On the Necessity of Dataset Characterization for Experimental Analysis

Towards artificial datasets

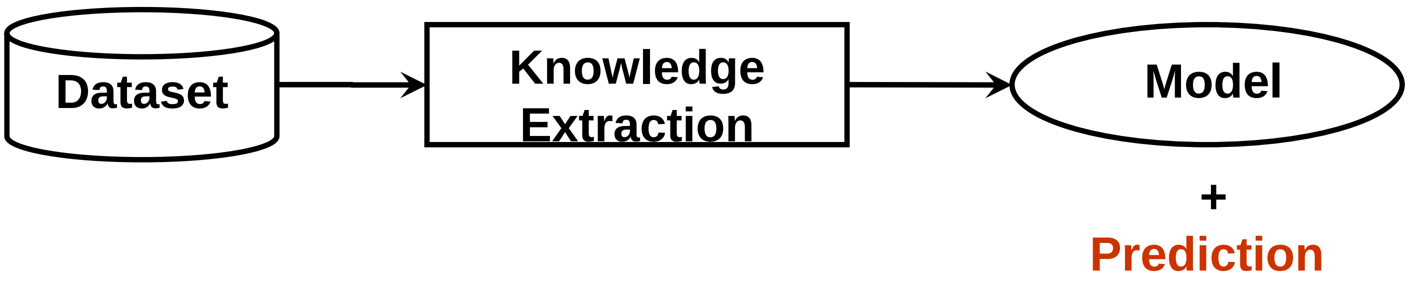
Núria Macià Antolínez
nmacia@salle.url.edu

Grup de Recerca en Sistemes Intel·ligents
Enginyeria i Arquitectura La Salle
Universitat Ramon Llull

LA SALLE
Universitat Ramon Llull

GRSI

Framework



- **Why this performance?**
  - Constraints of the method (classifier)
  - Complexity of the problem
- **Will other method be better suited to this problem?**
- **Is there an optimal learner?**

State of the Art

	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	-	96.1
sick	93.8	96.1	96.3	93.3	**98.4**	97.0	93.8	96.7
soyb	13.5	89.5	90.3	**92.8**	91.4	90.3	-	76.2
tao	49.8	**96.1**	96.0	80.8	95.1	93.6	83.6	88.4
thy	19.5	68.1	65.1	80.6	**92.1**	**92.1**	-	86.3
veh	25.1	69.4	69.7	46.2	73.6	72.6	-	72.2
vote	61.4	92.4	92.6	90.1	96.3	**96.5**	95.6	95.4
vow	9.1	99.1	**96.6**	65.3	80.7	78.3	-	87.6
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

Recent studies

- J. Demsar. Statistical comparisons of classifiers over multiple data sets. (JMLR06)
- J. Luengo, S. García, F. Herrera. A Study on the Use of Statistical Tests for Experimentation with Neural Networks. (IWANN07).

Literature Review

	JMLR	ML	PR	ICML
Total number of papers	380	230	665	393
Relevant paper for our study	85	20	45	56
Dataset selection				
Source of datasets				
Repositories	85	67	81	87
Synthetics	21	41	38	20
Specifics	14	15	0	0
Number of datasets				
1	0	2	10	0
(1,10]	57	60	72	68
(10,30]	29	30	36	21
>30	14	5	1	11
Number of instances				
(0,1000]	100	80	88	80
(1000,10000]	88	75	57	60
(10000,100000]	44	32	16	22
>100000	5	4	10	0
Number of attributes				
(0,10]	77	65	73	81
(10,25]	77	71	70	81
(25,100]	88	60	70	70
>100	15	8	12	11
Sampling method [%]				
cross validation, leave-one-out	82	80	72	75
Score function [%]				
Classification accuracy	75	84	88	83
ROC, AUC	18	15	4	12
Deviations, confidence intervals	35	41	25	38

