Some Results on the use of UCS in Imbalanced Domains

Albert Orriols Puig
aorriols@salle.url.edu

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

Outline

1. Why do we care about mining rarities?
2. The UCS Classifier System
3. Focusing the problem: Facet-wise analysis
4. Results on imbalanced data
5. Rebalancing the imbalanced data
6. Conclusions and further work

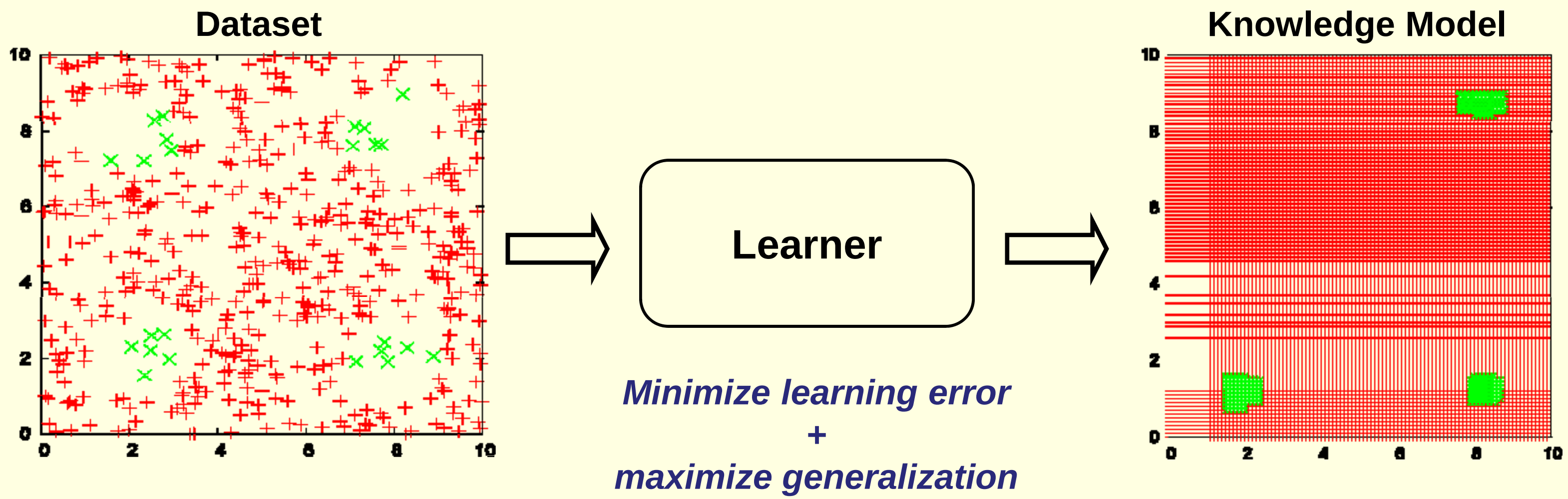
Is it usual to mine rarities?

- **The most interesting and novel knowledge tends to reside in the rarities**
- **Rarities in the classification realm:**
  - **E.g: Detection of infrequent diseases**
  - **Imbalanced datasets: Few ill patients and a high amount of healthy patients (luckily)**
- **Imbalanced real-world domains:**
  - Fraudulent credit card transactions **(Chan, P.K. and Stolfo, S.J., 1998)**
  - Learning word pronunciation **(Van den Bosch, A. et al, 1997)**
  - Prediction of pre-term births **(Grzymala-Busse, J.W. et al., 2000)**
  - Prediction of telecommunications equipment failures **(Grzymala-Busse, J.W. et al., 2000)**
  - Detection oil spills from satellite images **(Kubat, M. Et al., 1998)**
  - Detection of Melanomas

Should we care about rarities?

- Typically, standard learners appear to be biased toward the majority class.
- Classifiers attempt to reduce global quantities as error rate
- Result:
  - Examples of the overwhelmed class well-classified.
  - Examples of the minority class bad-classified.
- This problem has been hidden by some tricky metrics:
  - I reached a performance rate of 99% in test. I'm a genius!
    - **What is the distribution of the training dataset?**
    - Performance per class?
    - Has my classifier generalized correctly?
    - What is the purpose of my model?

Should we care about rarities?



Outline

1. Why do we care about mining rarities?
2. The UCS Classifier System
3. Focusing the problem: Facet-wise analysis
4. Results on imbalanced data
5. Rebalancing the imbalanced data
6. Conclusions and further work

Learning Classifier Systems

(Holland, 1976; Rolland & Reitman, 1978)

Rule Evaluation Procedure

Typically, Reinforcement Learning (Sutton & Barto, 1998)

Population

Classif. 1
Cassif. 2
Classif. 3
...
Classif. n

if *condition* **then** *action* **with** error ε

Knowledge discovery

Typically GA (Holland,1989)

