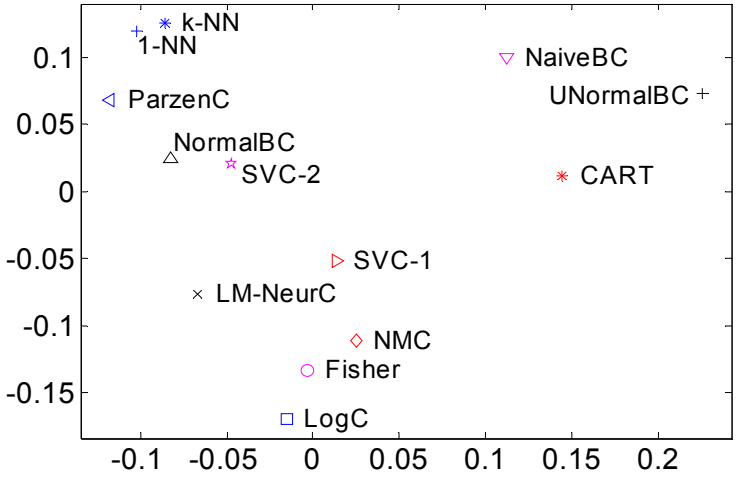
Spirals



Sonar, 208 60

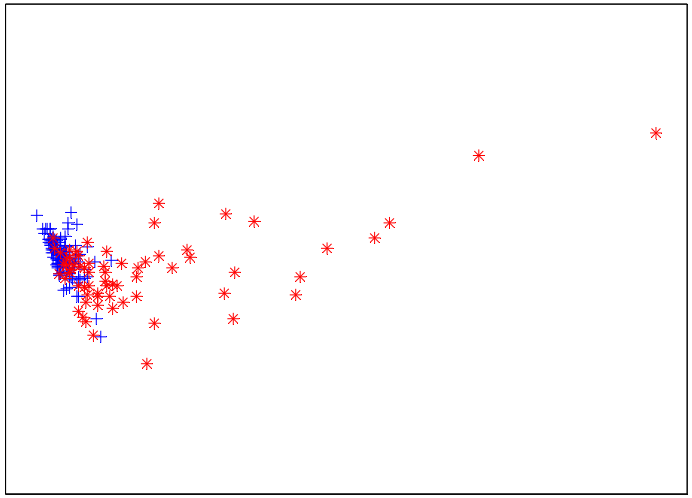


Sonar

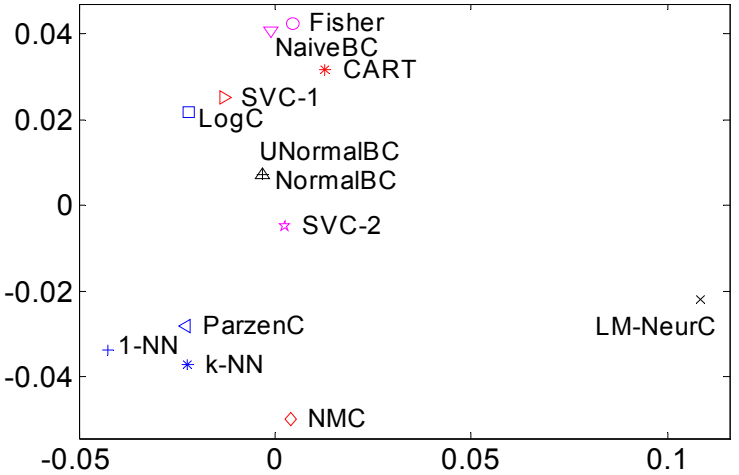


CSP Examples (11,12)

