UNIVERSITÀ DI PISA

Dottorato di Ricerca in Ingegneria dell'Informazione

Data Mining and Soft Computing

Francisco Herrera

Research Group on Soft Computing and Information Intelligent Systems (SCI²S)
Dept. of Computer Science and A.I.
University of Granada, Spain

Email: herrera@decsai.ugr.es
http://sci2s.ugr.es
http://decsai.ugr.es/~herrera

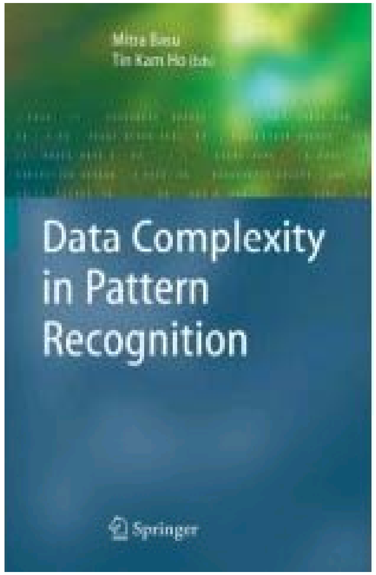
DECSAI
Universidad de Granada

Summary

1. Introduction to Data Mining and Knowledge Discovery
2. Data Preparation
3. Introduction to Prediction, Classification, Clustering and Association
4. Data Mining - From the Top 10 Algorithms to the New Challenges
5. Introduction to Soft Computing. Focusing our attention in Fuzzy Logic and Evolutionary Computation
6. Soft Computing Techniques in Data Mining: Fuzzy Data Mining and Knowledge Extraction based on Evolutionary Learning
7. Genetic Fuzzy Systems: State of the Art and New Trends
8. Some Advanced Topics I: Classification with Imbalanced Data Sets
9. Some Advanced Topics II: Subgroup Discovery
10. Some advanced Topics III: Data Complexity
11. Final talk: How must I Do my Experimental Study? Design of Experiments in Data Mining/Computational Intelligence. Using Non-parametric Tests. Some Cases of Study.

Slides used for preparing this talk:

Data Complexity:
An Overview and New Challenges



Tin Kam Ho
Bell Labs, Alcatel-Lucent

Joint work with Mitra Basu,
Ester Bernado, Martin Law, Albert Orriols

Outline

- ✓ Motivation
- ✓ Class ambiguity, dimensionality and boundary complexity
- ✓ Measures of Geometric Complexity
- ✓ Domains of Competence of Classifiers
- ✓ Other studies
- ✓ Concluding Remarks

Motivation

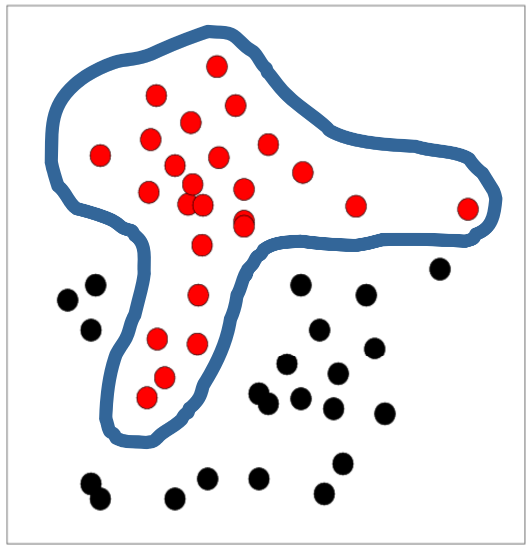
Automatic Classification

Tin Kam Ho
Bell Labs, Alcatel-Lucent

- Classifiers
  - Bayesian classifiers
  - polynomial discriminators
  - nearest-neighbor methods
  - decision trees & forests
  - neural networks
  - genetic algorithms
  - Fuzzy Rule Based Systems
  - support vector machines
  - ensembles and classifier combination
- Why are machines still far from perfect?
- What is still missing in our techniques?

samples

features

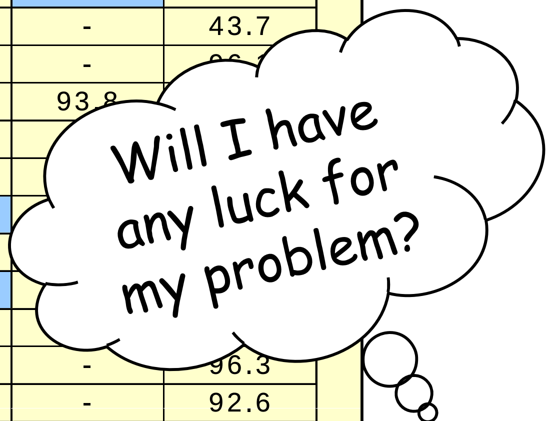


Large Variations in Accuracies of Different Classifiers

classifier

application

	ZeroR	NN1	NNK	NB	C4.5	PART	SMO	XCS
aud	25.3	76.0	68.4	69.6	79.0	**81.2**	-	57.7
aus	55.5	81.9	85.4	77.5	85.2	83.3	84.9	**85.7**
bal	45.0	76.2	87.2	**90.4**	78.5	81.9	-	79.8
bpa	58.0	63.5	60.6	54.3	65.8	65.8	58.0	**68.2**
bps	51.6	83.2	82.8	78.6	80.1	79.0	86.4	**83.3**
bre	65.5	96.0	**96.7**	96.0	95.4	95.3	**96.7**	96.0
cmc	42.7	44.4	46.8	50.6	52.1	49.8	-	52.3
gls	34.6	66.3	66.4	47.6	65.8	69.0	-	**72.6**
h-c	54.5	77.4	83.2	**83.6**	73.6	77.9	-	79.9
hep	79.3	79.9	80.8	83.2	78.9	80.0	**83.9**	83.2
irs	33.3	**95.3**	**95.3**	94.7	**95.3**	95.3	-	94.7
krk	52.2	89.4	94.9	87.0	98.3	98.4	96.1	98.6
lab	65.4	81.1	92.1	**95.2**	73.3	73.9	93.2	75.4
led	10.5	62.4	75.0	74.9	**74.9**	75.1	-	74.8
lym	55.0	83.3	83.6	**85.6**	77.0	71.5	-	79.0
mmg	56.0	63.0	**65.3**	64.7	64.8	61.9	67.0	63.4
mus	51.8	**100.0**	**100.0**	96.4	**100.0**	**100.0**	**100.0**	99.8
mux	49.9	78.6	99.8	61.9	99.9	**100.0**	61.6	**100.0**
pmi	65.1	70.3	73.9	75.4	73.1	72.6	**76.7**	76.0
prt	24.9	34.5	42.5	**50.8**	41.6	39.8	-	43.7
seg	14.3	**97.4**	96.1	80.1	97.2	96.8	-	[illegible]
sick	93.8	96.1	96.3	93.3	**98.4**	97.0	93.8	[illegible]
soyb	13.5	89.5	90.3	**92.8**	91.4	90.3	[illegible]	[illegible]
tao	49.8	**96.1**	96.0	80.8	95.1	93.6	[illegible]	[illegible]
thy	19.5	68.1	65.1	80.6	**92.1**	**92.1**	[illegible]	[illegible]
veh	25.1	69.4	69.7	46.2	73.6	72.6	[illegible]	[illegible]
vote	61.4	92.4	92.6	90.1	96.3	**96.5**	[illegible]	[illegible]
vow	9.1	99.1	**96.6**	65.3	80.7	78.3	[illegible]	[illegible]
wne	39.8	95.6	96.8	**97.8**	94.6	92.9	-	96.3
zoo	41.7	94.6	92.5	**95.4**	91.6	92.5	-	92.6
Avg	44.8	80.0	82.4	78.0	82.1	81.8	84.1	81.7



Many classifiers are in close rivalry with each other. Why?

- Do they represent the limit of our technology?
- What do the new classifiers add to the methodology?
- Is there still value in the older methods?
- Have they used up all information contained in a data set?

When I face a new recognition task …

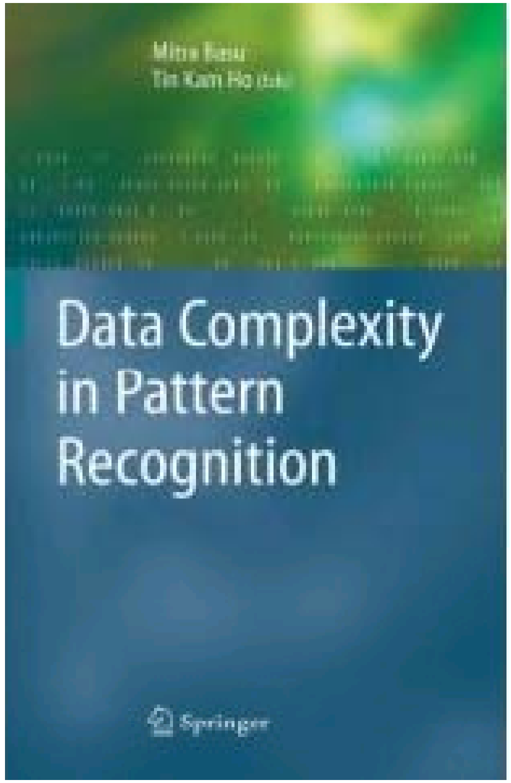
- How much can automatic classifiers do?
- How should I choose a classifier?
- Can I make the problem easier for a specific classifier?

Complexity Measures

Sources of Difficulty in Classification

Tin Kam Ho
Bell Labs, Alcatel-Lucent

- Class ambiguity
- Sample size and dimensionality
- Boundary complexity



We need metrics for analizing problems features and the limits of every learning model.

