

# On the dimensions of data complexity through synthetic data sets

Onzè Congrés Internacional de l'Associació Catalana d'Intel·ligència Artificial

---

Núria Macià  
Ester Bernadó-Mansilla  
Albert Orriols-Puig  
{nmacia,esterb,aorriols}@salle.url.edu

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



# Motivation (I)

---

- **Competitive learners that extract accurate models from data have been developed**
- **Which is the best technique?**
- **My learner has outperformed another in 1000 data sets. Will it outperform it again in the 1001<sup>st</sup> data set?**

**Toward:**

**Data characterization to know which and why a method is the best for a given problem**

# Motivation (II)

Dataset	#Attr	#Inst	C4.5	IB3	PART	SMO	My approach
Abalone	8	4177	0.9998	0.9998	0.9998	0.9998	<b>0.9999</b>
Balance Scale	4	625	0.8467	0.8866	0.8625	0.9297	<b>0.9999</b>
Breast Cancer Wisconsin	30	569	0.9367	0.9719	0.9332	0.9771	<b>0.9999</b>
Chess (King-Rook vs. King)	6	28056	0.9791	0.9010	0.9975	0.9003	<b>0.9999</b>
Glass Identification	9	214	0.8071	0.8355	0.8407	0.7104	<b>0.9999</b>
Heart Disease	13	303	0.8037	0.7963	0.7444	0.8407	<b>0.9999</b>
Hepatitis	19	155	0.8008	0.7996	0.8020	0.8804	<b>0.9999</b>
Ionosphere	34	351	0.9120	0.8518	0.9118	0.8776	<b>0.9999</b>
Lenses	4	24	0.8167	0.9333	0.8833	0.8000	<b>0.9999</b>
Letter Recognition	16	20000	0.9960	0.9994	0.9965	0.9916	<b>0.9999</b>
Lung Cancer	56	32	0.7583	0.7333	0.7583	0.6500	<b>0.9999</b>
Optical Recognition	64	5620	0.9939	0.9996	0.9957	0.9977	<b>0.9999</b>
Pima Indians Diabetes	8	768	0.7566	0.7382	0.7161	0.7669	<b>0.9999</b>
Statlog (Image Segmentation)	19	2310	0.9939	0.9952	0.9931	0.9965	<b>0.9999</b>
Statlog (Vehicle Silhouettes)	18	846	0.7695	0.7731	0.7766	0.7494	<b>0.9999</b>
Thyroid Disease	5	215	0.9344	0.9485	0.9299	0.7907	<b>0.9999</b>
Waveform Database Generator	21	5000	0.8290	0.8496	0.8360	0.8588	<b>0.9999</b>
Wine	13	178	0.9604	0.9833	0.9604	0.9889	<b>0.9999</b>
Yeast	8	1484	0.7223	0.7069	0.7156	0.6880	<b>0.9999</b>
One more problem	?	?	?	?	?	?	?

- **The purpose of this work is to:**
  - Analyze the impact of different data sets characteristics
  - Consider two frequent measures
    - **Number of attributes**
    - **Number of instances**
  - Consider a measure about class geometry
    - **Length of the class boundary**

# Outline

---

- 1. Approach**
- 2. Complexity dimensions**
- 3. Synthetic data sets**
- 4. Design of experiments**
- 5. Results**
- 6. Conclusions**
- 7. Further work**

