

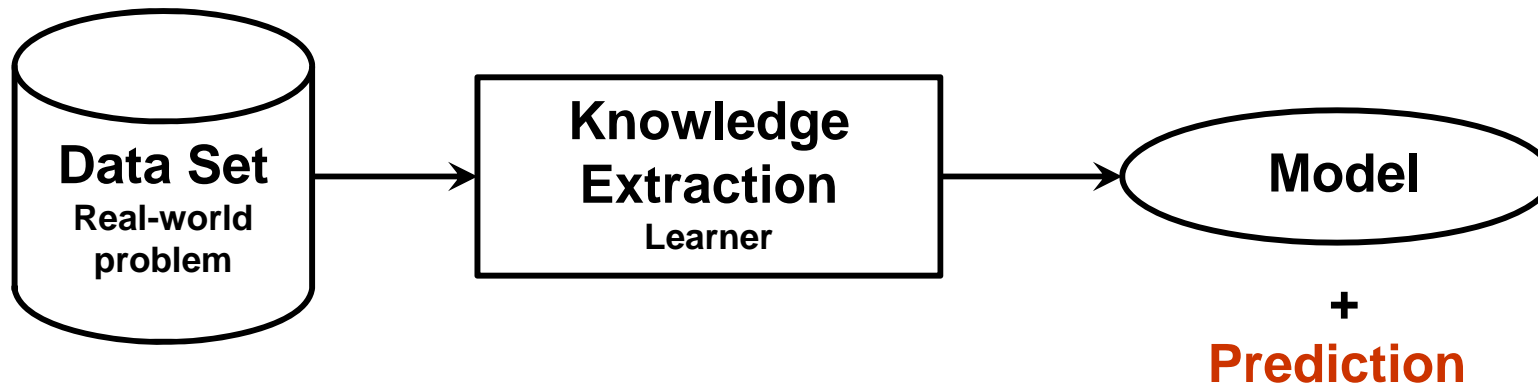
Genetic-based Synthetic Data Sets for the Analysis of Classifiers Behavior

8th International Conference on Hybrid Intelligent Systems

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

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





- **Necessity of synthetic data sets**
 - To evaluate real learners performance under controlled scenarios
- **How to generate synthetic data sets?**
 - Data complexity (Ho & Basu, 2002)
 - Length of the class boundary (Macià et al., 2008)

Objective: **Set of benchmark problems to analyze learners behavior**

Outline

- 1. Data complexity**
- 2. Synthetic data sets**
- 3. Design of GA**
- 4. Experiments and results**
- 5. Conclusions and further work**

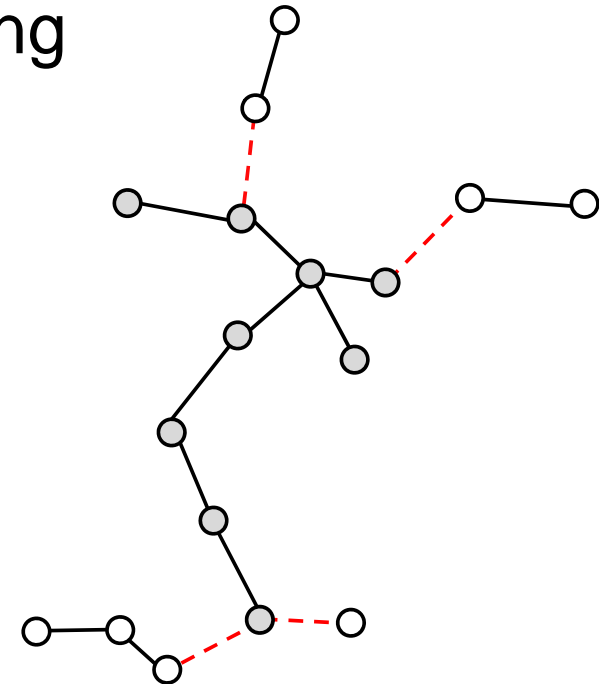
1. Data complexity

□ Length of the class boundary

- Build minimum spanning tree (MST) connecting all the points regardless of class
- Count the number of edges joining opposite classes