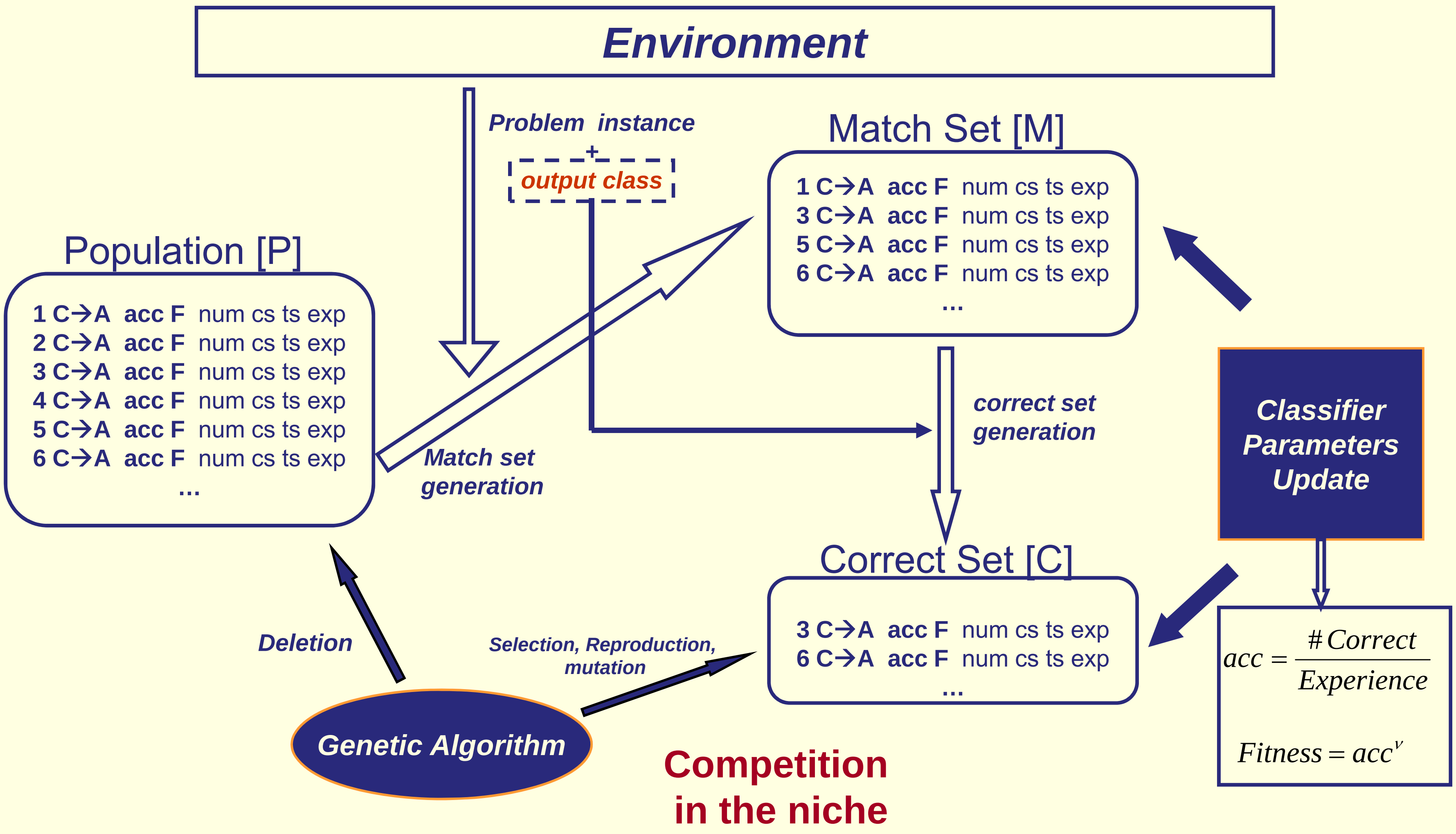
- **XCS** **(Wilson,1995 & 1998)**
- **UCS** **(Bernadó & Garrell, 2003)**

Three main pilars:

- **Rule-based systems**
- **Rule evaluation**
- **Knowledge discovery**

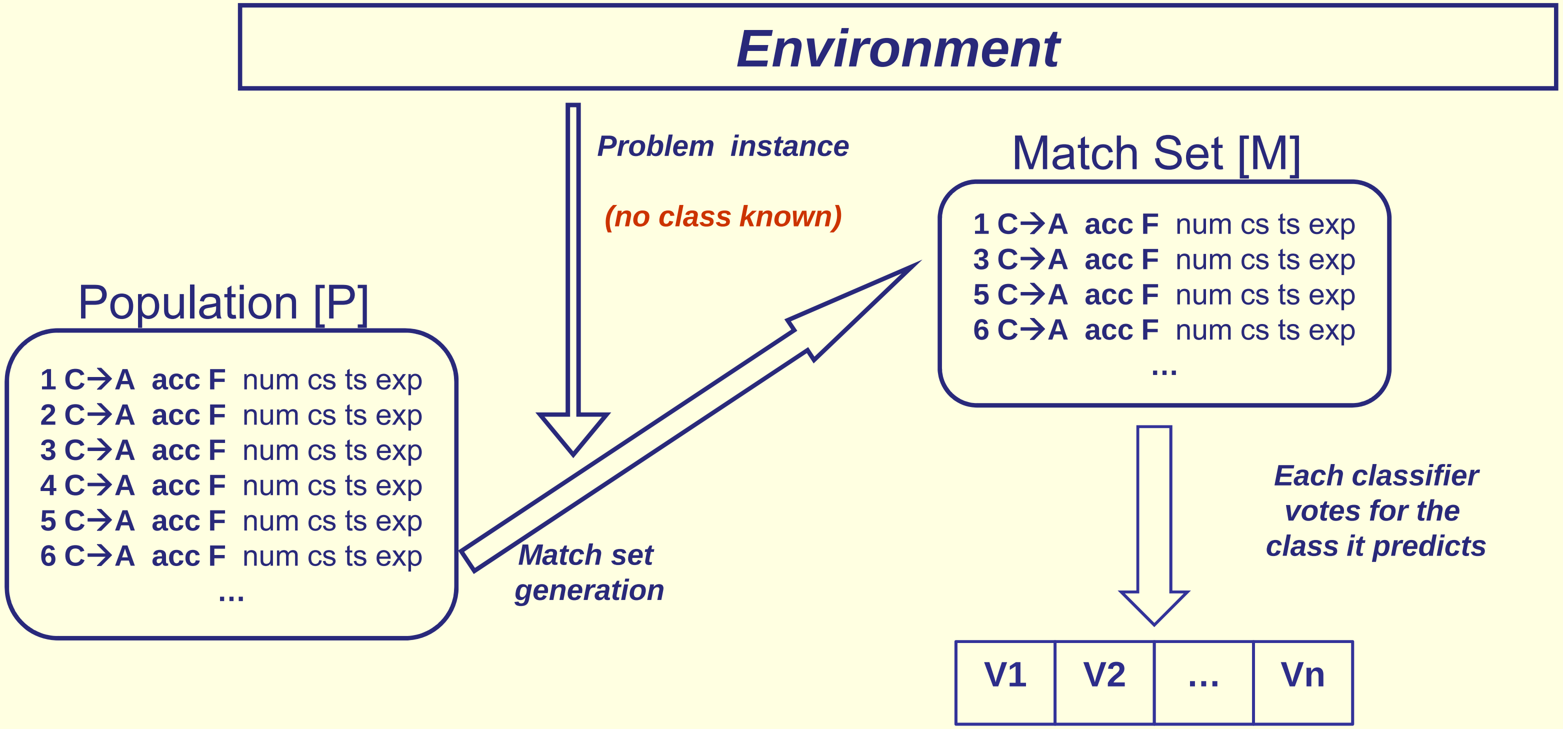
Description of UCS

- **In training mode**



Description of UCS

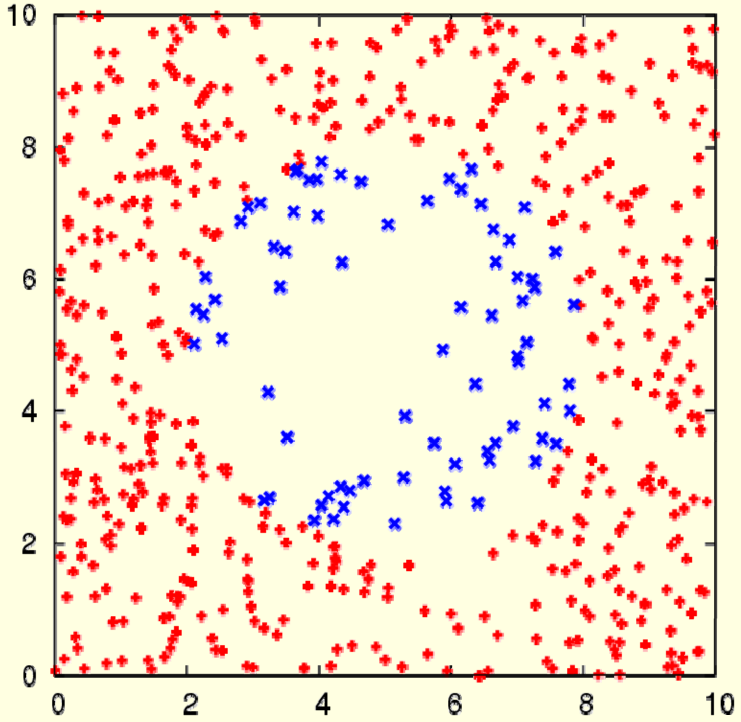
- **In test mode**



Classifier/niches collaboration to cover all the feature space and define class boundaries

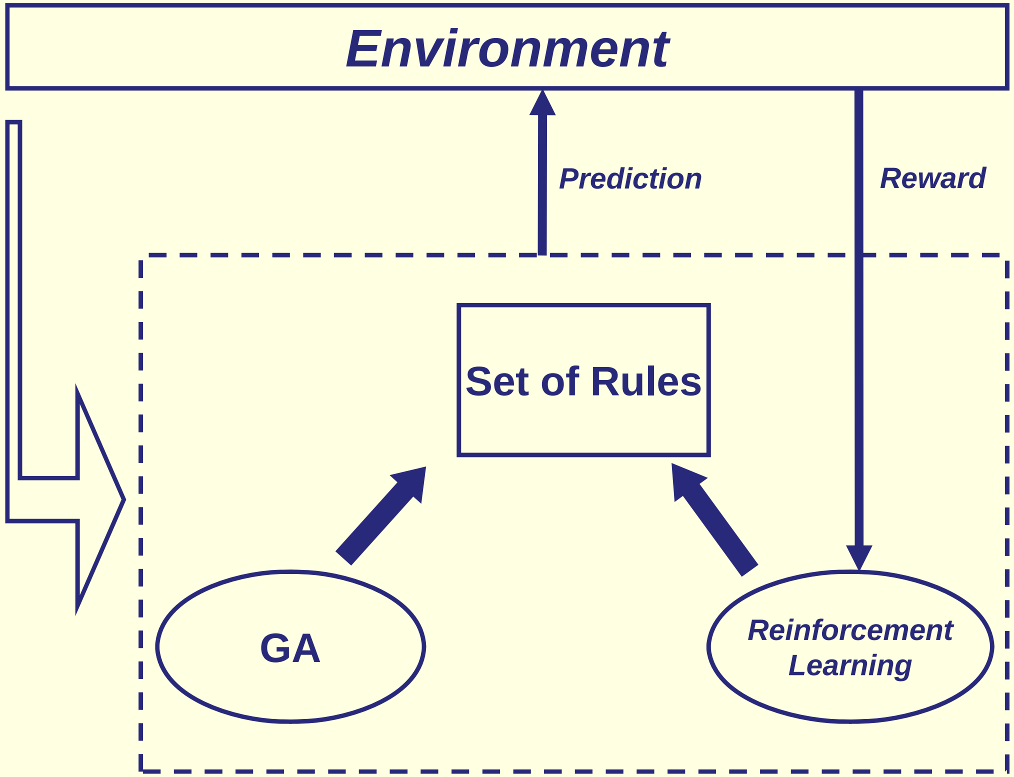
Description of UCS

Learning domain



Imbalance ratio* = Ratio of num. inst. majority class to num. inst. Minority class = *525:75

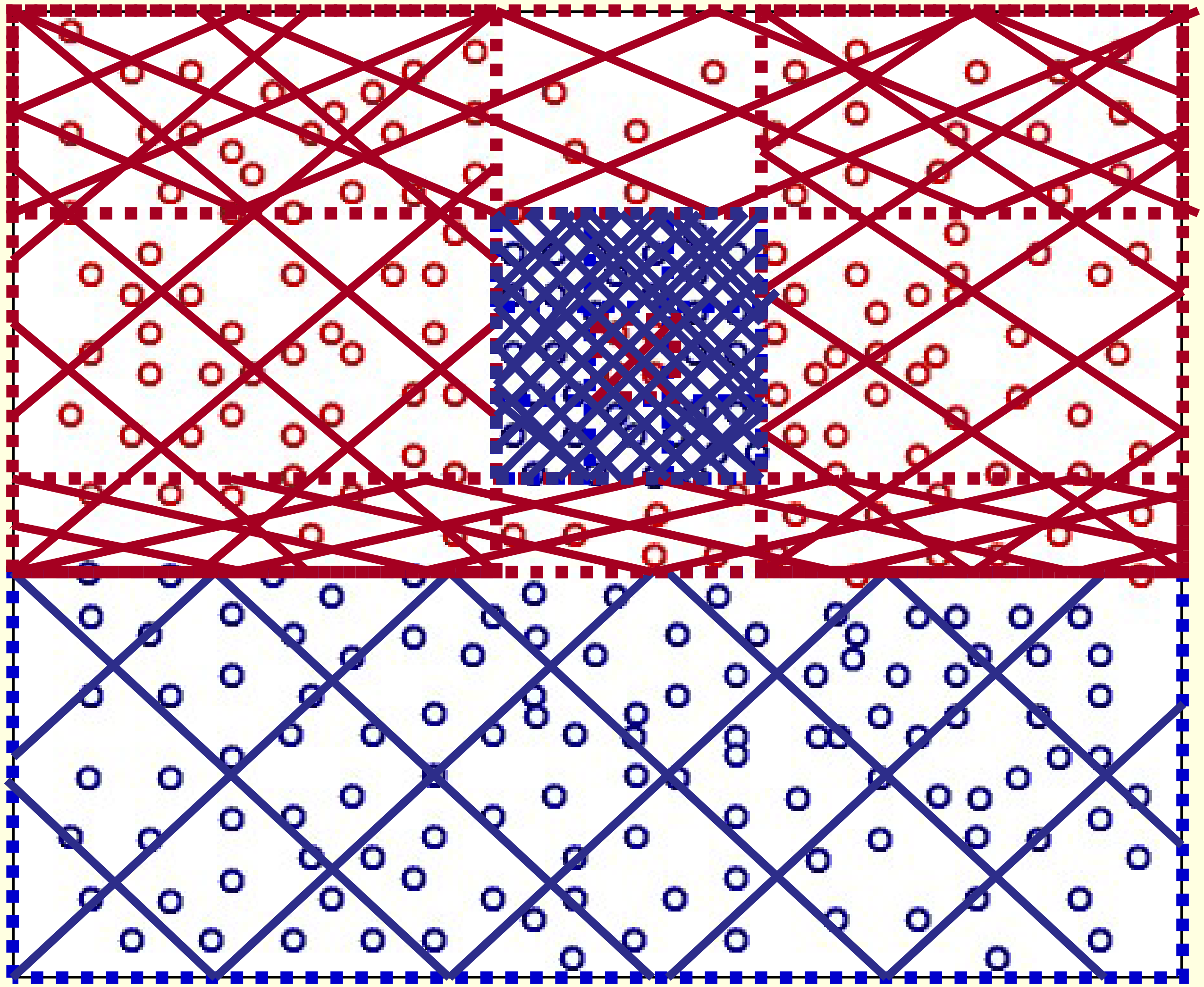
- **1 minority class example**
- **7 majority class examples**