Limits of Current Learning Algorithms

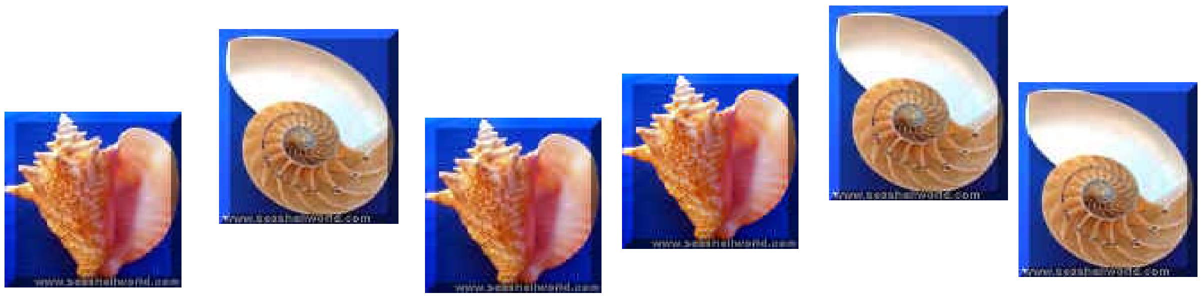
Outline

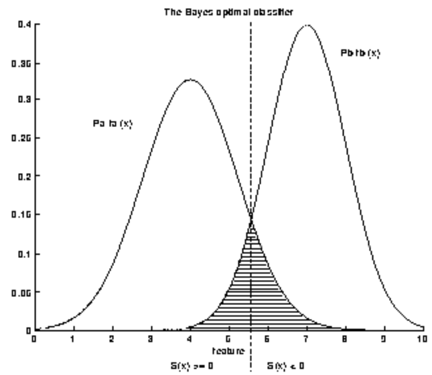
- ✓ Motivation
- ✓ Class ambiguity, dimensionality and boundary complexity
- ✓ Measures of Geometric Complexity
- ✓ Domains of Competence of Classifiers
- ✓ Other studies
- ✓ Concluding Remarks

Class Ambiguity

- Is the concept intrinsically ambiguous?
- Are the classes well defined?

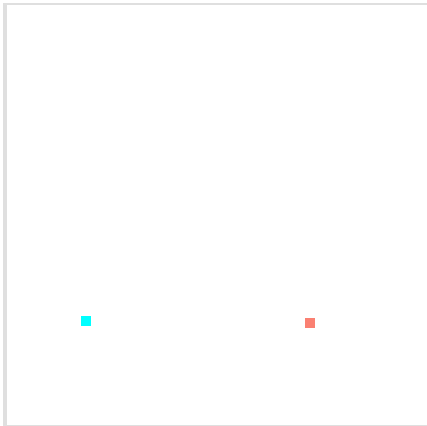
- What information do the features carry?
- Are the features sufficient for discrimination?



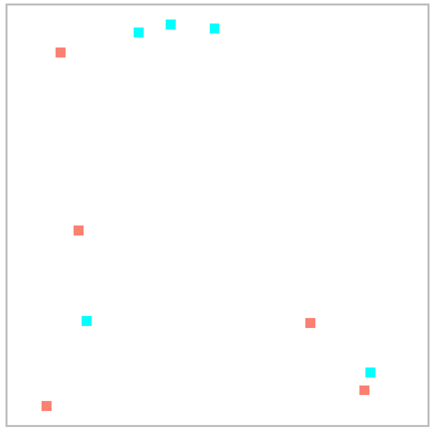


Bayes error

Sampling Density

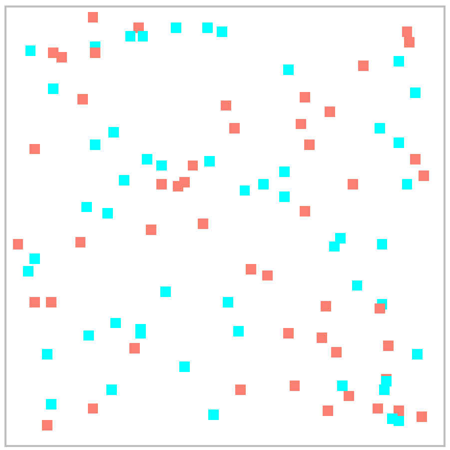


2 points

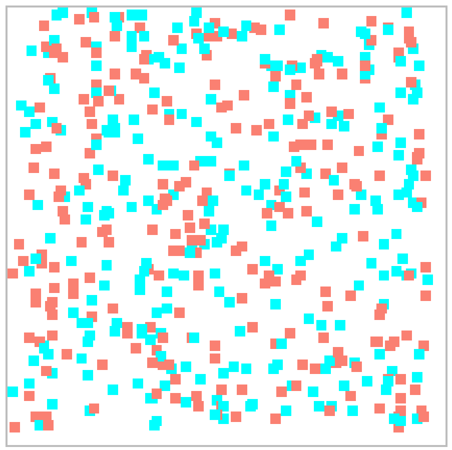


10 points

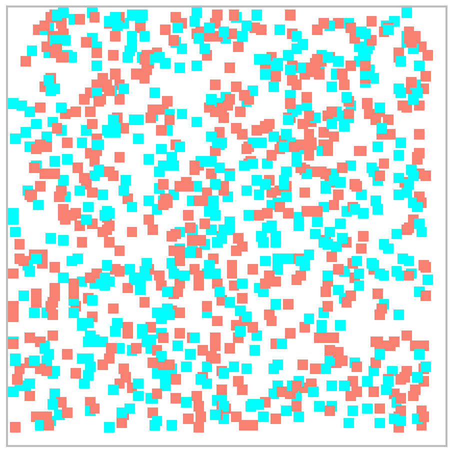
Problem may appear deceptively simple or complex with small samples



100 points



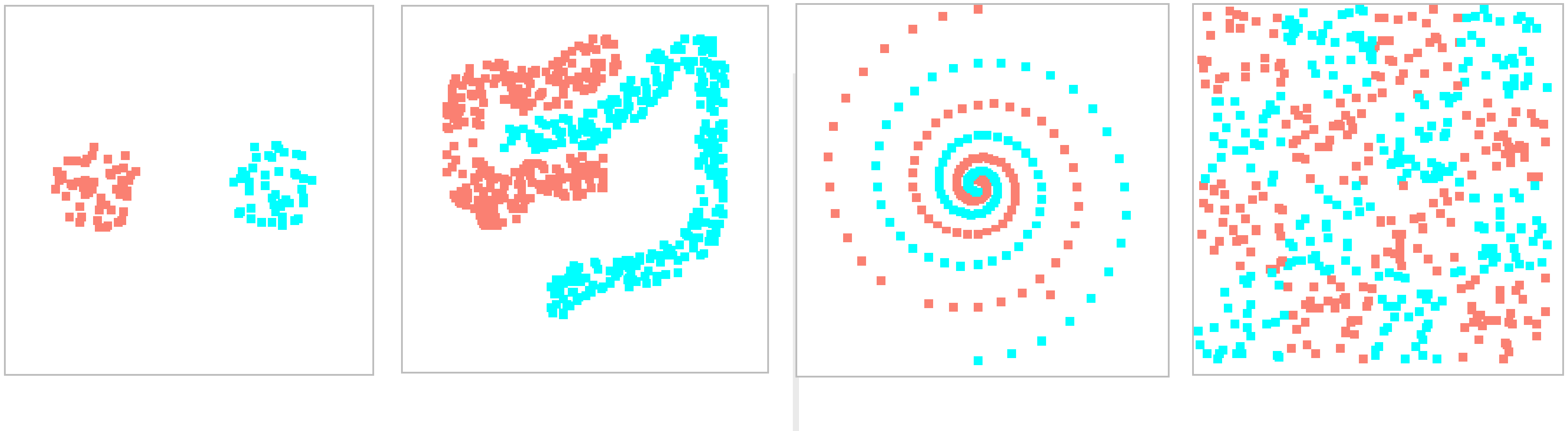
500 points



1000 points

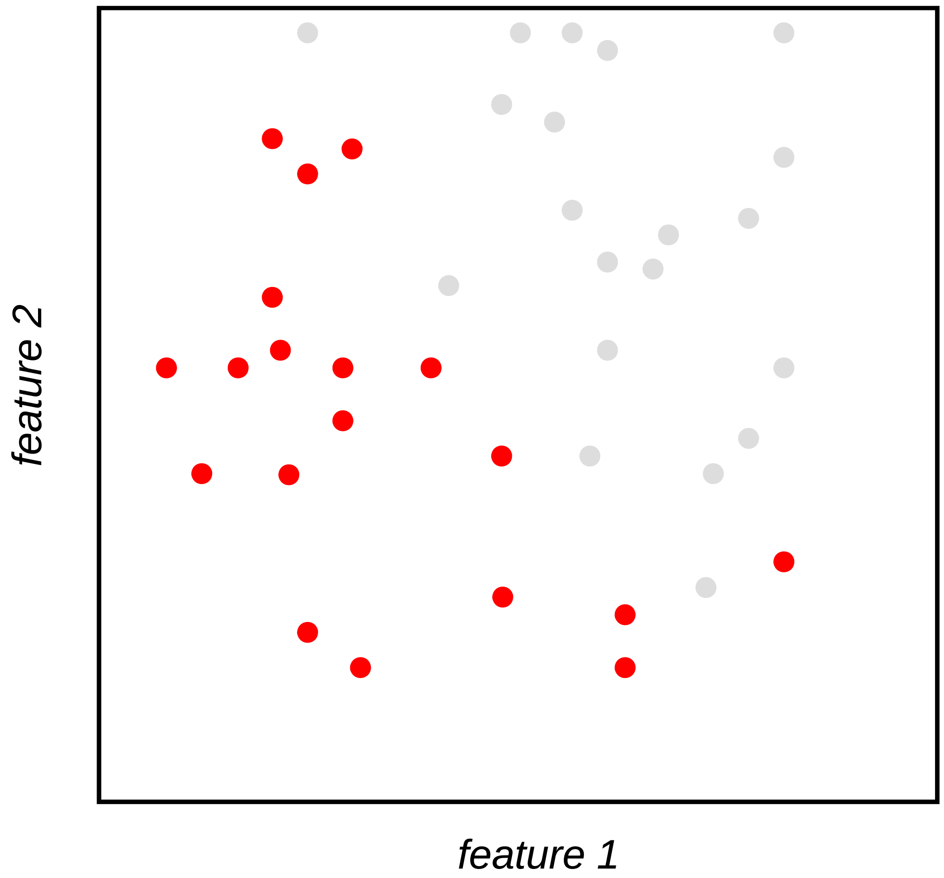
Boundary Complexity

- Kolmogorov complexity
- Length can be exponential in dimensionality
- A trivial description is to list all points & class labels
- Is there a shorter description?