- **A. Orriols et al. *On Data Set Characterization.* (In preparation)**

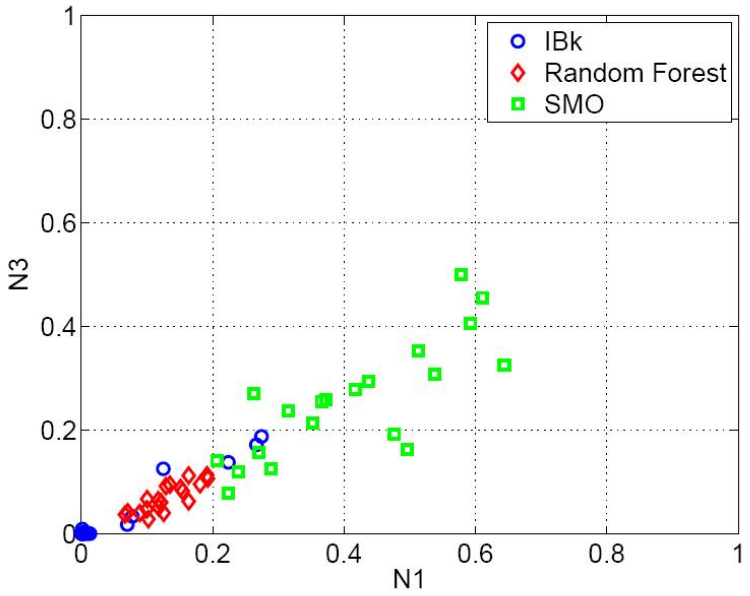
Is there a global learner?

- **Experimentation**
  - 3 collections of 20 data sets from the UCI Repository
  - IBk, Random Forest, and SMO from Weka
  - Bonferroni-Dunn test
- **Results**

	IBk	RF	SMO	Friedman
Collection 1	**1.33**	2.33	2.35	0.00015
Collection 2	2.33	**1.3**	2.38	0.00059
Collection 3	2.45	2.21	**1.34**	0.00038
All data sets	**2.03**	**1.96**	**2.02**	0.91430

Data Set Characterization

Which is the best and why?



IBk
- Instances of the different classes: lowly interleaved
- Instances of the same class: close in the feature space

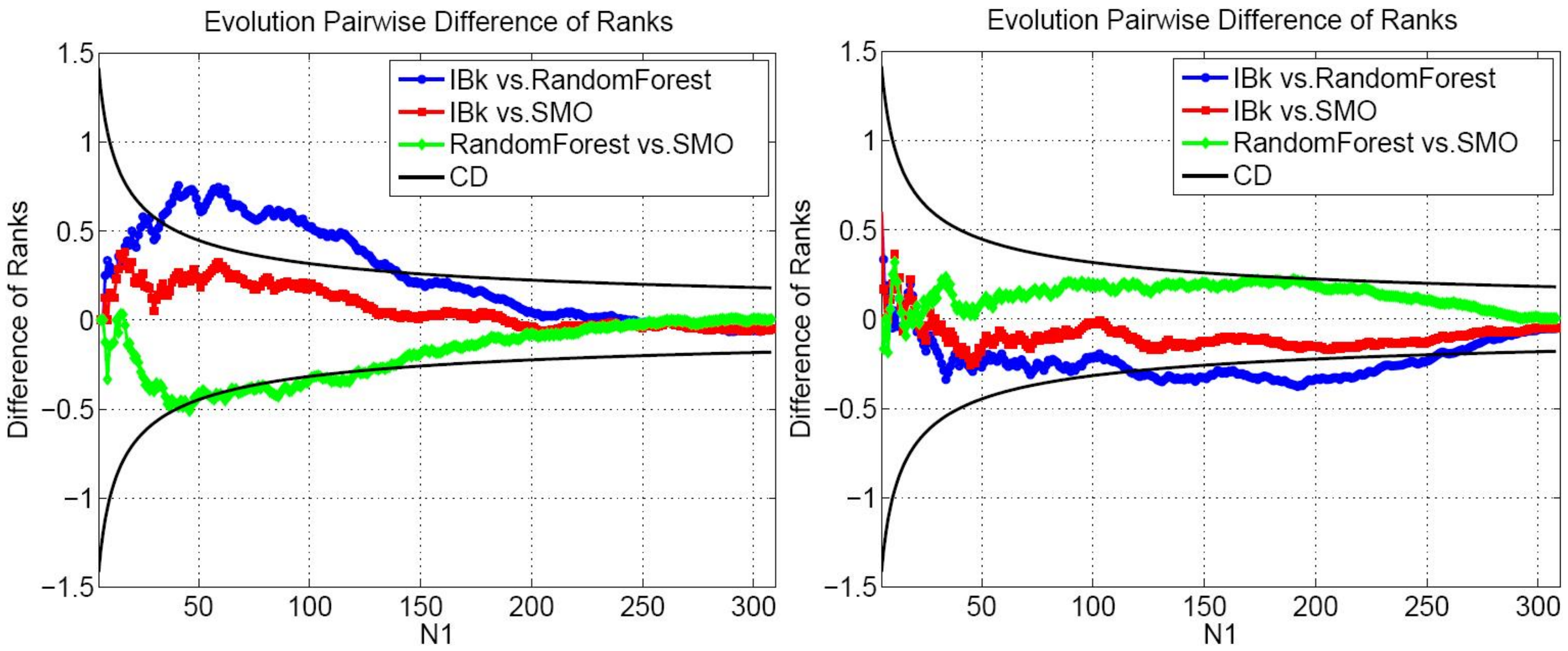
RF
- Instances of the different classes: High interleaved
- Instances of the same class: slightly disperse