Outline

1. Why do we care about mining rarities?
2. The UCS Classifier System
3. Focusing the problem: Facet-wise analysis
4. Results on imbalanced data
5. Rebalancing the imbalanced data
6. Conclusions and further work

Focusing the problem



Small Disjunct **or**
Starved niche

Again
more small disjuncts

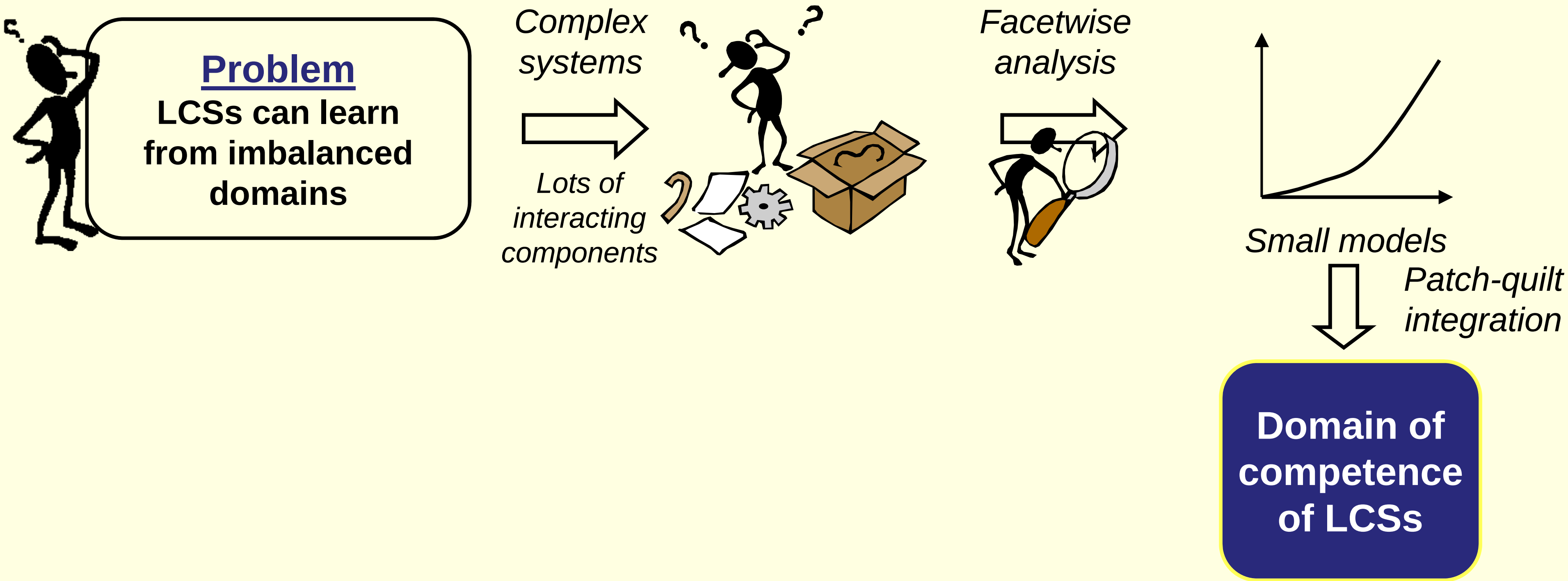
Overgeneral Classifier

How to face this problem?

- **First approach: Re-sampling techniques**
  - Problem: few instances of the minority class
  - Solution: resample to balance the data set
  - Advantages:
    - Many authors have shown that learning improves in some limited domains **(Holmes, 1998; Orriols & Bernadó 2005; Batista, 2004)**
  - Disadvantages:
    - Changing the learning domain with poor control
    - And still, we have not found the problems, limitations, and virtues of our LCSs.

How to face this problem?