# 1. Approach

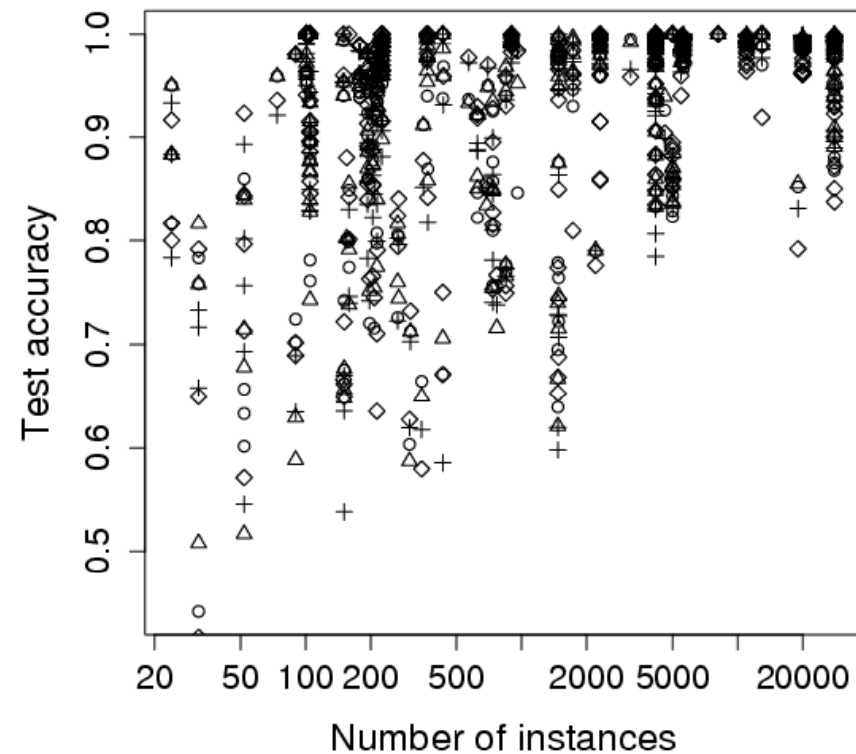
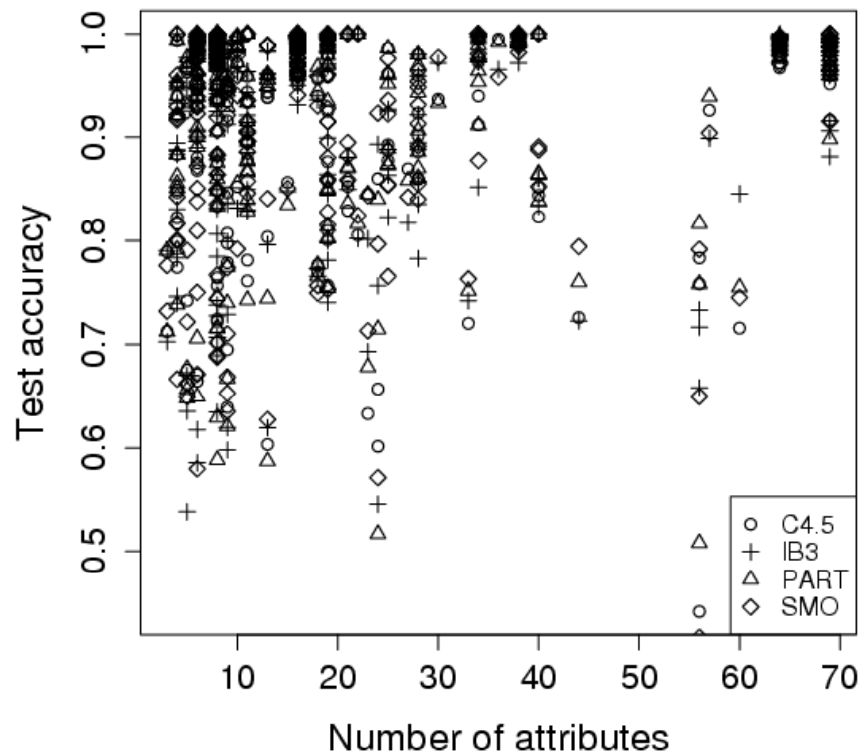
---

- **Study how different dimensions influence classifier behavior**
  - Number of attributes
  - Number of instances
  - Length of the class boundary
- **Through synthetic data sets**
  - They provide a controlled scenario with known characteristics