SMO
- Half of the instances have lay in the class boundary.

Statistical Conclusions

Experimentation

- 300 datasets from the UCI Repository
- IBk, Random Forest, and SMO



Summary

- **My learner has outperformed another in 1000 data sets. Will it outperform it again in the 1001st data set?**
  - Much emphasis is being done on providing statistical analysis to the results
  - The selection of the datasets may not be representative enough
  - Claims on a best performer algorithm can be mislead by the current selection of datasets
  - If we use a high number of datasets, then all the algorithms perform the same

Synthetic Datasets (I)

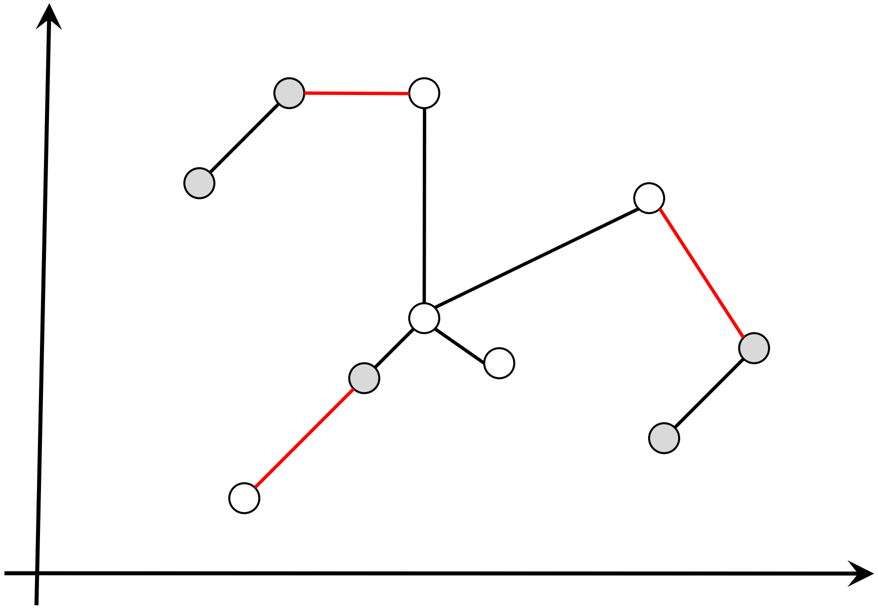
- **Use of synthetic datasets**
  - With known and controlled complexity
- **Two goals**
  - As benchmarking problems, whose complexity can be characterized. Problems can also be “grouped by types of underlying complexities”
  - Evaluation of current set of complexity metrics

Synthetic Datasets (II)