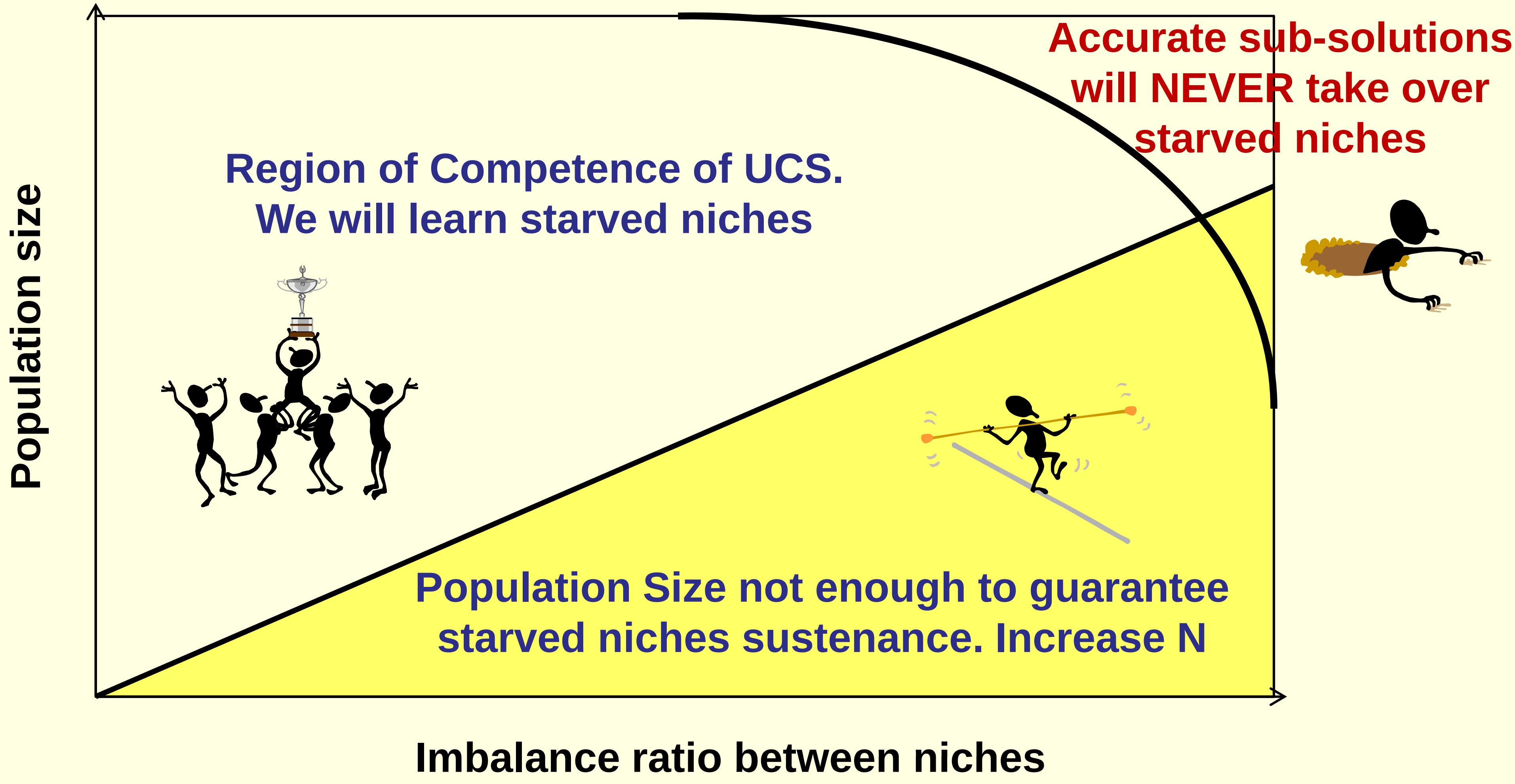
- **Second approach: Modeling principle**
  - Goldberg's way:
    - Understand what your system is capable of
    - Then, design approaches to improve it



Facet-wise Analysis

- Conditions to obtain classifiers that represent starved niches
- Estimation of the parameters of over-general classifiers
- The effect of the GA frequency on different niches
- Take-over time of starved niches

Patch-quilt integration



- Crucial for better understanding of the system
- Providing insights on how the system must be configured

see (Orriols et al, 2008)
(to be submitted in brief)

A brief example: The imbalanced multiplexer

(11-bit) Multiplexer

Selection bits

Position bits

- Example: **000 10010100:1**
- Complexity related to the number of address bits
- Completely balanced
- UCS should evolve:

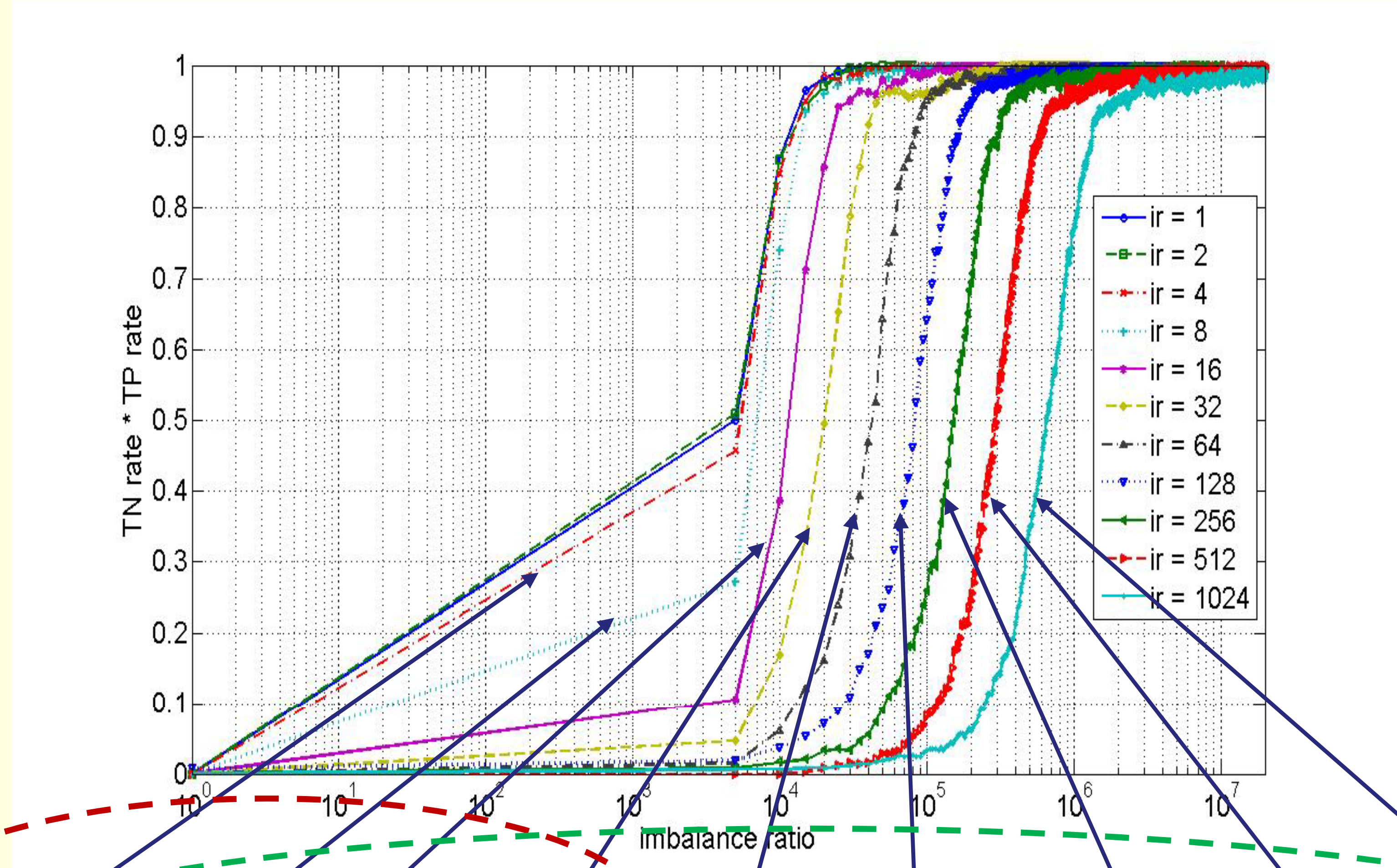
000 0#######:0	000 1#######:1
001 #0######:0	001 #1######:1
010 ##0#####:0	010 ##1#####:1
011 ###0####:0	011 ###1####:1
100 ####0###:0	100 ####1###:1
101 #####0##:0	101 #####1##:1
110 ######0#:0	110 ######1#:1
111 #######0:0	111 #######1:1