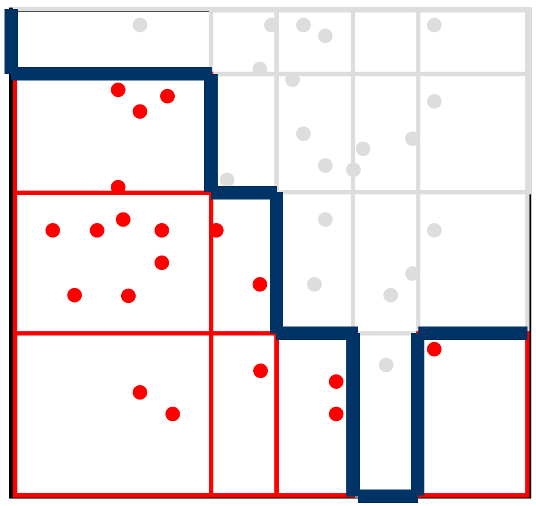
Classification Boundaries As Decided by Different Classifiers

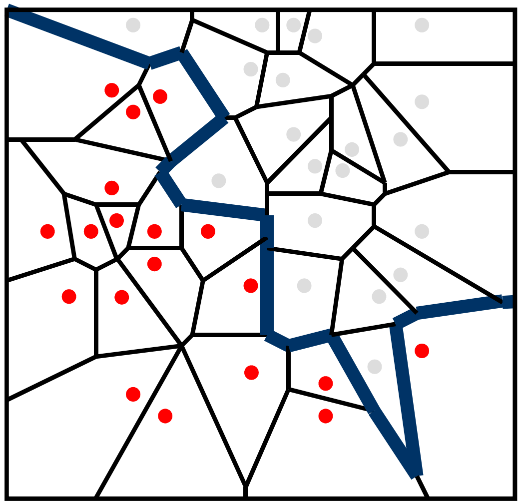
Training samples for a 2D classification problem

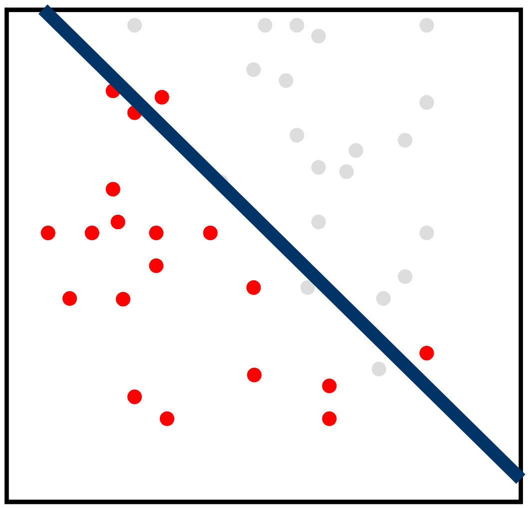


Classification Boundaries Inferred by Different Classifiers

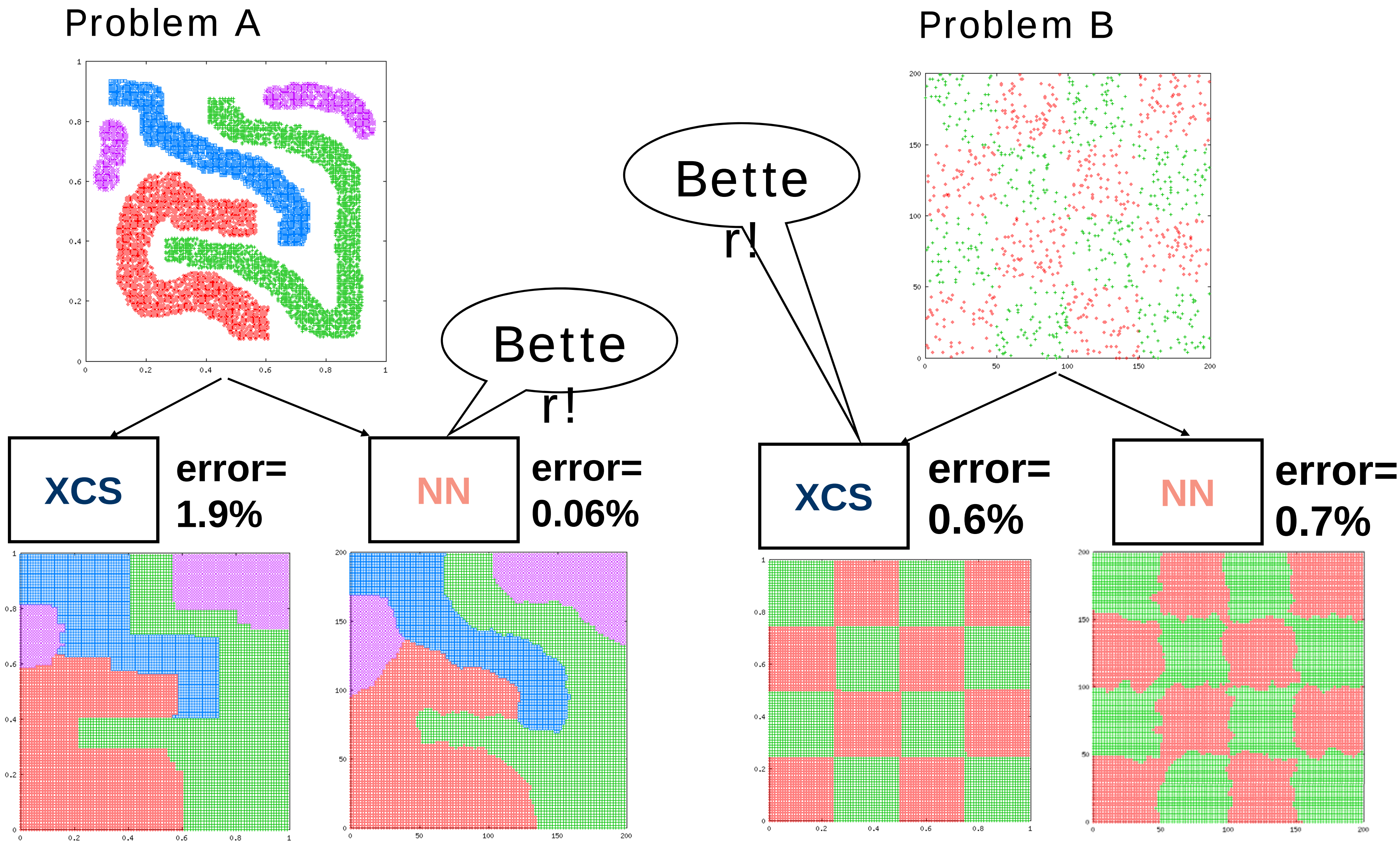
- XCS: a genetic algorithm
- Nearest neighbor classifier
- Linear classifier







Match between Classifiers and Problems



Outline

- Motivation
- Class ambiguity, dimensionality and boundary complexity
- Measures of Geometric Complexity
- Domains of Competence of Classifiers
- Other studies
- Concluding Remarks

Measures of Geometrical Complexity of Classification Problems

The approach: develop mathematical language and algorithmic tools for studying

- Characteristics of geometry & topology of high-dim data
- How they change with feature transformations, noise conditions, and sampling strategies
- How they interact with classifier geometry

Focus on descriptors computable from real data and relevant to classifier geometry

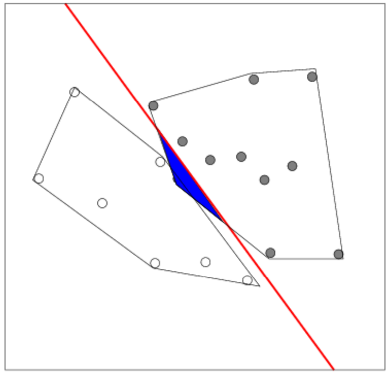
Geometry of Datasets and Classifiers

- Data sets:
  - length of class boundary
  - fragmentation of classes / existence of subclasses
  - global or local linear separability
  - convexity and smoothness of boundaries
  - intrinsic / extrinsic dimensionality
  - stability of these characteristics as sampling rate changes
- Classifier models:
  - polygons, hyper-spheres, Gaussian kernels, axis-parallel hyper-planes, piece-wise linear surfaces, polynomial surfaces, their unions or intersections, ...