- **Synthetic datasets (ADS) can help to understand the complexity of real-world problems.**
- **Previous studies on data complexity found some limitations:**
  - Incomplete coverage of the measurement space
  - Unknown properties of data may not be characterized by the current set of metrics
  - Apparent complexity
    - **Synthetic dataset allow us to work on real complexities**
  - High correlation between some metrics (e.g. N1 and N2)
    - **Constraints in the experimental testbed**

Preliminary Study on ADS

- ADS were built based on boundary



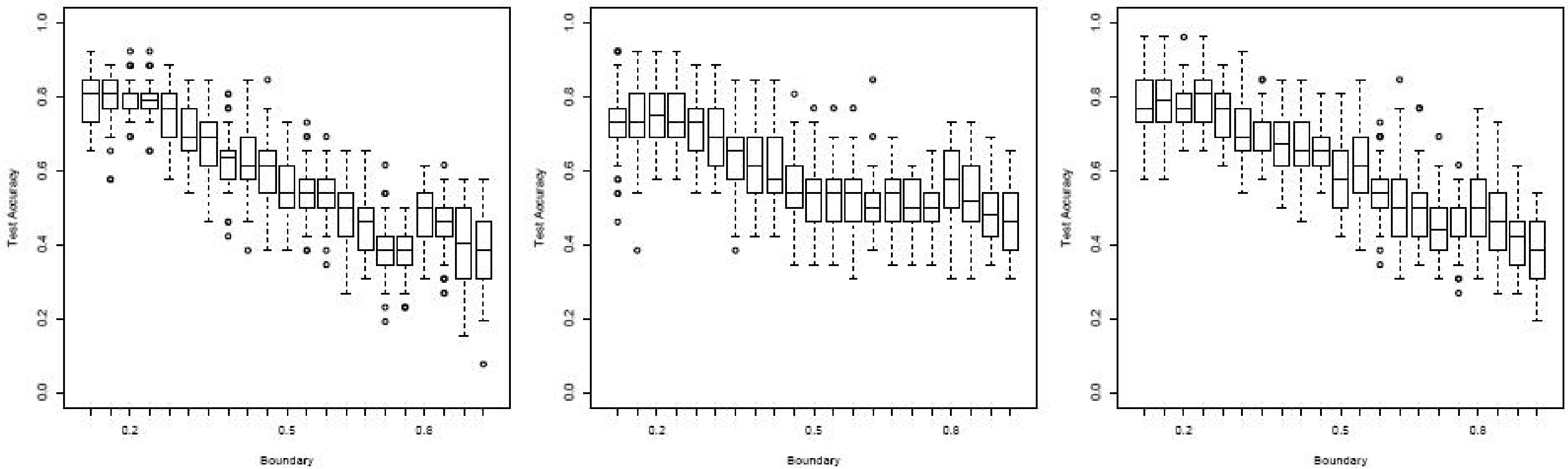
Preliminary Study on ADS

Generation procedure

- Set the number of instances *n*, the number of attributes *m*, and the length of the class boundary *b*
- Generate n points distributed randomly
- Build the MST
- Label the point to obtain the required boundary

Experiment and Results (I)