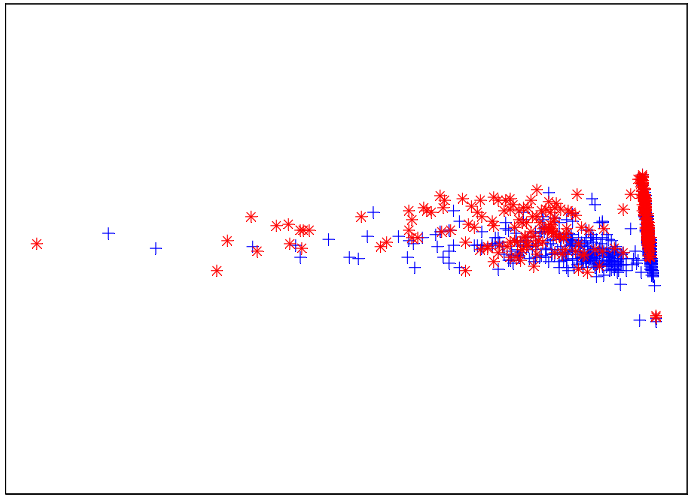
Biomed, 194 5



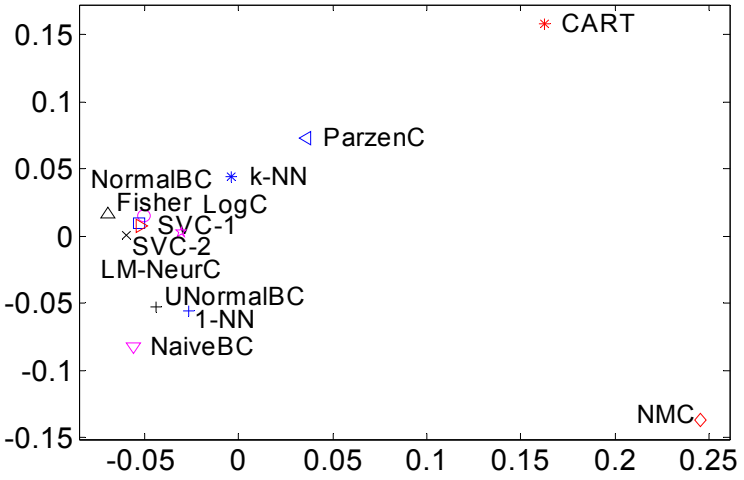
Biomed



Diabetes, 768 8

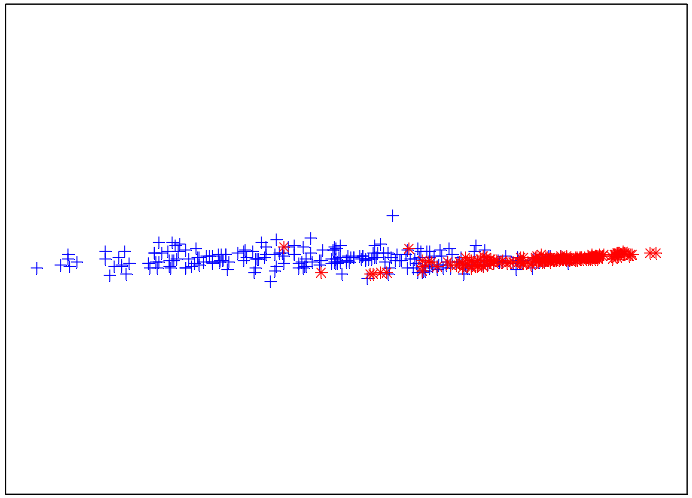


Diabetes

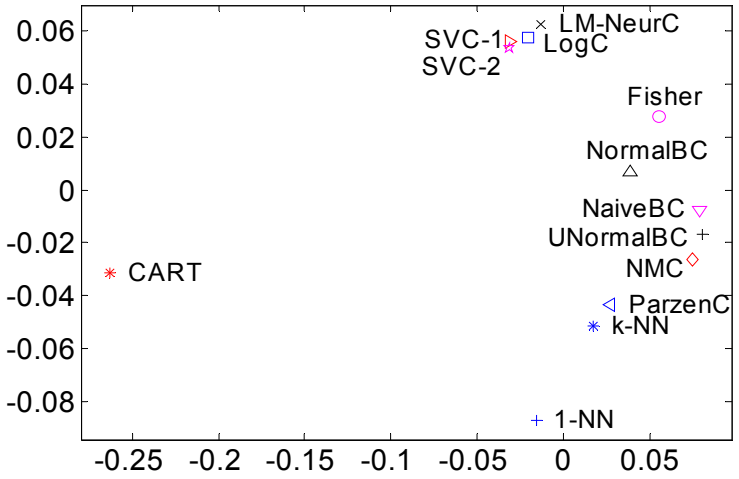


CSP Examples (13,14)