Measures of Geometric Complexity

Degree of Linear Separability

- Find separating hyper-plane by linear programming
- Error counts and distances to plane measure separability



Fisher's Discriminant Ratio

- Classical measure of class separability
- Maximize over all features to find the most discriminating

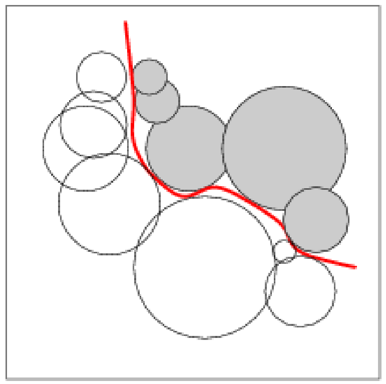
$$f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

Length of Class Boundary

- Compute minimum spanning tree
- Count class-crossing edges

Shapes of Class Manifolds

- Cover same-class pts with maximal balls
- Ball counts describe shape of class manifold



Measures of Geometrical Complexity

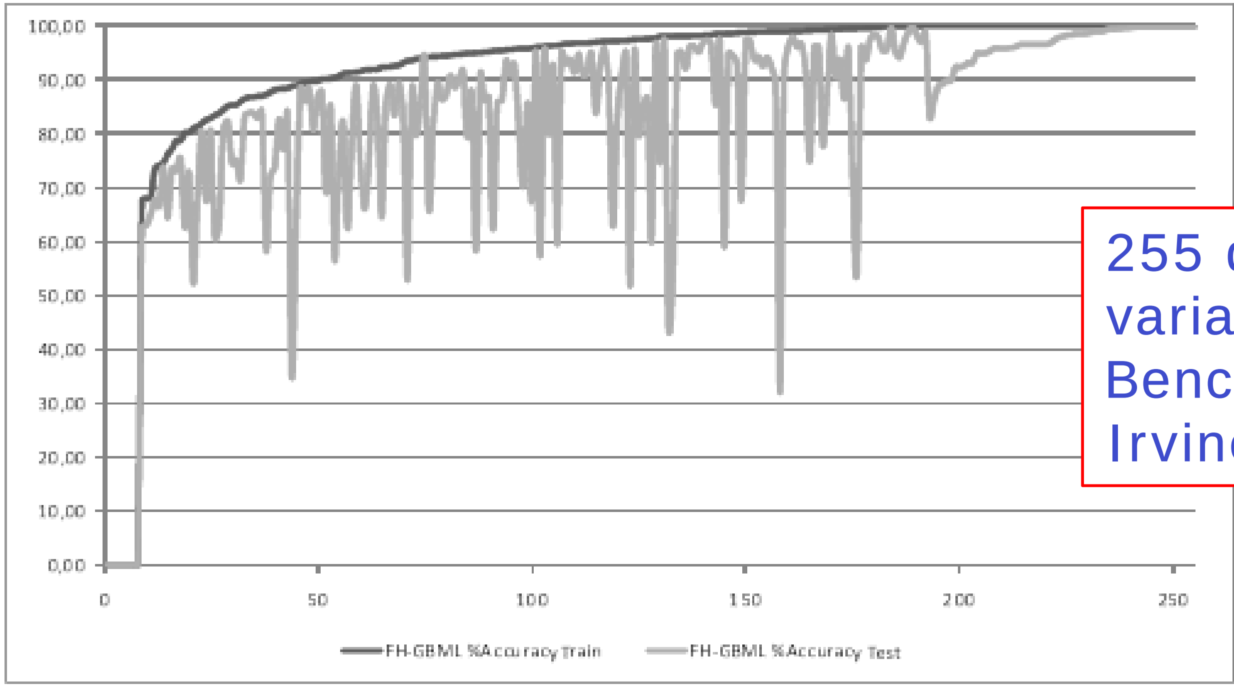
F1	maximum Fisher's discriminant ratio
F2	volume of overlap region
F3	maximum (individual) feature efficiency
L1	minimized error by linear programming (LP)
L2	error rate of linear classifier by LP
L3	nonlinearity of linear classifier by LP
N1	fraction of points on boundary (MST method)
N2	ratio of average intra/inter class NN distance
N3	error rate of 1NN classifier
N4	nonlinearity of 1NN classifier
T1	fraction of points with associated adherence subsets retained
T2	average number of points per dimension

Example

Method Ishibuchi FH-GGBML, 2005, IEEE TSMC

Measure	Description
F2	volume of overlap region
L1	minimized sum of error distance by linear programming
L2	error rate of linear classifier by Linear Programming
N2	ratio of average intra/inter class NN distance
N3	error rate of 1NN classifier
N4	non-linearity of 1NN classifier
T2	average number of points per dimension

Table 1: Complexity metrics used in this study



255 data sets with 2-variables, nonseparable Benchmarking data from UC-Irvine archive