Imbalanced Multiplexer

•**We under-sampled class 1**

•***ir:*** Proportion between majority and minority class instances

A brief example: The imbalanced multiplexer



ir = 16:1 *ir = 16:1* *ir = 16:1* *ir = 32:1* *ir = 64:1* *ir = 128:1* *ir = 256:1* *ir = 512:1* *ir = 1024:1*

Problems solved with first-order XCS before the analysis

Problems solved following the guidelines provided by the facet-wise anlaysis

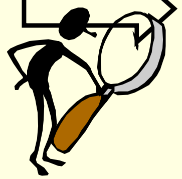
What’s next?

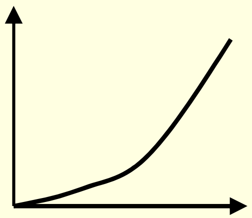
Problem
LCSs can learn from imbalanced domains

Complex systems

Lots of interacting components

Facetwise analysis





Small models

Competence domain of LCSs

Aplication of LCSs to a new real-life problem

Outline

1. Why do we care about mining rarities?
2. The UCS Classifier System
3. Focusing the problem: Facet-wise analysis
4. Results on imbalanced data
5. Rebalancing the imbalanced data
6. Conclusions and further work

Id.	Dataset	#Ins.	#At.	%Min.	%Maj.	ir
bald1	*balance-scale disc. 1*	625	4	7.84%	92.16%	11.76
bald2	*balance-scale disc. 2*	625	4	46.08%	53.92%	1.17
bald3	*balance-scale disc. 3*	625	4	46.08%	53.92%	1.17
bpa	*bupa*	345	6	42.03%	57.97%	1.38
glsd1	*glass disc. 1*	214	9	4.21%	95.79%	22.75
glsd2	*glass disc. 2*	214	9	6.07%	93.93%	15.47
glsd3	*glass disc. 3*	214	9	7.94%	92.06%	11,59
glsd4	*glass disc. 4*	214	9	13.55%	86.45%	6.38
glsd5	*glass disc. 5*	214	9	32.71%	67.29%	2.06
glsd6	*glass disc. 6*	214	9	35.51%	64.49%	1.82
h-s	*heart-disease*	270	13	44.44%	55.56%	1.25
pim	*pima-inidan*	768	8	34.90%	65.10%	1.87
tao	*tao-grid*	1888	2	50.00%	50.00%	1.00
thyd1	*thyroid disc. 1*	215	5	13.95%	86.05%	6.17
thyd2	*thyroid disc. 2*	215	5	16.28%	83.72%	5.14
thyd3	*thyroid disc. 3*	215	5	30.23%	69.77%	2.31
wavd1	*waveform disc. 1*	5000	40	33.06%	66.94%	2.02
wavd2	*waveform disc. 2*	5000	40	33.84%	66.16%	1.96
wavd3	*waveform disc. 3*	5000	40	33.10%	66.90%	2.02
wbcd	*Wis. breast cancer*	699	9	34.48%	65.52%	1.90
wdbc	*Wis. diag. breast cancer*	569	30	37.26%	62.74%	1.68
wined1	*wine disc. 1*	178	13	26.97%	73.03%	2.71
wined2	*wine disc. 2*	178	13	33.15%	66.85%	2.02
wined3	*wine disc. 3*	178	13	39.89%	60.11%	1.51
wpbc	*wine disc. 4*	198	33	23.74%	76.26%	3.21

	C4.5	SMO	IBk	XCS	UCS
bald1	0,00	0,00	0,00	0,00	0,00
bald2	69,28 ••	83,98 ◦◦◦	81,16 ◦◦◦	71,22 ••	69,77 ••
bald3	71,21 ••	85,69 ◦◦◦	82,11 ◦◦◦	70,07 ••	73,65 ••
bpa	33,50 ••◦	0,00 ••••	32,40 ••◦	47,22 ◦◦◦	47,21 ◦◦◦
glsd1	79,60 ◦◦	0,00 •••	69,32 ◦◦	20,00 ••	59,11 ◦
glsd2	33,95 •	15,00 ••	24,13 •	59,40 ◦	74,25 ◦◦◦
glsd3	28,78 ◦◦◦	0,00 ••	0,00 ••	0,00 ••	19,39 ◦◦◦
glsd4	73,36	80,33	77,07	80,33	83,61
glsd5	65,35 ◦	9,58 ••••	62,26 ◦	67,82 ◦	64,45 ◦
glsd6	52,03 ◦	0,00 ••••	61,74 ◦	61,08 ◦	57,90 ◦
h-s	63,70	68,80 ◦	64,40 ◦	60,32	54,87 ••
pim	44,96	48,36	46,91	46,06	47,88
tao	91,00 •◦◦◦	70,57 ••••	94,25 ◦◦◦◦	82,90 ••◦◦	78,79 •••◦
thyd1	87,53	76,67	76,67	78,69	92,32
thyd2	93,12 ◦	54,17 ••••	77,90 ◦	82,50 ◦	93,12 ◦
thyd3	87,31 ◦	33,81 ••••	81,12 ◦	89,74 ◦	87,97 ◦
wavd1	67,80 ••••	78,65 ◦◦	72,28 •••◦	80,43 ◦◦◦	76,35 •◦◦
wavd2	62,54 ••••	72,35 ◦◦	67,49 •••◦	73,48 ◦◦	71,50 ◦◦
wavd3	68,61 ••••	79,61 ◦◦	74,14 ••◦	81,01 ◦◦◦	76,62 •◦
wbcd	89,10 •••	92,72 ◦	92,72 ◦	92,29	94,11 ◦
wdbc	88,83 ••	94,27 ◦◦◦	93,47 ◦	90,30 •	89,67 •
wined1	85,58 •••	98,46 ◦	94,98	99,23 ◦	99,23 ◦
wined2	91,83 •	97,51	97,50	99,17 ◦◦	91,76 •
wined3	87,64 •	97,14 ◦◦	87,94	93,43 ◦	85,36 ••
wpbc	33,96 ◦	9,37 ••	28,98 ◦	20,99	16,97
Avg.	66,02	53,88	65,64	60,17	68,23
Score	30-14	33-20	13-24	11-22	15-22
Score$_{ir>5}$	1-6	11-0	3-3	4-2	0-8