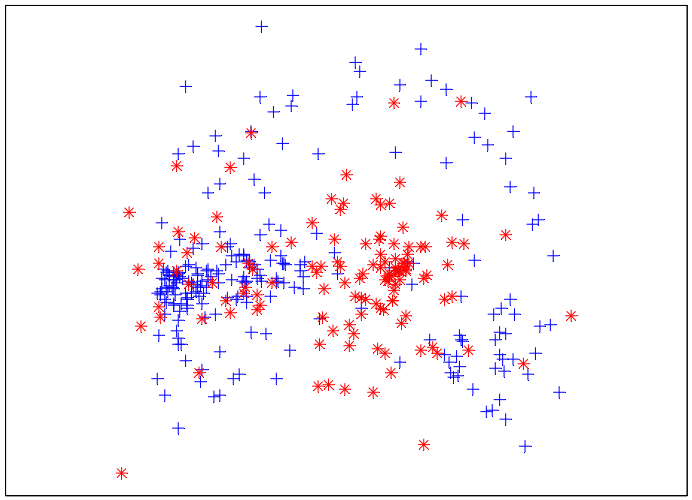
Auto-mpg, 398 6



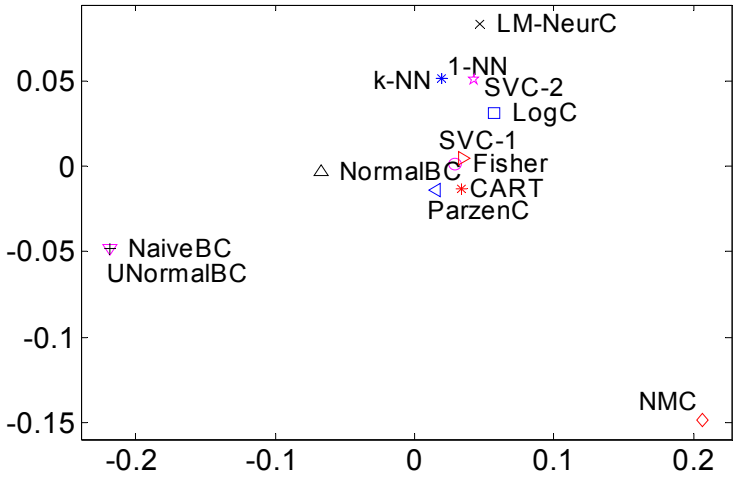
Auto-mpg



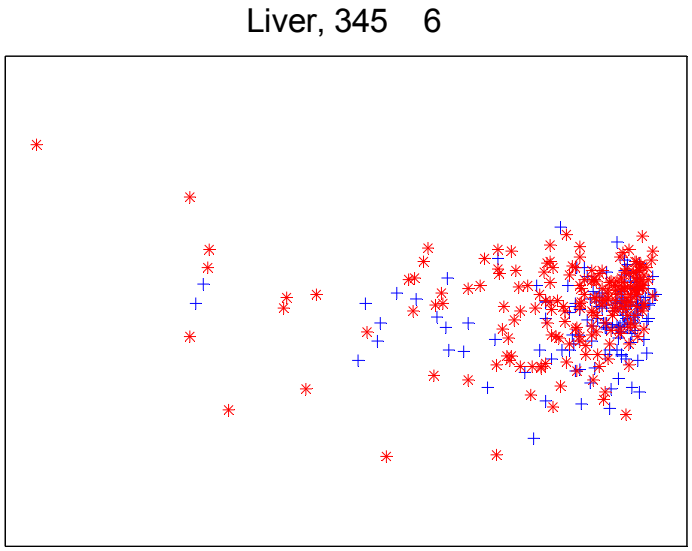
Ionosphere, 351 34