Figure 2: Accuracy in Train/Test for FH-GBML sorted by train accuracy

Method Ishibuchi FH-GGBML

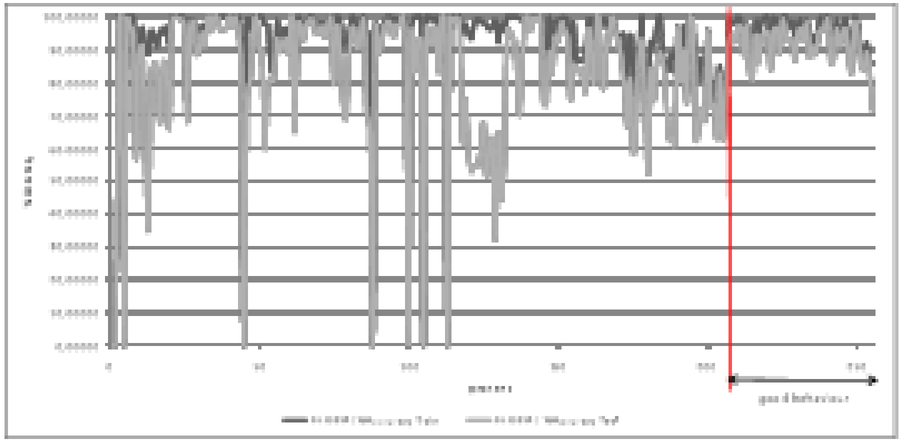


Figure 3: Accuracy in Train/Test sorted by F2

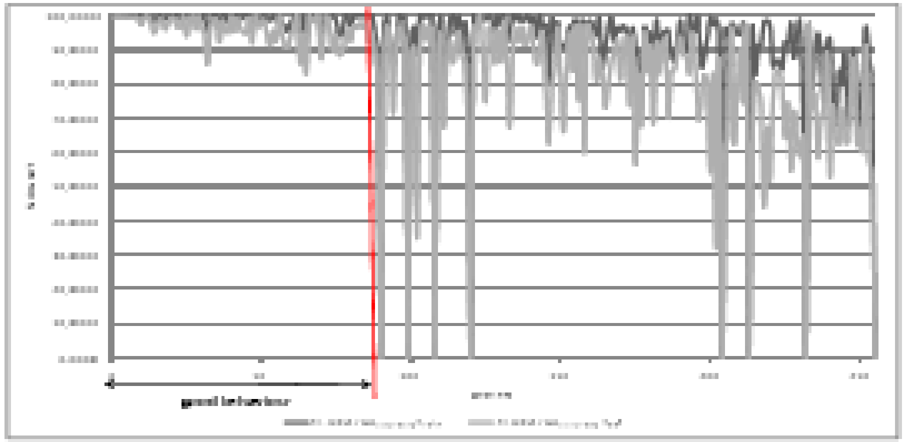


Figure 4: Accuracy in Train/Test sorted by N2

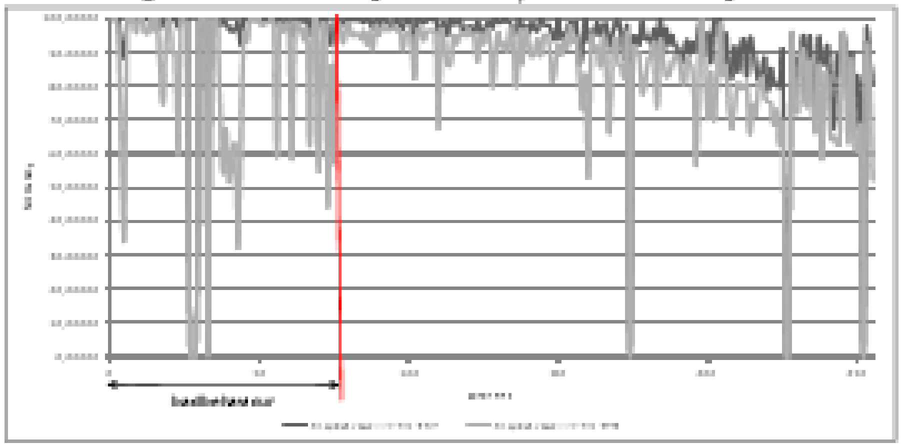


Figure 5: Accuracy in Train/Test sorted by N3

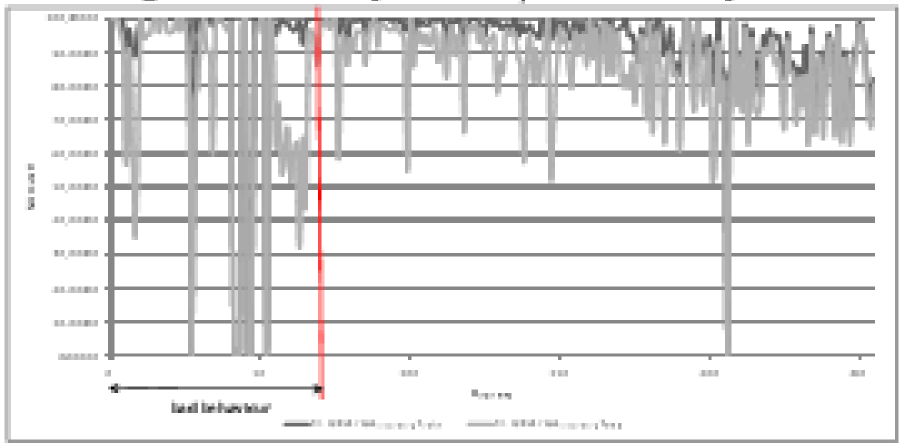


Figure 6: Accuracy in Train/Test sorted by N4

Method Ishibuchi FH-GGBML

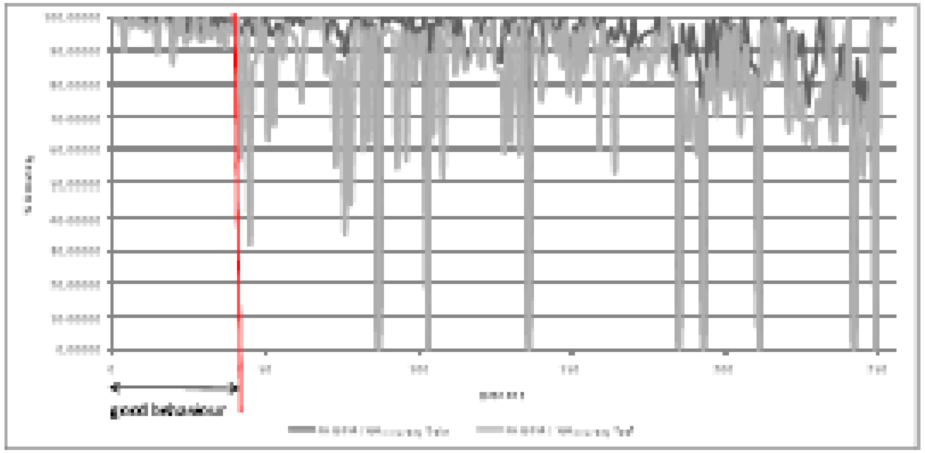


Figure 7: Accuracy in Train/Test sorted by L1

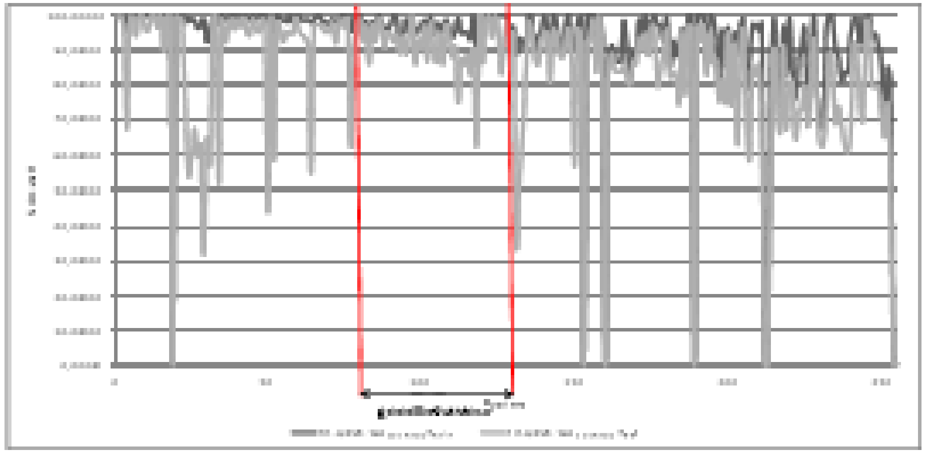


Figure 8: Accuracy in Train/Test sorted by L2

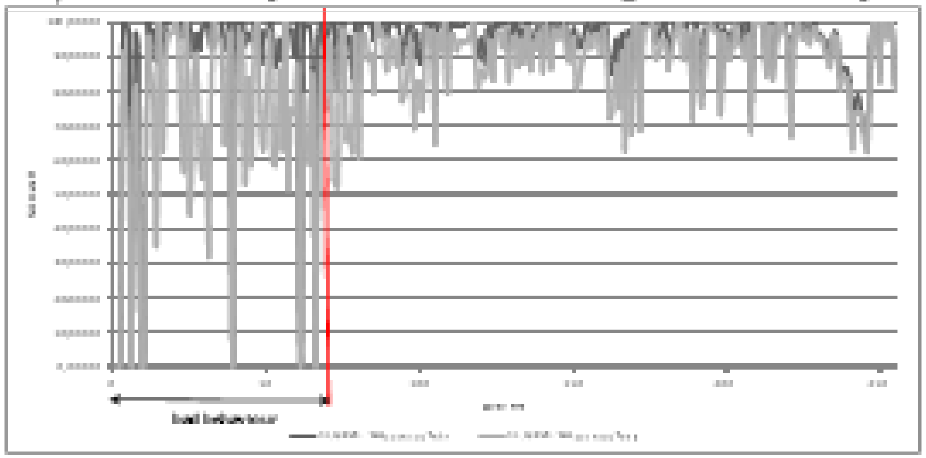


Figure 9: Accuracy in Train/Test sorted by T2

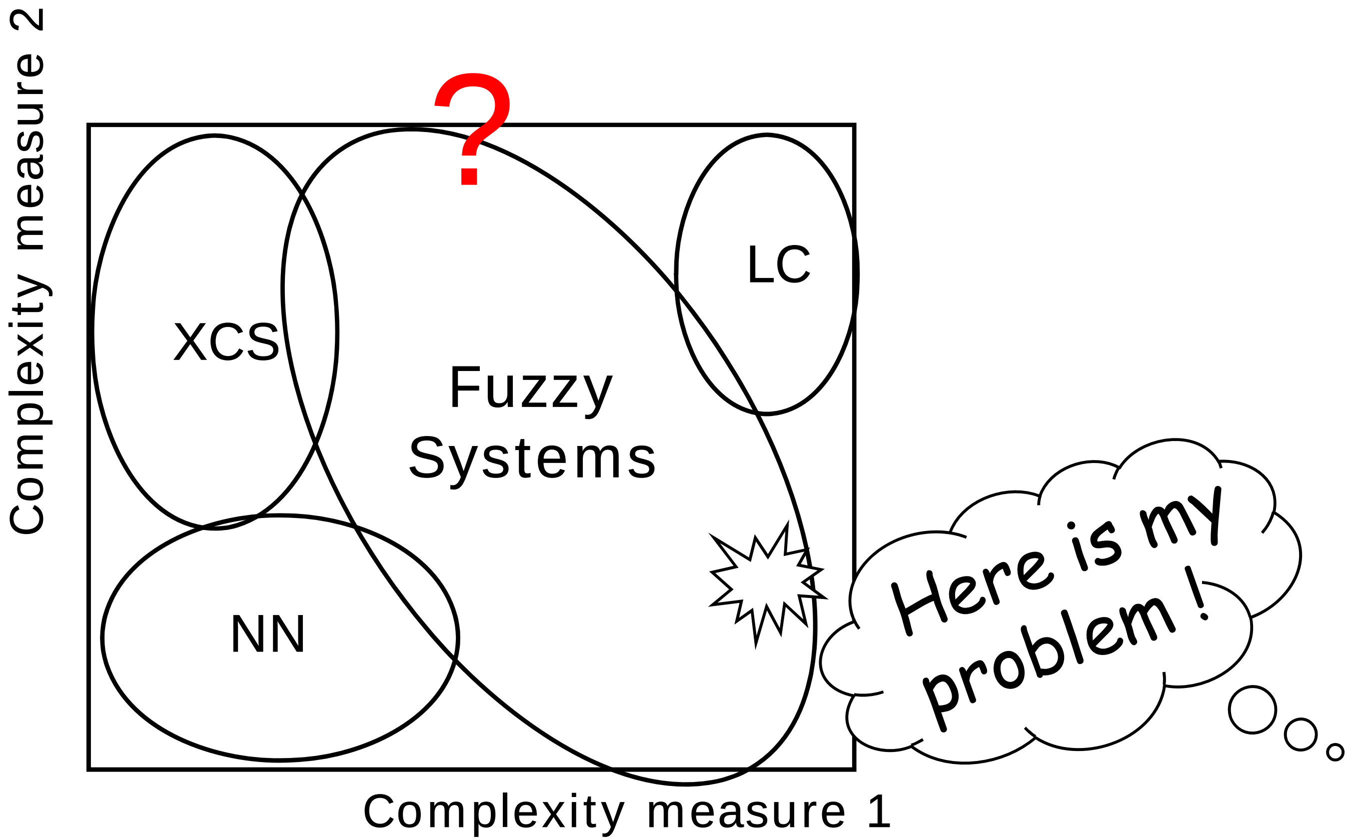
Some Advanced Topics III: Data Complexity

Outline

- ✓ Motivation
- ✓ Class ambiguity, dimensionality and boundary complexity
- ✓ Measures of Geometric Complexity
- ✓ Domains of Competence of Classifiers
- ✓ Other studies
- ✓ Concluding Remarks

Domains of Competence of Classifiers

- Given a classification problem, determine which classifier is the best for it



Domains of Competence of Classifiers

Method Ishibuchi FH-GGBML

Some interesting intervals

Interval	FH-GBML Behaviour
$N2 < 0.23$	*good behaviour*
$L1 < 0.1585$	*good behaviour*
$F2 = 1$	*good behaviour*
$0.04 < L2 < 0.1$	*good behaviour*
$N3 = 0$	*bad behaviour*
$N4 = 0$	*bad behaviour*
$T2 < 7$	*bad behaviour*

Table 2: Significant intervals

Domains of Competence of Classifiers

Method Ishibuchi FH-GGBML

Rules with a metric

Id.	Rule	Support	Avg. % Train St.Dev	Train Diff.	Avg. % Test St.Dev	Test Diff.
R1+	If N2[X] < 0.23 then good behaviour	32.549%	99.10000% 1.56873	6.8880%	96.40400% 3.73928	12.6190%
R2+	If L1[X] < 0.1585 then good behaviour	16.471%	98.79382% 1.88762	6.5810%	96.63110% 6.92474	12.8459%
R3+	If F2[X] = 1 then good behaviour	19.216%	95.99478% 4.08713	3.7820%	91.47715% 5.74098	7.6919%
R4+	If 0.04 < L2[X] < 0.1 then good behaviour	19.608%	97.07823% 2.46866	4.8654%	91.73752% 6.76988	7.9523%
R1-	If N3[X] = 0 then bad behaviour	18.039%	90.17976% 28.26869	-2.03303%	78.79163% 30.81635	-4.99360%
R2-	If N4[X] = 0 then bad behaviour	27.059%	88.73440% 30.12516	-3.47839%	77.14338% 31.48844	-6.64185%
R3-	If T2[X] < 7 then bad behaviour	30.588%	86.47399% 29.72216	-5.73880%	69.42453% 28.89741	-14.36070%

Table 3: Rules with one metric obtained from the intervals

Domains of Competence of Classifiers

Method Ishibuchi FH-GGBML

Rules with a metric

Id.	Rule	Support	Avg. % Train St.Dev	Train Diff.	Avg. % Test St.Dev	Test Diff.
R5+	If L1[X] < 0.1585 and not T2[X] < 7 then good behaviour	10.196%	98.72043% 1.72081	6.5076%	97.29695% 2.3808	13.5117%
R6+	If N2[X] < 0.1585 and not N3[X] = 0 then good behaviour	22.353%	98.68990% 1.74822	6.4771%	95.46808% 3.88134	11.6829%
R7+	If 0.04 < L2[X] < 0.1 and not T2[X] < 7 then good behaviour	14.902%	96.88916% 2.22073	4.6764%	93.02681% 4.67047	9.2416%
R4-	If N3[X] = 0 and not L1[X] < 0.19 then bad behaviour	12.941%	86.45058% 32.71477	-5.76221%	71.02749% 33.35019	-12.75774%
R5-	If N3[X] = 0 and not N2[X] < 0.23 then bad behaviour	7.843%	77.64346% 39.94092	-14.56933%	53.22919% 32.01059	-30.55604%
R6-	If N4[X] = 0 and not L1[X] < 0.19 then bad behaviour	20.000%	84.82022% 34.26575	-7.39257%	69.74147% 33.6301	-14.04376%
R7-	If N4[X] = 0 and not N2[X] < 0.23 then bad behaviour	14.510%	79.00123% 38.78489	-13.21156%	59.12644% 33.87836	-24.65879%