□ Two cases of many points in boundary:

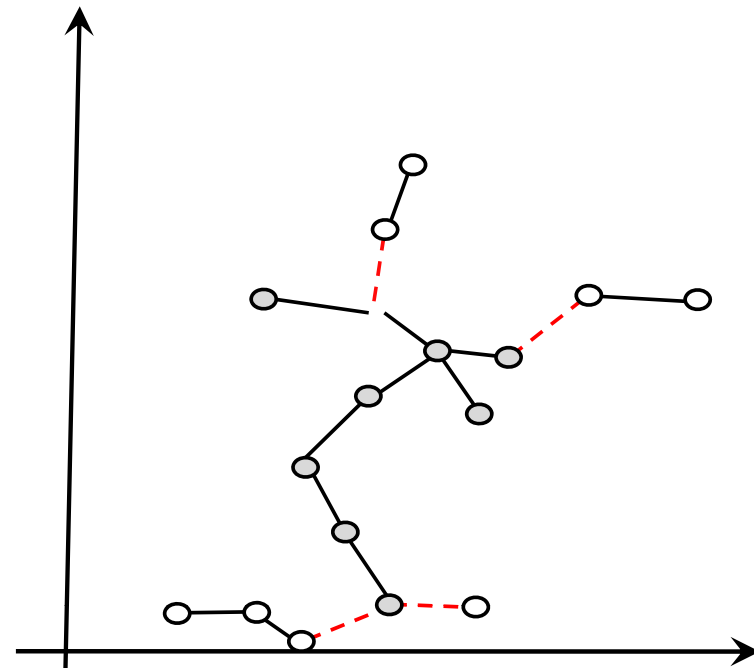
- Very interleaved or random data
- Linearly separable problem with narrow margins



2. Synthetic data sets

□ 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 and build the MST.
- Label the class of each instances



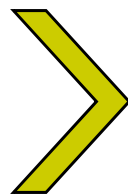
2. Synthetic data sets

- **Exhaustive search**

- Labelings grow exponentially with the number of instances

- **Heuristic search**

- Demanded length of the class boundary is not always achieved
- No diverse solutions



Genetic algorithm

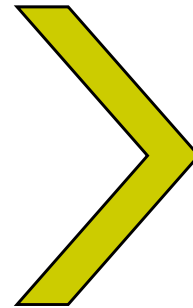
3. Design of GA

□ Knowledge representation

- k -ary string where the bit i stores the class label of the i th instance

Data set i

Att. 1	Att. 2	...	Att. N	Class
0.4	0.5		0.4	0
0.2	1.0		0.2	1
0.5	0.3		0.4	1
0.6	0.5		0.4	0
0.7	0.1		1.0	1
0.5	0.3		0.9	1



Individual i

0	1	1	0	1	1
---	---	---	---	---	---

3. Design of GA

□ Genetic operators

- s-wise tournament selection
- Two-point crossover
- Bit-wise mutation

□ Fitness function

- $fitness_i = |b_{obj} - b_i|$

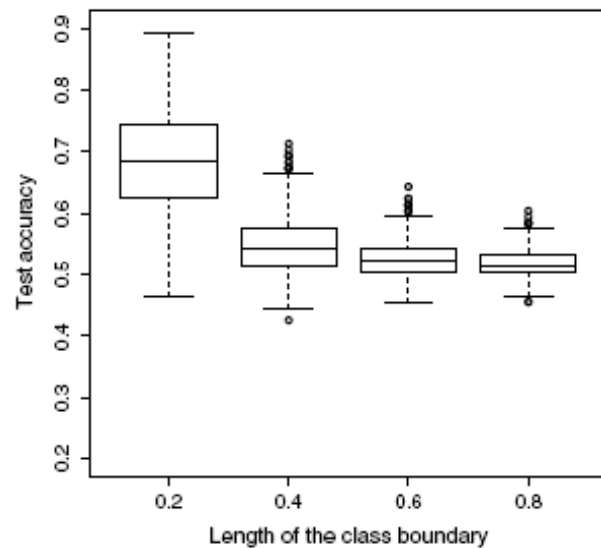
4. Experiment and results (I)