# 2. Complexity dimensions

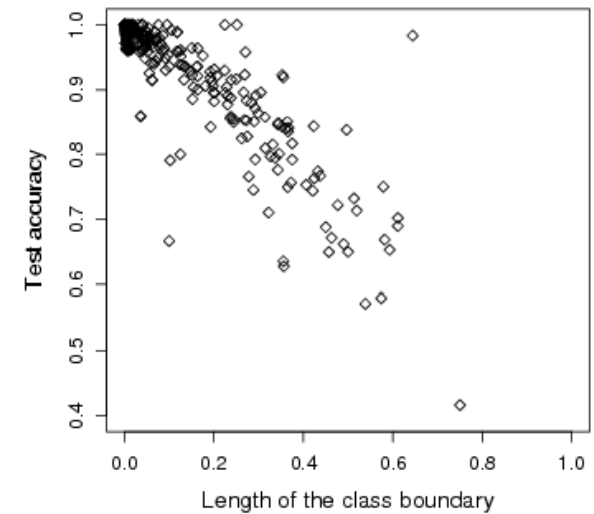
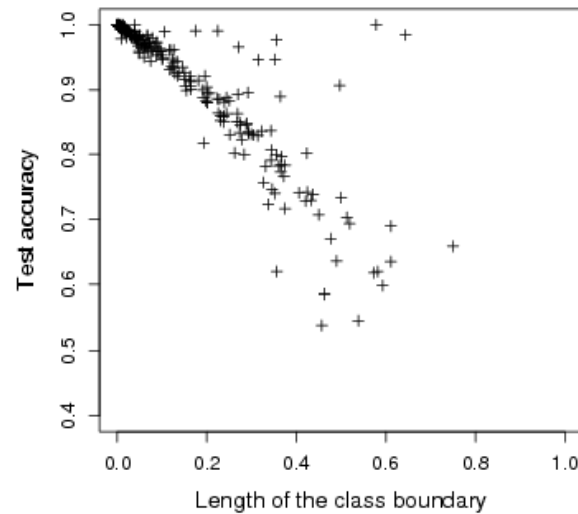
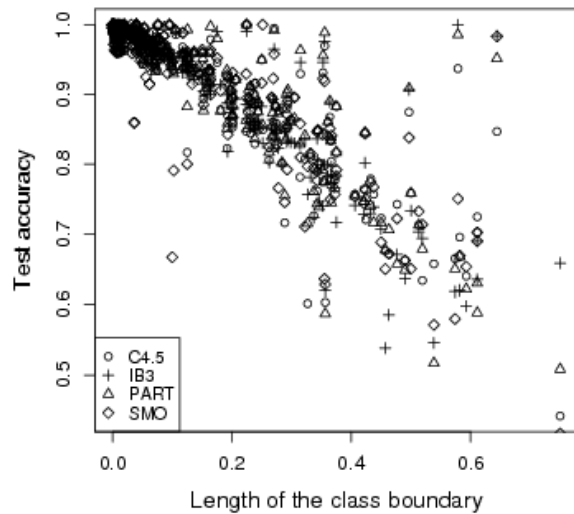
## □ Number of attributes and instances

- No relation is observed among these characteristics and classifiers' accuracy



# 2. Complexity dimensions

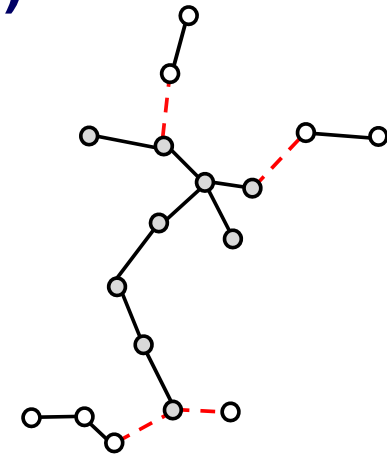
- Length of the class boundary
  - Linear correlation with classifiers' accuracy



# 2. Complexity dimensions

## □ Length of the class boundary

- It estimates the density of points located near the class boundary
- To compute the measure:
  - **Build the minimum spanning tree (MST) connecting all the points regardless of class**
  - **Count the number of edges joining opposite classes and divide this value by the total number of points**



# 3. Synthetic data sets

---

- **Real-world problems**
  - Not cover all the complexity spectrum
  - Are characterized by uncertainty, missing values, class imbalance...
  