Table 4: Rules with two metrics obtained from the intervals

Domains of Competence of Classifiers

Method Ishibuchi FH-GGBML

Combination of Rules

Id.	Rule	Support	Avg. % Train St.Dev	Train Diff.	Avg. % Test St.Dev	Test Diff.
RDP	If R1+ or R2+ or R3+ or R4+ or R5+ or R6+ or R7+ then *good behaviour*	50.196%	97.87275% 3.24086	5.65996%	94.10161% 6.21307	10.31638%
RDN	If R1- or R2- or R3- or R4- or R5- or R6- or R7- then *bad behaviour*	41.176%	89.77980% 26.23892	-2.43299%	76.29024% 27.76105	-7.49499%

Table 6: Disjunction Rules from all rules

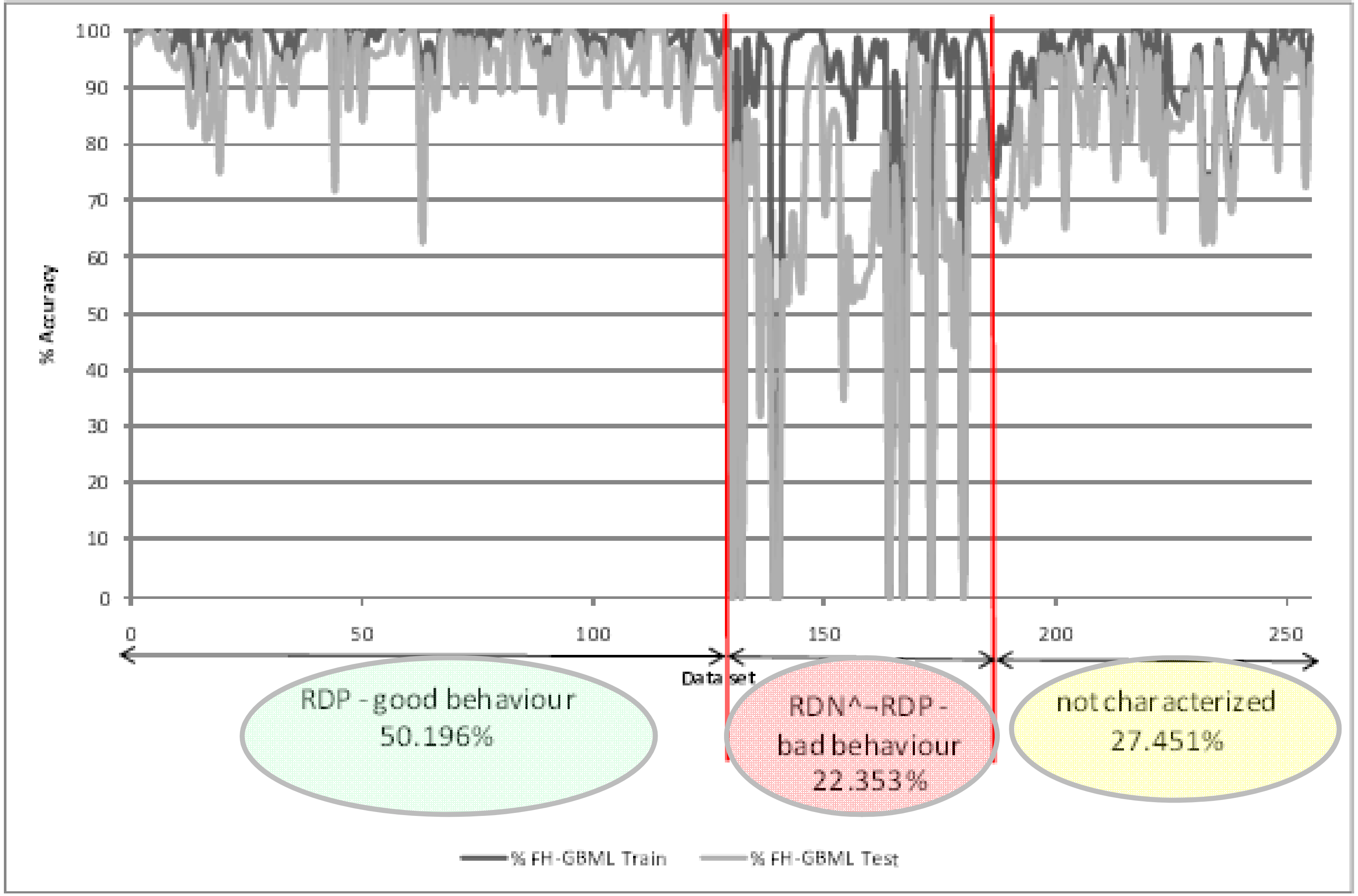
Id.	Rule	Support	Avg. % Train St.Dev	Train Diff.	Avg. % Test St.Dev	Test Diff.
RDP∧ RDN	If RDP and RDN then good behaviour	18.824%	99.41149% 1.83755	7.19870%	95.18826% 7.09706	11.40303%
RDP∧ ¬RDN	If RDP and not RDN then good behaviour	31.373%	96.94950% 3.546016	4.73671%	93.44961% 5.56263	9.66438%
RDN∧ ¬RDP	If RDN and not RDP then bad behaviour	22.353%	81.66890% 33.60499	-10.54389%	60.37611% 28.72427	-23.40912%

Table 7: Intersections of the disjunction rules

Domains of Competence of Classifiers

Method Ishibuchi FH-GGBML

Combination of Rules – Behaviour Caracterization



Domains of Competence of Classifiers

Method Ishibuchi FH-GGBML

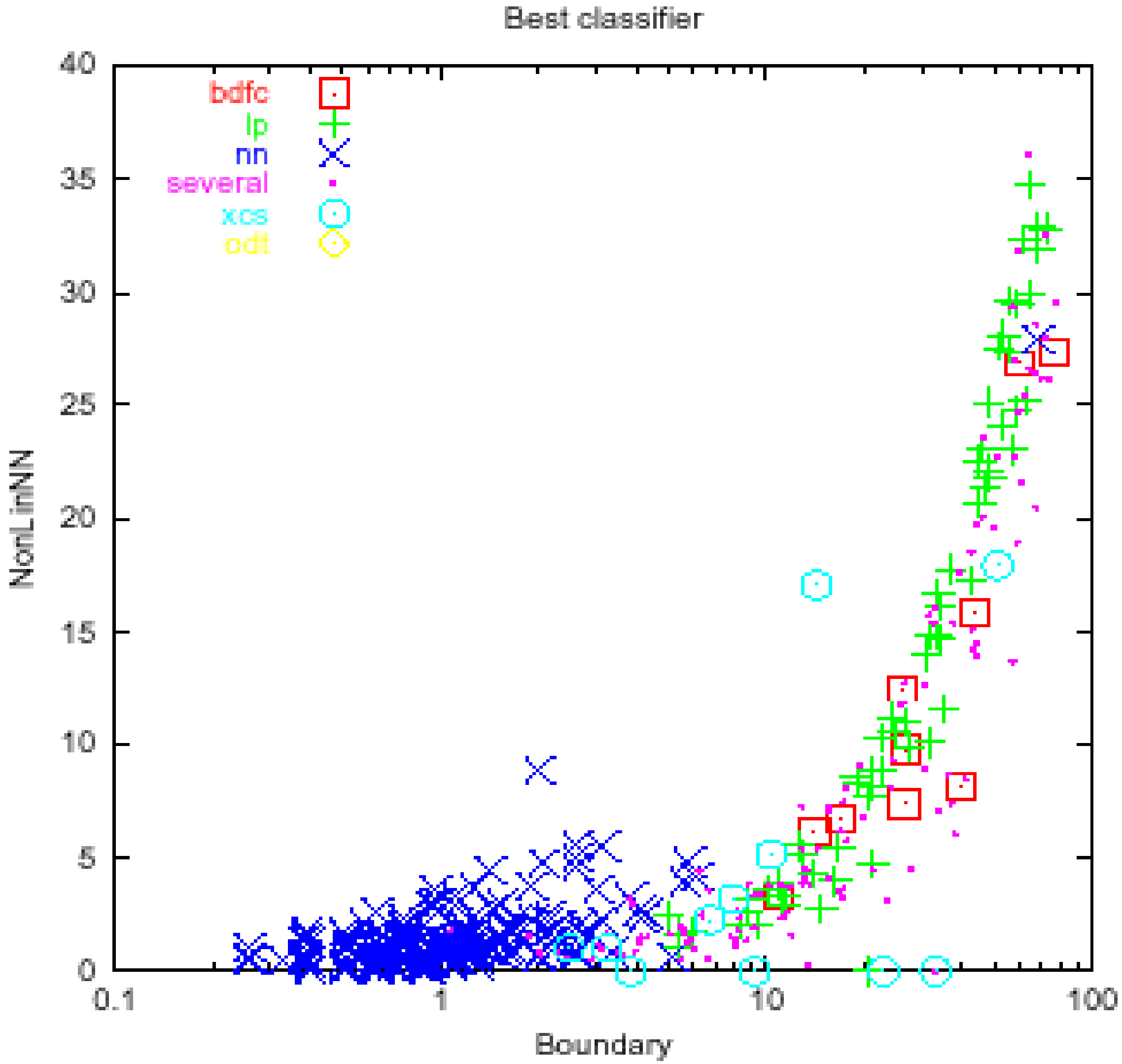
Combination of Rules – Behaviour Caracterization