- **Synthetic data set generation**
 - Different solutions < Solutions
 - Population converge to the same solution
 - {0100,1011} are equivalent individuals
 - Intermediate complexity are obtained in early generations

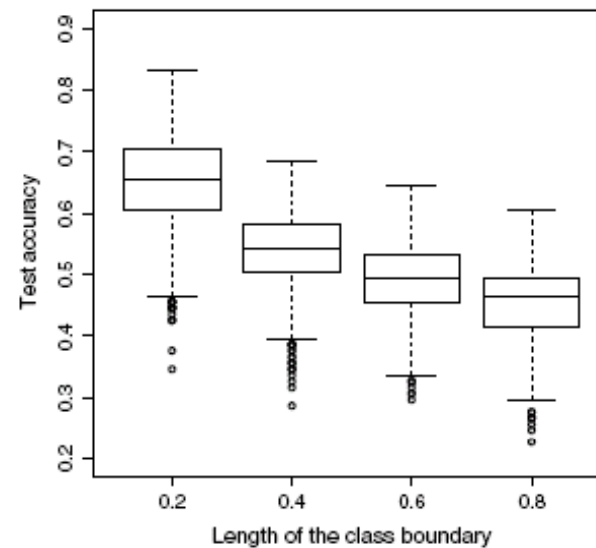
# Instances	Population size	Runs		0.2	0.4	0.6	0.8
101	500	50	First solution in generation	22.82	0	0	22.24
			Solutions (%)	10.43	18.73	18.89	10.67
			Different solutions (%)	9.04	16.93	17.29	9.26
501	1000	10	First solution in generation	85.5	7.7	8.2	85.4
			Solutions (%)	9.41	15.05	14.24	10.82
			Different solutions (%)	8.46	13.90	13.06	9.73

4. Experiment and results (II)

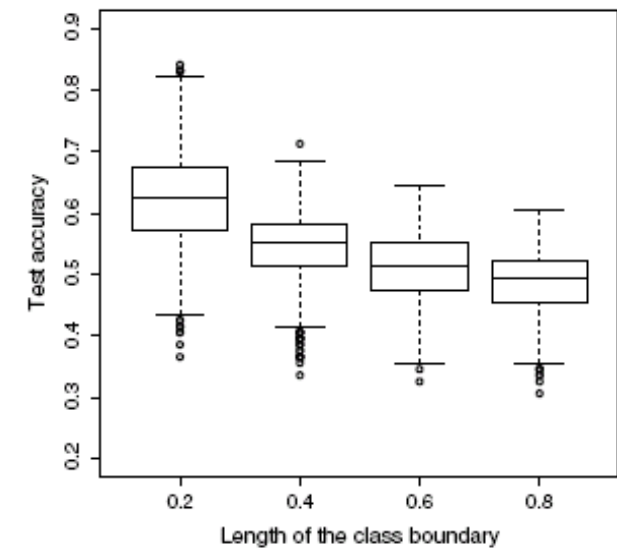
- **Analysis of classifiers behavior**
 - Three different paradigms: C4.5, Naïve Bayes, and SMO
 - Similar accuracy rates with noticeable variability



(a) C4.5



(b) Naïve Bayes



(c) SMO

5. Conclusions

- **The GA allows us to generate data sets with the demanded length of the class boundary**

6. Further work

- **Efficiency and scalability**
 - Move from simple GA to competent GA
- **Capacity of satisfying multiple criteria**
 - Multi-objective strategy
- **Achieve structure of real-world problems**
 - Provide a set of benchmark problems

Genetic-based Synthetic Data Sets for the Analysis of Classifiers Behavior

8th International Conference on Hybrid Intelligent Systems

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

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