● → **significant degradation with respect to another method**

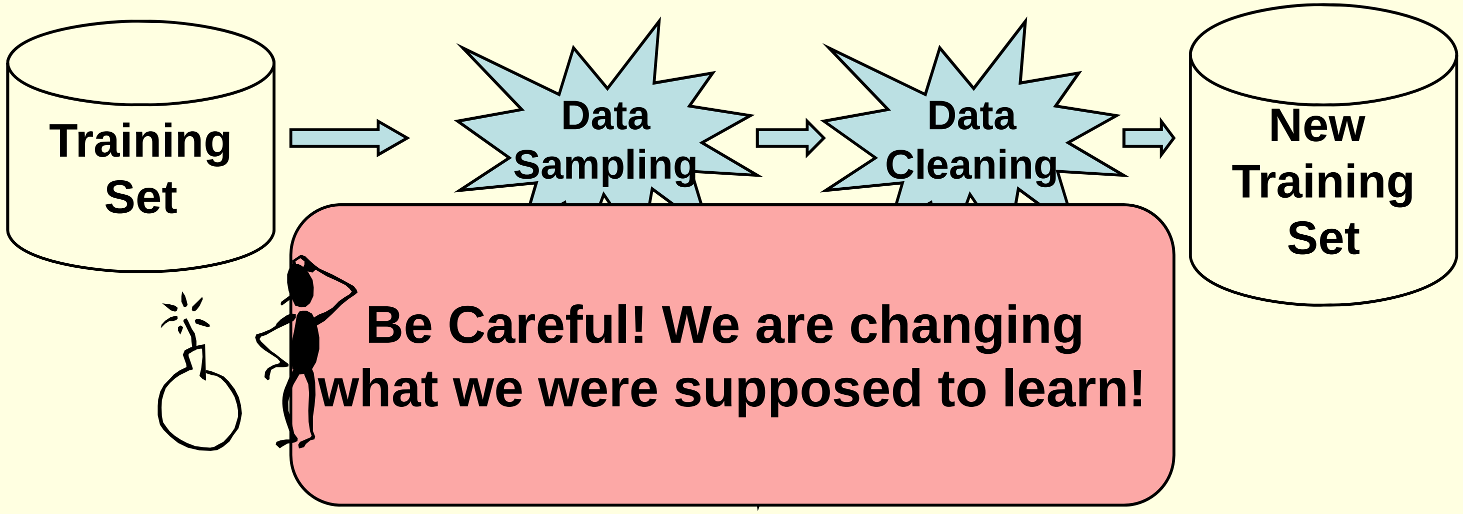
○ → **Significant improvement with respect to another method**

Outline

1. Why do we care about mining rarities?
2. The UCS Classifier System
3. Focusing the problem: Facet-wise analysis
4. Results on imbalanced data
5. Rebalancing the imbalanced data
6. Conclusions and further work

Resampling techniques

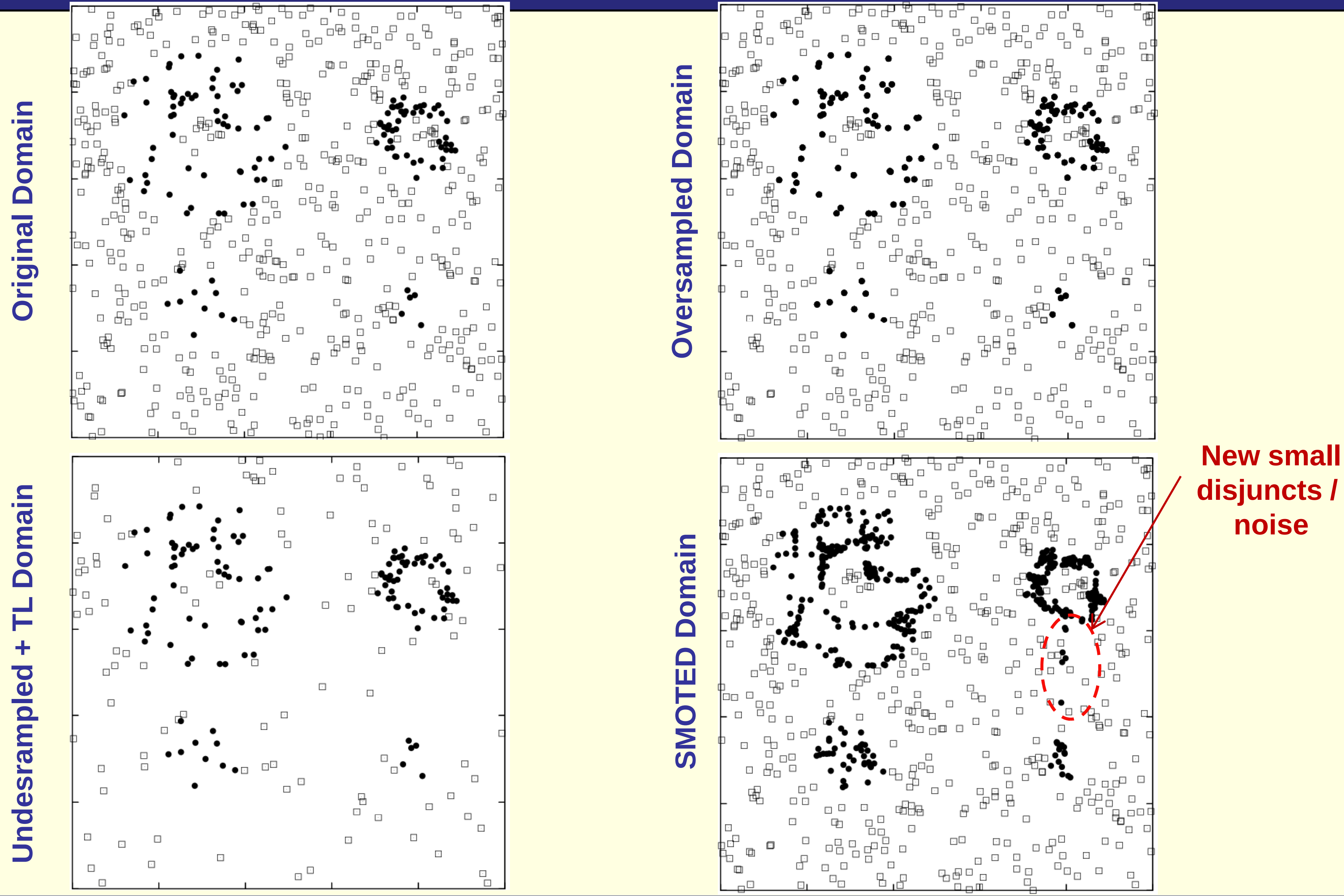
- **Resample methods**



Lots of approaches in the literature:

1. Oversampling
2. Undersampling
3. Tomek links
4. CNN
5. Smote
6. Any combination

Resampling techniques



	Resamp. Method	1st	2nd	3rd	4th	5th
C4.5	*original*	6	2	5	**9**	3
	oversampling	7	4	**8**	4	2
	undersampling TL	0	5	7	6	7
	smote	**10**	**8**	3	2	2
	csmote	2	6	2	4	**11**
SMO	*original*	6	2	2	4	**11**
	oversampling	**11**	**11**	3	0	0
	undersampling TL	2	8	**9**	3	3
	smote	3	3	8	7	4
	csmote	3	1	3	**11**	7
IBk	*original*	6	6	2	6	5
	oversampling	4	**8**	**11**	1	1
	undersampling TL	4	2	5	4	**10**
	smote	**10**	4	2	**7**	2
	csmote	1	5	5	**7**	7
XCS	*original*	3	5	2	**6**	**9**
	oversampling	7	5	4	1	8
	undersampling TL	1	**8**	**10**	**6**	0
	smote	**11**	3	2	**6**	3
	csmote	3	4	7	**6**	5
UCS	*original*	2	4	**8**	5	6
	oversampling	6	5	5	**7**	2
	undersampling TL	5	4	7	**7**	2
	smote	**7**	**11**	4	1	2
	csmote	5	1	1	5	**13**

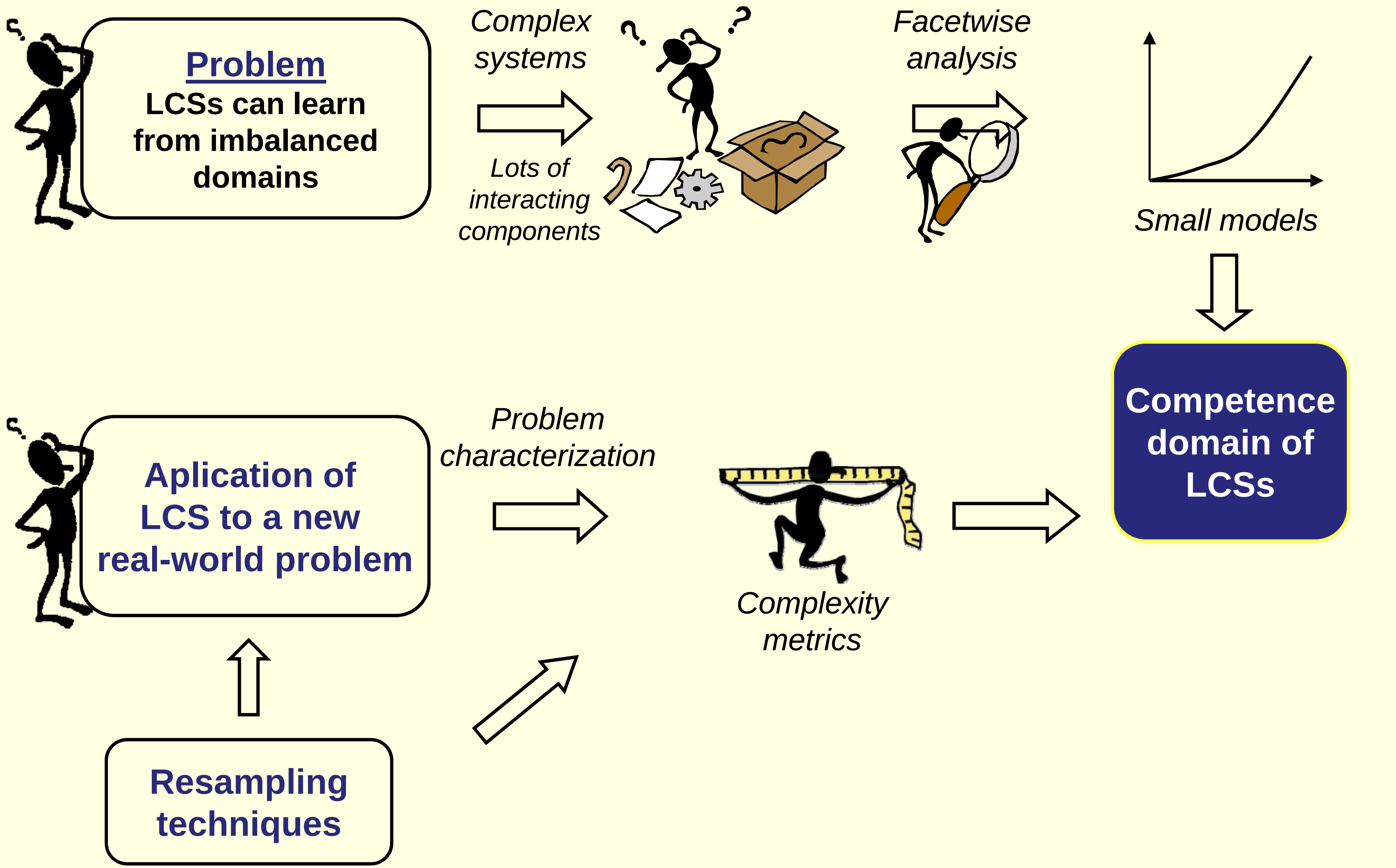
- **Best resampling technique depends on each method**
- **Oversampling methodologies lead to imporvements on average**
- **Still, depends on the data set**

- **What are really these resampling techniques doing to my domain?**

Overview

1. Description of XCS
2. Facetwise analysis of LCSs
3. Resampling techniques
4. Real-world problems
5. Further work

Current/Further work



Some Results on the use of UCS in Imbalanced Domains

Albert Orriols Puig
aorriols@salle.url.edu

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