Id.	Rule
RDP	If (N2[X] < 0.23) or (L1[X] < 0.1585) or (F2[X] = 1) or (0.04 < L2[X] <0.1) or (L1[X] < 0.1585 and not T2[X] < 7) or (N2[X] < 0.1585 and not N3[X] = 0) or (0.04 < L2[X] <0.1 and not T2[X] < 7) then good behaviour
RDN$\wedge\neg$RDP	If [(N3[X] = 0) or (N4[X] = 0) or (T2[X] < 7) or (N3[X] = 0 and not L1[X] < 0.19) or (N3[X] = 0 and not N2[X] < 0.23) or (N4[X] = 0 and not L1[X] < 0.19) or (N4[X] = 0 and not N2[X] < 0.23)] and not [(N2[X] < 0.23) or (L1[X] < 0.1585) or (F2[X] = 1) or (0.04 < L2[X] <0.1) or (L1[X] < 0.1585 and not T2[X] < 7) or (N2[X] < 0.1585 and not N3[X] = 0) or (0.04 < L2[X] <0.1 and not T2[X] < 7)] then bad behaviour

Table 8: RDP and RDN$\wedge\neg$RDP rules

Domains of Competence of Classifiers

Comparison of classifers with a measure

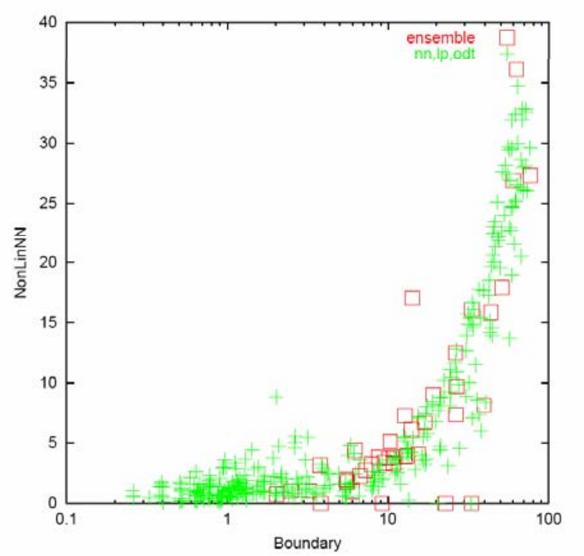


Best Classifier for Benchmarking Data

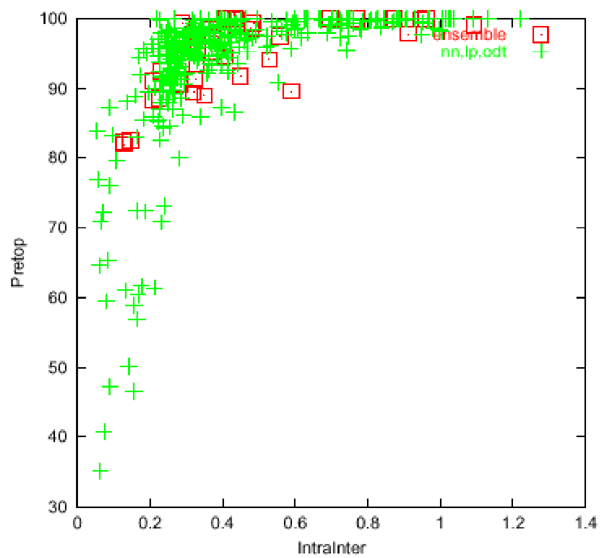
Domains of Competence of Classifiers

Best Classifier Being
nn,lp,odt vs an ensemble technique

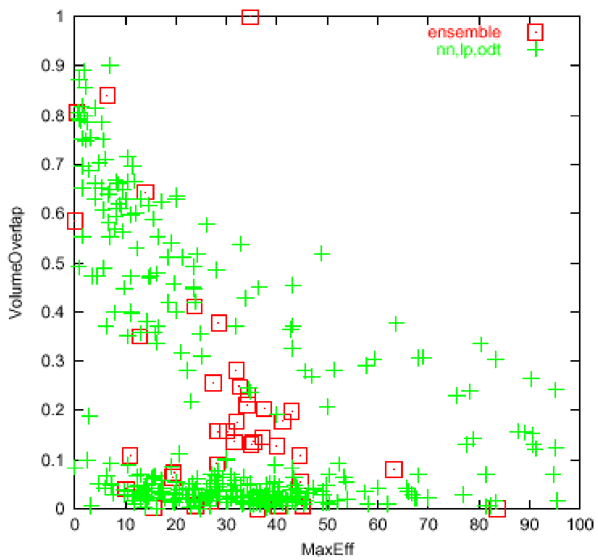
Boundary-NonLinNN



IntraInter-Pretop



MaxEff-VolumeOverlap



ensemble

+ nn,lp,odt

Outline

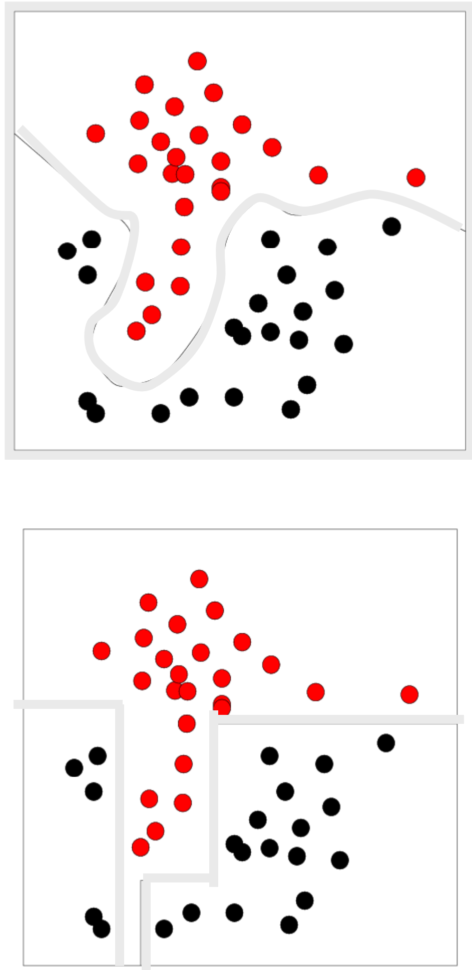
- ✓ Motivation
- ✓ Class ambiguity, dimensionality and boundary complexity
- ✓ Measures of Geometric Complexity
- ✓ Domains of Competence of Classifiers
- ✓ Other studies
- ✓ Concluding Remarks

Complexity and Sample Sparsity

Sparse Sample & complex geometry cause ill-posedness

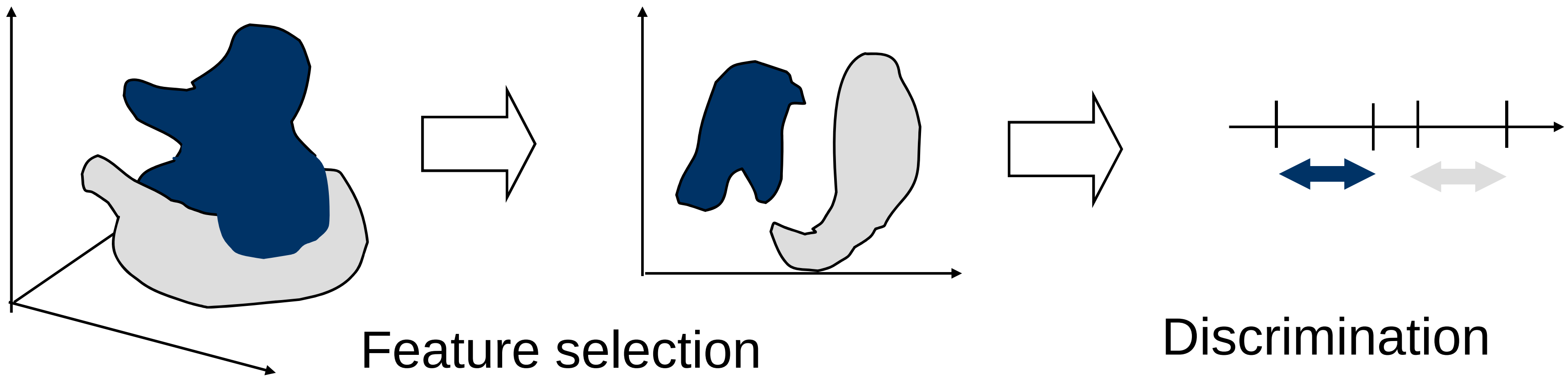
Careful statistical procedures are needed to infer complexity of the data population from those of the training samples

Complexity estimation requires further hypotheses on data geometry and sampling processes



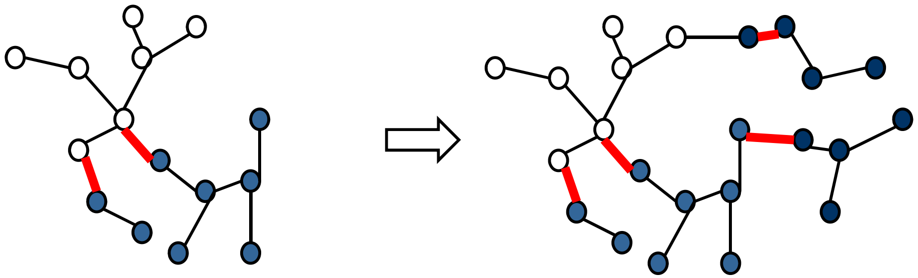
Complexity and Data Dimensionality:
Class Separability after Dimensionality Reduction

- Feature selection may change the difficulty of a classification problem
  - Widening the gap between classes
  - Compressing the discriminatory information
  - Removing irrelevant dimensions
- It is often unclear to what extent these happen
- We seek quantitative description of such changes