- **Synthetic data sets**
  - Provide a controlled framework
  - Permit varying different dimensions independently

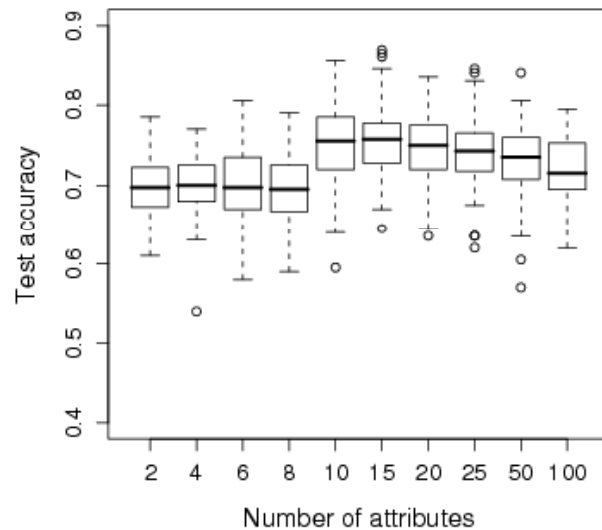


# 4. Design of experiments

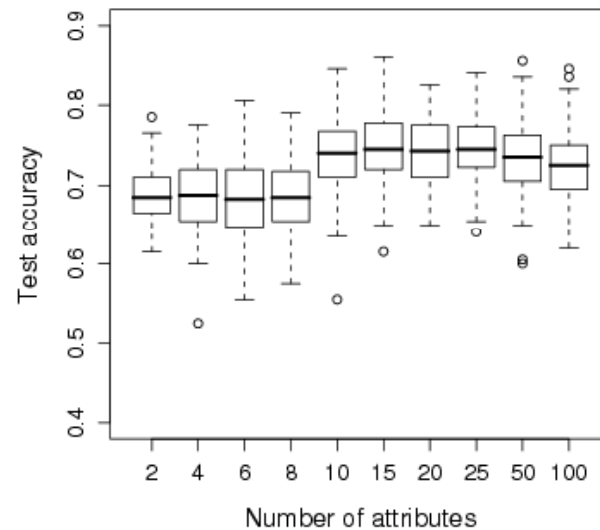
---

- **Three different paradigms of learners:**
  - Tree induction – C4.5
  - Rule learning – PART
  - Support vector machine – SMO
- **10-fold cross-validation**
- **1000 synthetic data sets**

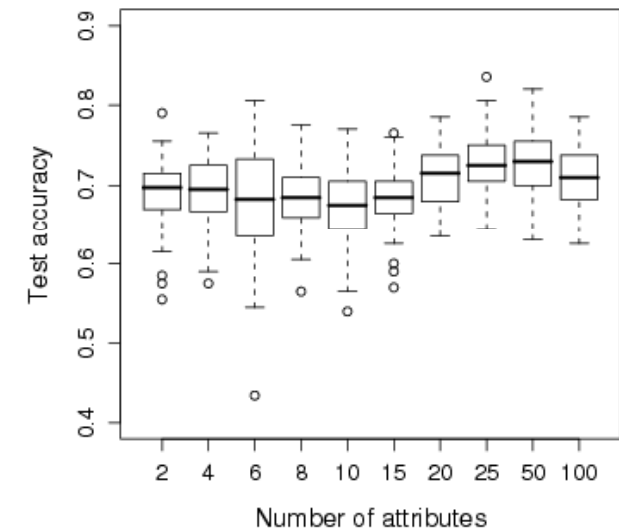
# 5. Results (I)