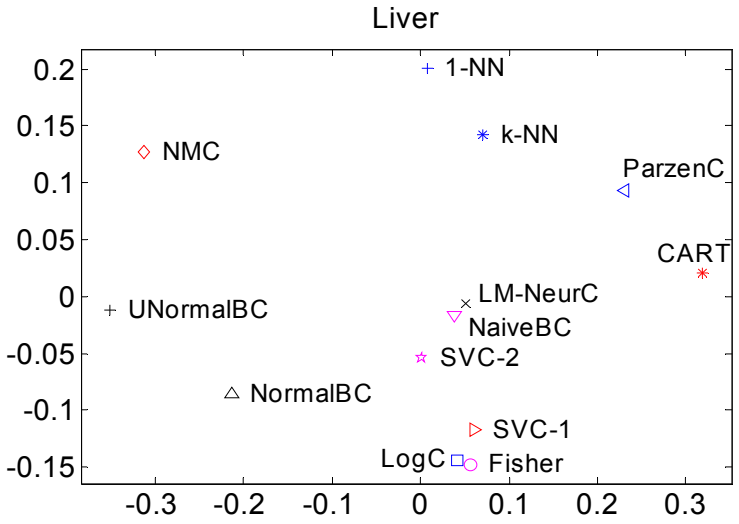


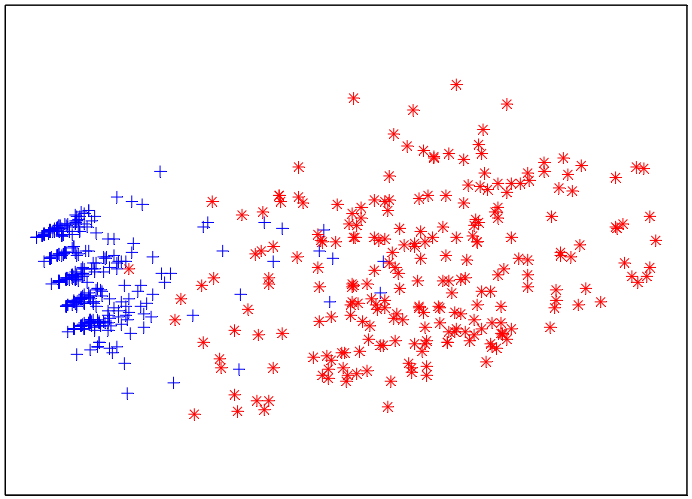
Ionosphere

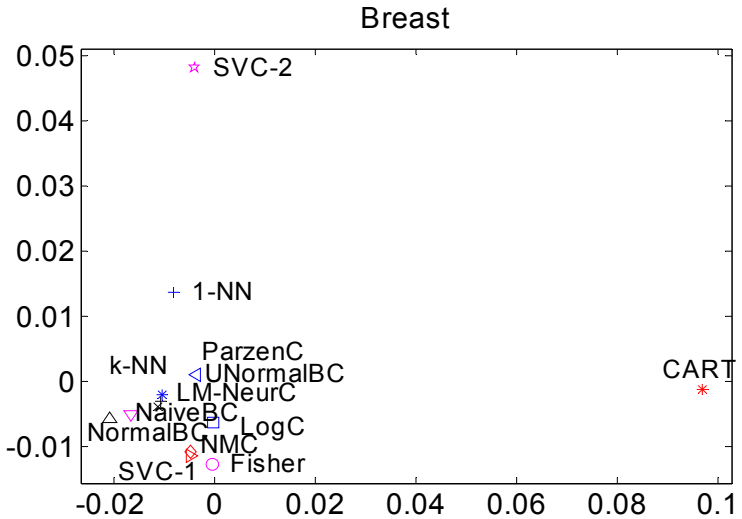


CSP Examples (15,16)









CSP Examples (17,18)