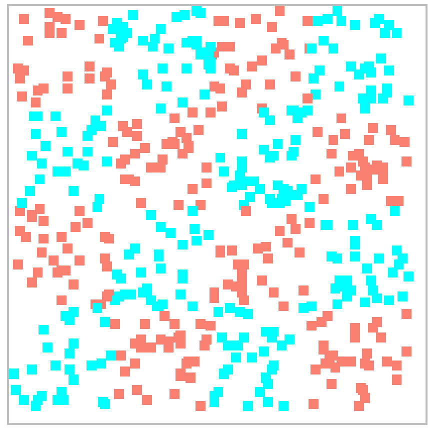


Extensions of the Study on Data Complexity

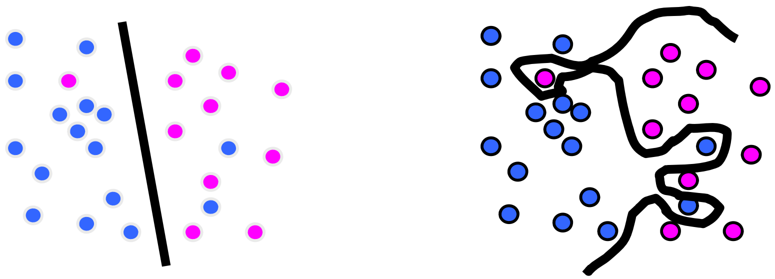
Multi-Class Measures



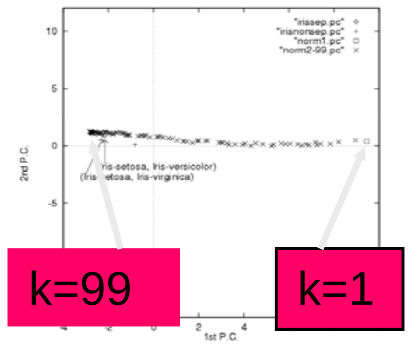
Global vs. Local Properties



Intrinsic Ambiguity & Mislabeling



Task Trajectory with Changing Sampling & Noise Conditions

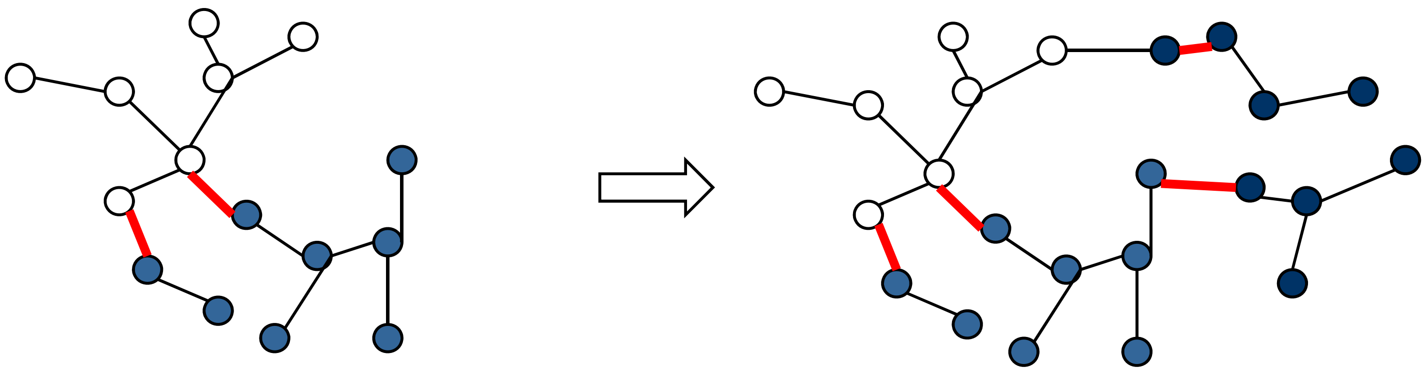


Extension to Multiple Classes

- Fisher's discriminant score ➔ Mulitple discriminant scores

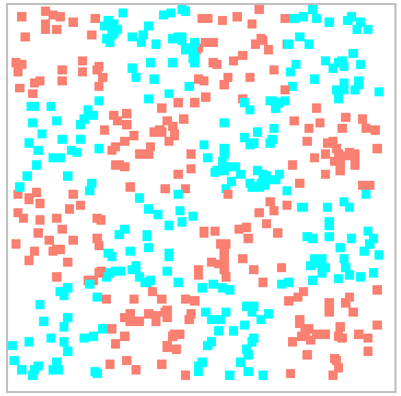
$$f = \frac{(\mu_1 - \mu_2)^2}{(\sigma_1^2 + \sigma_2^2)} \quad \Longrightarrow \quad f = \frac{\sum_{i=1,j=1,i\neq j}^{C} p_i p_j (\mu_i - \mu_j)^2}{\sum_{i=1}^{C} p_i \sigma_i^2}$$

- Boundary point in a MST: a point is a boundary point as long as it is next to a point from other classes in the MST



Comparing Global vs. Local Properties

- Boundaries can be simple locally but complex globally
  - These types of problems are relatively simple, but are characterized as complex by the measures



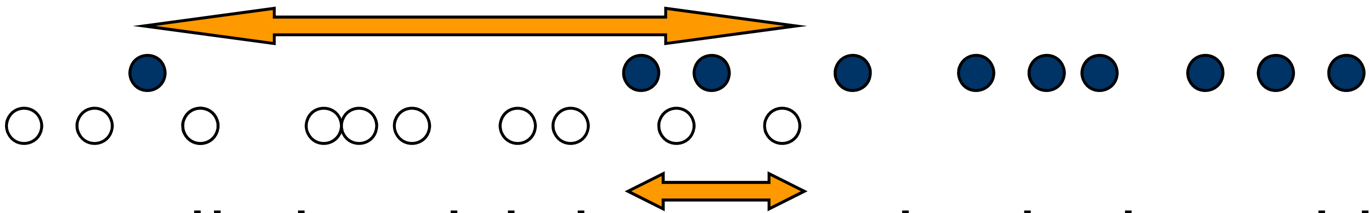
- Solution: complexity measure at different scales
  - This can be combined with different error levels

- Let $N_{i,k}$ be the k neighbors of the i-th point defined by, say, Euclidean distance. The complexity measure for data set D, error level ε, evaluated at scale *k* is

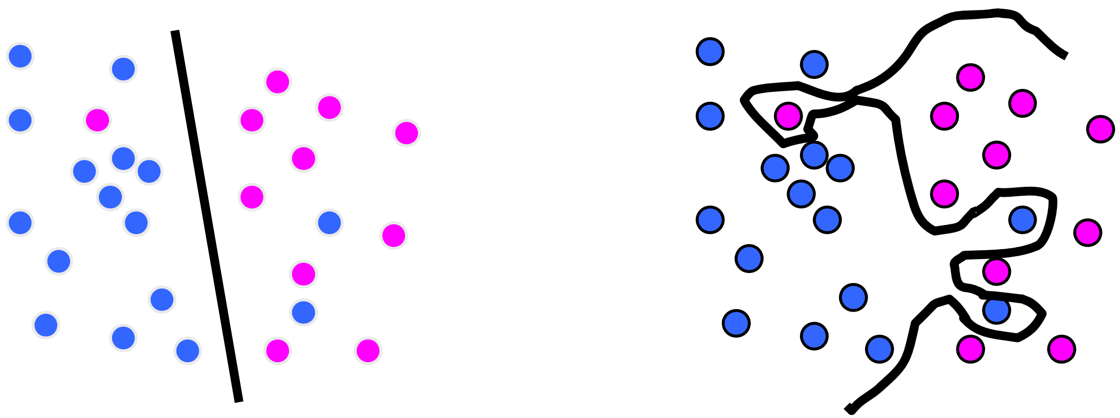
$$\bar{f}(D, \epsilon, k) = \frac{1}{n} \sum_{i=1}^{n} f(N_{i,k}, \epsilon)$$

Effects of Intrinsic Ambiguity

- The complexity measures can be severely affected when there exists intrinsic class ambiguity (or data mislabeling)
  - Example: FeatureOverlap (in 1D only)



- Cannot distinguish between intrinsic ambiguity or complex class decision boundary



Tackling Intrinsic Ambiguity

- Compute the complexity measure at different error levels
  - f(D): a complexity measure on the data set D
  - D^*: a "perturbed" version of D, so that some points are relabeled
  - $h(D, D^*)$: a distance measure between D and D^* (error level)
  - The new complexity measure is defined as a curve:

$$g(D, \epsilon) = \min_{D^*: h(D, D^*) \leq \epsilon} f(D^*)$$

  - The curve can be summarized by, say, area under curve

- Minimization by greedy procedures
  - Discard erroneous points that decrease complexity by the most

Outline

- ✓ Motivation
- ✓ Class ambiguity, dimensionality and boundary complexity
- ✓ Measures of Geometric Complexity
- ✓ Domains of Competence of Classifiers
- ✓ Other studies
- ✓ Concluding Remarks

Some Advanced Topics III: Data Complexity

Summary

- Automatic classification is useful, but can be very difficult.
- We know the key steps and many promising methods.

  But we have not fully understood how they work, what else is needed.
- Difficulties are class ambiguity, geometric complexity, & sample sparsity.
- Measures for geometric complexity are useful to characterize classifier domains of competence.

Some Advanced Topics III: Data Complexity

Summary

- Better understanding of how data and classifiers interact can guide practice.

- Further progress in statistical and machine learning will need systematic, scientific evaluation of the algorithms with problems that are difficult for different reasons.

Summary

1. Introduction to Data Mining and Knowledge Discovery
2. Data Preparation
3. Introduction to Prediction, Classification, Clustering and Association
4. Data Mining - From the Top 10 Algorithms to the New Challenges
5. Introduction to Soft Computing. Focusing our attention in Fuzzy Logic and Evolutionary Computation
6. Soft Computing Techniques in Data Mining: Fuzzy Data Mining and Knowledge Extraction based on Evolutionary Learning
7. Genetic Fuzzy Systems: State of the Art and New Trends
8. Some Advanced Topics I: Classification with Imbalanced Data Sets
9. Some Advanced Topics II: Subgroup Discovery
10. Some advanced Topics III: Data Complexity
11. Final talk: How must I Do my Experimental Study? Design of Experiments in Data Mining/Computational Intelligence. Using Non-parametric Tests. Some Cases of Study.