(a) C4.5



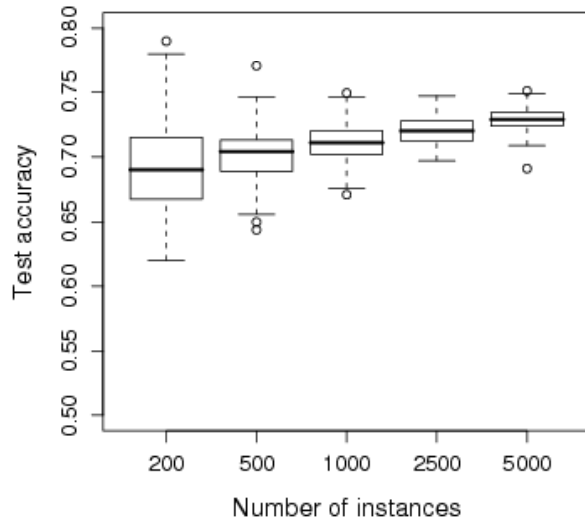
(b) PART



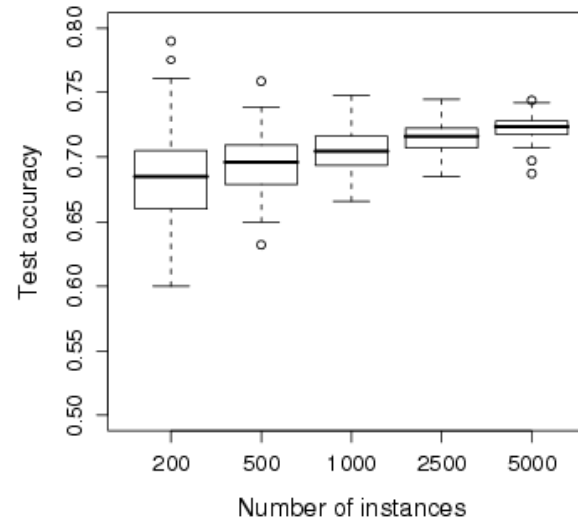
(c) SMO

- $n = 200, b = 0.3$
- Classifiers behave similarly
- Accuracy rates range in the interval  $[0.6-0.8]: 1 - b$
- Variability indicates that other dimensions are required to describe the data complexity

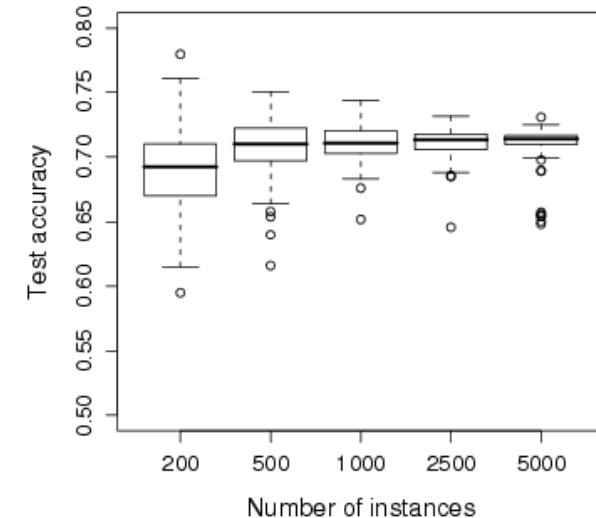
# 5. Results (II)



(a) C4.5



(b) PART



(c) SMO

- $m = 2, b = 0.3$
- Classifiers behave similarly
- Spread of accuracy rates decreases with increasing number of instances

# 6. Conclusions

---

- **Information provided by the number of attributes and the number of instances can be embedded in the measure of length of the class boundary**
- **Length of the class boundary is a relevant estimate of classifier accuracy, but it is not enough to describe the whole data set complexity**
- **Using synthetic data sets along the experiments allows us to vary different dimensions independently and work under a controlled scenario**

# 7. Further work

---

- **Synthetic data sets do not contain similar structures to those of real-world problems**
  - Generate data sets that follow physical processes
- **Study other complexity dimensions**

# On the dimensions of data complexity through synthetic data sets

Onzè Congrés Internacional de l'Associació Catalana d'Intel·ligència Artificial

---

Núria Macià  
Ester Bernadó-Mansilla  
Albert Orriols-Puig  
{nmacia,esterb,aorriols}@salle.url.edu

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