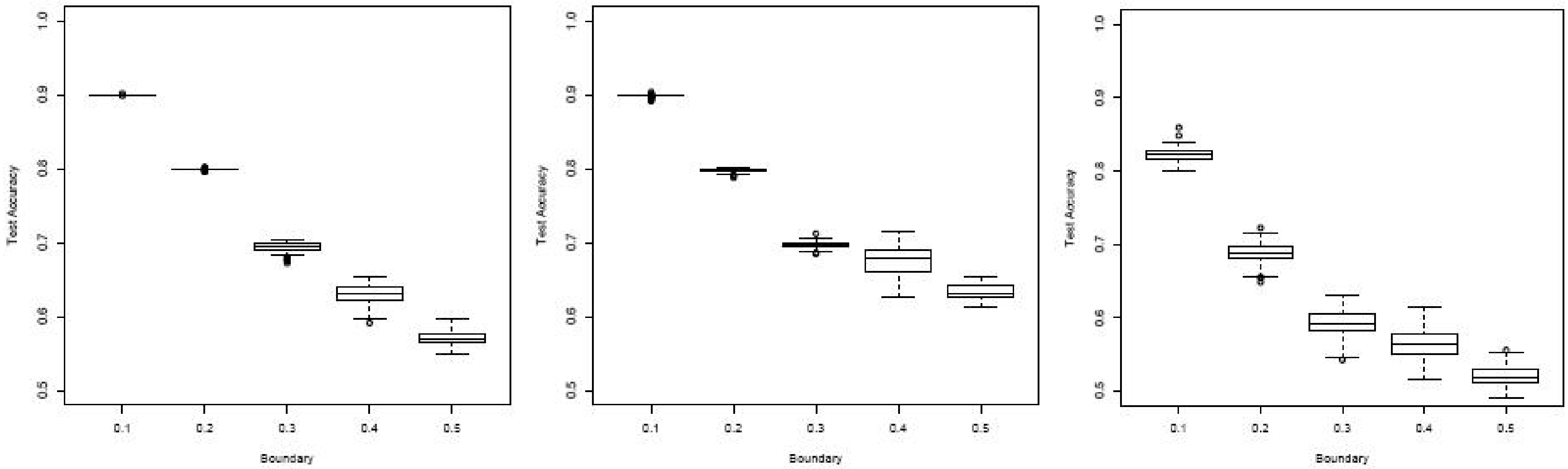
- 1150 artificial two-class problems (50 datasets for each complexity level)
- n=26, m=5



(a) Naive Bayes (b) PART (c) Random Tree

Experiment and Results (II)

- 500 artificial two-class problems (100 for each complexity level {0.1, 0.2, 0.3, 0.4, 0.5})
- n=1001, m=10

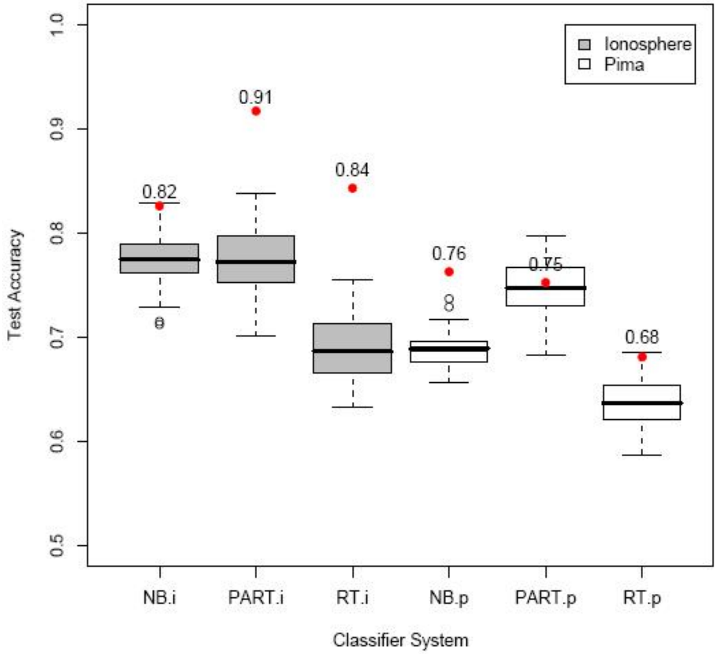


(a) Naive Bayes (b) PART (c) Random Tree

Experiment and Results (III)

Comparison of ADS with UCI Datasets

- Pima and Inosphere from the UCI Repository
- 100 datasets from each problem with the same characteristics
- Minimum accuracy bound



Limitations

- **Experiment (I)**
  - The behavior is somehow irregular from complexities greater than 0.8
  - Certain variability of the classifier's accuracy
- **Experiment (II)**
  - Similar structures due to the labeling of points at the extremes of the MST

On the Necessity of Dataset Characterization for Experimental Analysis

Towards artificial datasets

Núria Macià Antolínez
nmacia@salle.url.edu

Grup de Recerca en Sistemes Intel·ligents
Enginyeria i Arquitectura La Salle
Universitat Ramon Llull